Research Article

Fast Enhanced Exemplar-Based Clustering for Incomplete EEG Signals

Anqi Bi, Wenhao Ying, and Lu Zhao (ID)

School of Computer Science and Engineering, Changshu Institute of Technology, Changshu, Jiangsu, China

Correspondence should be addressed to Lu Zhao; constance@cslg.edu.cn

The diagnosis and treatment of epilepsy is a significant direction for both machine learning and brain science. This paper newly proposes a fast enhanced exemplar-based clustering (FEEC) method for incomplete EEG signal. The algorithm first compresses the potential exemplar list and reduces the pairwise similarity matrix. By processing the most complete data in the first stage, FEEC then extends the few incomplete data into the exemplar list. A new compressed similarity matrix will be constructed and the scale of this matrix is greatly reduced. Finally, FEEC optimizes the new target function by the enhanced α-expansion move method. On the other hand, due to the pairwise relationship, FEEC also improves the generalization of this algorithm. In contrast to other exemplar-based models, the performance of the proposed clustering algorithm is comprehensively verified by the experiments on two datasets.

1. Introduction

Epilepsy is a common disease of nervous system, which is characterized by sudden brain dysfunction. Although there are many other neuroimaging modalities for the recognition of brain activity, EEG signals have a high temporal resolution which is up to the millisecond level, and its acquisition equipment is inexpensive, portable, and noninvasive. Nowadays, most diagnoses of epilepsy are based on clinical experience and the analysis of electroencephalogram (EEG) signals. Compared with manual diagnostic method, machine learning methods are less time-consuming and more consistent [1–6]. Specifically, many machine learning methods such as support vector learning [7, 8], Takagi–Sugeno–Kang (TSK) fuzzy system [9, 10], and Naïve Bayes [11] have been applied.

As we know that brain activity is a nonlinear, unstable complex network system, EEG signals we usually get are complicated. That is to say, some EEG signals are complete while others may miss some features, namely, incomplete. Therefore, recognition of epilepsy based on machine learning models will be more promising compared with clinical diagnosis depending on experience. Moreover, EEG

signals have the characteristics of high dimension and stochasticity which limit the performance of most existing clustering models, such as k-means [11] and fuzzy c mean (fcm) [12]. K-means and fcm clustering models need to preset the number of clusters in advance. More specifically, the performance of the k-means model relies on the initialization of data, while the fcm model requires high interpretability. Thus, we focus on the exemplar-based clustering model [13] which is proposed by Frey in this paper. The exemplar-based clustering model has the advantages of automatically obtaining the cluster number, high efficiency, and not relying on the initialization of data.

In conclusion, we consider the scenario of EEG signals consisting of most complete data and few incomplete data in this paper, as shown in Figure 1. Based on the previous work about the recognition of epileptic signals, we propose a novel fast enhanced exemplar-based clustering (FEEC) model for incomplete EEG signals. As shown in Figure 1, different from existing exemplar-based clustering models, FEEC compresses the exemplar list and reduces the pairwise similarity matrix, and then FEEC optimizes the target model by the enhanced α-expansion move framework. Moreover, the contributions of this paper can be highlighted as follows:
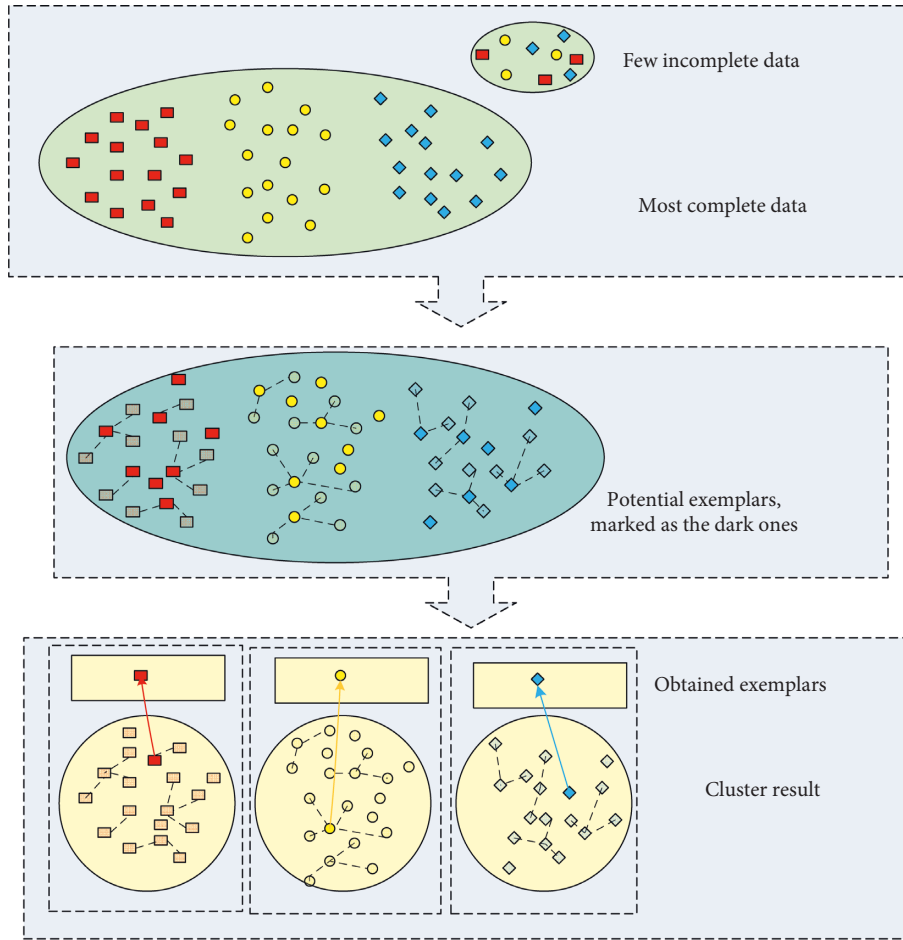
Figure 1: Clustering procedure of FEEC algorithm.

