Applied Soft Computing 71 (2018) 255–267

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

MTEEGC: A novel approach for multi-trial EEG clustering

Chenglong Dai[a], Dechang Pi[a,b,*], Lin Cui[a,c], Yanlong Zhu[a]

[a] College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China
[b] Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing 211106, China
[c] Intelligent Information Processing Laboratory, Suzhou University, Suzhou 234000, China

ARTICLE INFO

Article history:
Received 3 March 2018
Received in revised form 3 May 2018
Accepted 2 July 2018
Available online 11 July 2018

Keywords:
Multi-trial EEG
Clustering
Cross correlation
Local penalty
Feature extraction

ABSTRACT

This paper explores multi-trial EEG (Electroencephalography signal) clustering and proposes a novel centroid-based approach for it. It firstly utilizes an improved cross correlation to measure similarities of multi-trial EEGs and then proposes an optimal EEG feature extraction to seek cluster centroids based on the improved cross correlation similarities. Finally, it leads to a novel algorithm called MTEEGC for multi-trial EEG clustering. MTEEGC yields high-quality multi-trial EEG clustering with respect to the intra-cluster compactness as well as the inter-cluster scatter. Meanwhile, it also demonstrates the superiority of MTEEGC in clustering accuracy over 10 state-of-the-art time series clustering algorithms through a detailed experimentation using standard cluster validity criteria on 5 real-world multi-trial EEG datasets. Especially, compared with the worst and the best algorithms in the 10 baseline algorithms, MTEEGC respectively achieves 36.11% and 2.53% mean improvements with clustering accuracy (i.e., RI) on 5 multi-trial EEG datasets.

1. Introduction

With the globally ever-increasing demand for more affordable and effective clinical and healthcare services, new advanced techniques therefore should be developed to aid in disease diagnosis [1,2], monitoring [3], and treatment of abnormalities or diseases of the human body [4,5]. Biosignals in their manifold forms are useful information sources, which have the potential to facilitate such advancements. Furthermore, the brain electroencephalography (EEG), generated by population of neurons in human brain [6], is one of the most widely applied biosignals.

It is believed that from early stage and throughout life, EEG signals generated by the brain represent not only the brain function but also the status of the whole body. Therefore, EEG nowadays is widely used, with classification techniques, to diagnose neurocognitive disorders or cerebral diseases such as Alzheimer's Disease (AD) [7–9], epileptic seizure [10–13], stroke [14,15], Amyotrophic Lateral Sclerosis (ALS) [16,17] and so on [18] in a non-invasive way. However, these applications require EEG labels. But in fact, with the increasing amount of unlabeled EEGs, especially from the subjects suffering with the cerebral disorders mentioned above, unlabeled EEG analysis becomes a tough task due to the time and human resource costing. Moreover, the absence of labels also limits the applications of those supervised methods in practice such as classification. Among the techniques potentially applied to unlabeled EEG, clustering is the most popular and widely used since it performs with neither the supervision of human nor time consuming of labeling. Also, meaningful patterns and correlations in unlabeled EEG can be identified based on clustering. In this paper, we explore the tough but valuable problem of unlabeled multi-trial EEG clustering and we propose a novel approach to cluster multi-trial EEG in an unsupervised way.

1.1. Motivation

It is clear that the studies of EEG, based on classification techniques [13,19,20], pave the way for diagnosis of many neurological disorders and other abnormalities in the human body, but they belong to the supervised methodology. In fact, the amount of unlabeled EEG is increasing rapidly and its labeling is time and human resource consuming. Due to the lack of labels, the previously proposed classification approaches are unsuitable or inapplicable for unlabeled EEG analysis. With the growing interest in unlabeled EEG analysis, clustering becomes a new direction to research unlabeled EEG signals.

Recently, most researches on EEG analysis focus on single trial [21–25] which probably ignores the correlation among multi-trial EEG signals from the same subject, not mention to different

* Corresponding author at: College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China.
E-mail address: nuaacs@126.com (D. Pi).

https://doi.org/10.1016/j.asoc.2018.07.006
1568-4946/

subjects. In other words, multi-trial EEGs [26,27] contain more interesting and valuable information of healthy and patient subjects, which roundly reflects the status of “healthy” and “patient” of subjects. Unfortunately, as far as we know, multi-trial EEG clustering is not well addressed and there are few studies on it. Although EEG can be regarded as time series and traditional time series clustering methods have made many achievements in various applications, they may be not applicable for multi-trial EEG clustering because of its high weakness, complexity, high dimension, non-stationary, oscillation, low signal-to-noise ratio, that are different with most traditional time series. Consequently, it is necessary to propose new techniques to analyze unlabeled multi-trial EEG signals.

To the end, this paper aims to cluster multi-trial EEG and proposes a novel algorithm named MTEEGC, which utilizes normalized cross correlation to measure similarities among multi-trial EEGs and extract multi-trial EEG’s features (cluster centroids) to cluster.