(1) We extend the existing exemplar-based clustering algorithm into a fast version by compressing the potential exemplar lists in this study. FEEC compresses the number of potential exemplars by processing the most complete data in the first stage and extends the few incomplete data into the exemplar list. So, the complexity of FEEC is reduced as well.

(2) Along with most existing exemplar-based clustering models, FEEC is built on the pairwise similarity matrix of data. Thus, after compression, FEEC would construct a new reduced similarity matrix, and the generalization of this algorithm is improved.

(3) Moreover, this paper also considers the fact that the graph cuts [14] based optimization performs better than those loopy belief propagation (LBP) [15] based structure. So, the proposed FEEC algorithm optimizes the target model by the enhanced α-expansion move framework [16, 17].

(4) Experimental results of both synthetic and real-world datasets indicate the promising efficiency of the proposed FEEC algorithm.

The rest of this paper is listed as follows. In Section 2, we introduce some static exemplar-based clustering models. Section 3 discusses the proposed FEEC algorithm step by step. In Section 4, we analyze the experimental results and the comparison of FEEC and other existing methods. Section 5 concludes this whole paper.

2. Background

Since EEG signal feature extraction methods and exemplar-based clustering models are two important supporting theories for the FEEC model in this study, we will briefly introduce several feature extraction methods and exemplar-based clustering models in this section.

2.1. Feature Extraction Methods. Original EEG signals have the characteristics of high dimensionality, stochasticity, and nonlinearity. It would be computationally very expensive to extract features from raw EEG signals; nowadays, many feature extraction methods have been proposed to handle this problem. In sum, there are three categories, i.e., time-domain features, frequency-domain features, and time-frequency features.

More specifically, in time-domain analysis, statistics component features of the raw EEG signals will be analyzed [18]. In frequency-domain analysis, power spectrum analysis and short-time Fourier transform (STFT) [19, 20] are commonly used. In time-frequency analysis, time and

frequency domain are simultaneously extracted from non-stationary EEG signals. Wavelet and other improved versions [21, 22] are widely used in EEG signal processing. We utilize KPCA to extract feature in this paper.

2.2. Exemplar-Based Clustering Models.

Exemplar-based clustering models select cluster centers, namely, exemplars, from existing actual data. We focus on exemplar-based clustering models in this paper and briefly introduce affinity propagation (AP) [13] and enhanced α-expansion move (EEM) [17] in this section. And several extended versions for different scenarios are shown in Table 1. The target fucntion defined by exemplar-based clustering model equals to the minimization problem of energy function of Markov random field(MRF). Two existing optimization startegies have been utilized and evolved into AP and EEM frameworks accordingly. Moreover, loopy belief propagation (LBP) [23] is used in AP, while graph cuts technique [15] is used in EEM, respectively.