1.2. Related works

Clustering, widely applied to analyze time series data, is worthy of trying to handle unlabeled multi-trial EEGs. Furthermore, to the best of our knowledge, there are few studies on multi-trial EEG clustering due to its characteristics and difficulties [26–28]. Fortunately, multi-trial EEG can be regarded as one type of time series since it is a collection of observations ordered in time [29] and time series clustering approaches may provide solutions to it. In recent years, many time series clustering approaches are emerged out and have made many achievements [30,31], which probably can be divided into 5 categories.

1 The classic time series clustering methods: k -means and its improved variant kmTSC (k -means++ [32]). kmTSC specifies a procedure with probability proportional to initialize the cluster centers before proceeding with the standard k -means algorithm. Both of k -means and kmTSC randomly initialize cluster centers and then align time series into their closest centers. Consequently, they are highly sensitive to the initial centroid time series.
2 The feature selection-based time series clustering methods such as udTSC [33], ndTSC [34], ruTSC [35], rsTSC [36]. These approaches firstly select discriminative features of time series and then apply k -means to cluster with such features. These feature selection-based time series methods depend on not only the selected features, but also the number of selected features which is set in supervision. Therefore, they rely on many choices with different numbers of features for the time series clustering, so it requires extra time consumption and supervision.
3 Artificial neural networks (ANN)-based time series clustering methods such as ART (Adaptive resonance theory) networks [37] and its invariants [38,39], self-organizing map (SOM) networks [40,41], and so on [38]. Although they have made many achievements in clustering, they are mainly available to low-dimensional data and encounter some barriers for some longer higher dimensional series, such as EEG signals. Meanwhile, they likely cost much time to train (slow learning) the network and lose accuracy during the network processing [37,38].
4 The distance-based time series clustering methods such as kcTSC [42], kdTSC [43]. kcTSC uses spectral matrix to select cluster centers and kdTSC utilizes DTW barycenter averaging (DBA) which is based on dynamic time warping (DTW [44]) to select cluster centers. And then they based on spectral distance and DBA distance respectively to cluster time series into such centers. However, these distance measures cost lots of time to search the cluster centers through iteratively computing DBA distances among time series.
5 The shape- or shapelet-based time series clustering algorithms such as ksTSC [45] and usTSC [46]. They firstly search the shapes and shapelets from original time series as the reference features, and then align time series into clusters based on these extracted shapes and shapelets. Although they are domain independent and perform to cluster traditional time series well, they may not suitable for multi-trial EEG clustering due to the special characteristics of multi-trial EEG signals. Moreover, the shapelet-based time series clustering methods are probably impacted by the number of shapelets and likely requires lots of time to extract optimal shapelets.

Due to the specific characteristics of multi-trial EEG: high weakness, complexity, high dimension, non-stationary, oscillation, low signal-to-noise ratio, these promising approaches for conventional time series clustering may not suitable for multi-trial EEG clustering. To solve this problem, we propose a novel approach for multi-trial EEG clustering in this paper, which is called MTEEGC. In detail, MTEEGC applies the cross correlation with local penalty to measure the similarities among multi-trial EEGs and defines a cluster centroid function to search the optimal centers for multi-trial EEG assignment, so as to cluster these multi-trial EEGs to the proper cluster centroids in a way that EEG signals in the same cluster are highly compact while those in different clusters are highly separated.

1.3. Contributions and outiline

This paper explores multi-trial EEG clustering and proposes a novel multi-trial EEG clustering method based on an improved cross-correlation similarity. The contributions of this paper are highlighted as follows.

- Multi-trial EEG clustering is explored in this paper which is not well addressed by traditional time series clustering approaches, and this is the first try to cluster multi-trial EEG signals as far as we know. Finally an improved cross correlation-based multi-trial EEG clustering approach is proposed, which is named MTEEGC.
- A local penalty is brought in to modify the traditional cross correlation. It improves the measuring capacity by weighing the local tendency as well as cross correlation to measure the similarities among multi-trial EEGs.
- With the improved cross correlation-based similarity measurement, a multi-trial EEG feature extraction is proposed. It is based on a centroid sequence searching strategy. It is a global optimization that can obtain the most representative centroid sequences (EEG features) for multi-trial EEGs.
- The efficacy and superiority of MTEEGC are demonstrated via a detailed experimentation compared with 10 state-of-the-art time series clustering algorithms on real-world multi-trial EEG datasets through using standard cluster validity criteria including intra-cluster compactness, inter-cluster scatter, integrated ratio, rand index (RI), F-score, and Fleiss’ kappa.

The rest of the paper is organized as follows. In Section 2, a succinct background of cross correlation and data normalization is introduced. The proposed method of multi-trial EEG clustering (MTEEGC) based on the improved cross correlation with local penalty is presented in Section 3, along with time complexity analysis. Then a detailed experimentation for multi-trial EEG clustering on real-world multi-trial EEG datasets is carried out in Section 4. Finally, a summary of our work is presented in Section 5 and some directions for future work are also highlighted in this section.