$$\mathbf{R}(i,k) = \mathbf{S}(i,k) - \max_{j,\text{s.t. } j \neq k} \{\mathbf{A}(i,j) + \mathbf{S}(i,j)\}, \tag{3}$$

$$\mathbf{A}(i,k) = \begin{cases} \min\{;0, \mathbf{R}(k,k) + \sum_{j,\text{s.t. } j \neq \{i,k\}} \{\max(0, \mathbf{R}(j,k))\}, & i \neq k, \\ \sum_{j,\text{s.t. } j \neq \{i,k\}} \{\max(0, \mathbf{R}(j,k))\}, & i = k, \end{cases} \tag{4}$$

where $\mathbf{S}$ is the similarity matrix of data points and is defined as $\mathbf{S}(i,j) = -\|\mathbf{x}_i - \mathbf{x}_j\|^2$. Meanwhile, $\mathbf{S}(k,k) = p$ where p is named as preferences in this framework. Moreover, its value should be independent and can be set to a constant.

AP does not require presetting the number of the cluster and the performance is stable. Considering these advantages, many extended versions of AP have been proposed [24, 25]. Specifically, AP defines fading factor to adjust the iteration speed; adAP [24] is proposed to determine this fading factor adaptively. Moreover, several extended versions of AP methods have been proposed to deal with large data and link constraints. For instance, IAPKM, IAPNA, and IAPC [26, 27] employ incremental strategy and semisupervised AP and SSAP [28] concentrate on instance-level constraints. A two-stage fast version of AP (FAP) [29] is also proposed to improve the efficiency. However, although AP has been obtaining its success

2.2.1. Affinity Propagation.
AP is based on message passing among data points, and its target function is defined as follows:

$$\max_E \sum_{p=1}^{N} \mathbf{S}(i,k) - \sum_{p=1}^{N} \delta_p(E), \tag{1}$$

where

$$\delta_p(E) = \begin{cases} \infty, & \text{if } E(x_p) \neq p, \text{ but } \exists x_q: E(x_p) = p, \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

where $\mathbf{X} = \{x_1, x_2, \dots, x_N\} \in \mathbb{R}^{N*D}$ is an input dataset and N is the total number of D-dimensional data points. $\mathbf{S}E$ is the output of this framework, and the element $E(x_p)$ is referred to the exemplar for each x_p.

According to AP, each point receives availability message $\mathbf{A}(i,k)$ and sends responsibility $\mathbf{R}(i,k)$ message simultaneously, which are defined as follows:

in various applications, when we attempt to directly apply AP to incomplete EEG signals, the performance is unsatisfactory.

2.2.2. Enhanced α-Expansion Move.
In 2014, Zheng and Chen [17] utilized enhanced α-expansion move framework to optimize the object function of exemplar-based clustering models and accordingly proposed the EEM clustering model. In line with the mathematical symbols above, the target function of EEM is defined as follows:

$$\max_E \sum_{p=1}^{N} s\left(x_p, x_{E(x_p)}\right) - \sum_{p=1}^{N}\sum_{q>p}^{N} \theta_{p,q}\left(E(x_p), E(x_q)\right), \tag{5}$$

where

$$\theta_{p,q}\left(E(x_p), E(x_q)\right) = \begin{cases} M, E(x_p) = q, E(x_q) \neq q, \text{ or } E(x_q) = p, & E(x_p) \neq p, \\ 0, & \text{otherwise}, \end{cases} \tag{6}$$

$$\theta_{p,q}\left(E(x_p), E(x_q)\right) + \theta_{p,q}(\alpha, \alpha) \leq \theta_{p,q}\left(E(x_p), \alpha\right) + \theta_{p,q}\left(\alpha, E(x_q)\right). \tag{7}$$

In terms of [17], α-expansion move algorithm has been proved to be effective in the optimization of the target function equation (5). Specifically, when $\forall p, q, E(\mathbf{x}_p), E(\mathbf{x}_q), \alpha \in \{1, 2, \dots N\}$, equation (7) is verified.

TABLE 1: Descriptions of several state-of-the-art exemplar-based algorithms.

	Optimization	Extended version	Descriptions
AP	Message-passing	adAP	Determine fading factor adaptively
		IAPKM, IAPNA, IAPC	Incremental version for data stream
		SAP,SSAP	Deal with instance-level constraints
		FAP	Two-stage fast version
EEM	Graph cuts	DSC	Dynamic data stream cluster
		IEEM	Deal with link constraints
		FEEC (newly proposed)	Two-stage fast version

Furthermore, according to graph theory, in the fast α-expansion move algorithm, the expansion range is limited in a one exemplar. To break this limit, the EEM model enlarges the range to the whole exemplar set E when optimizing and defines a second exemplar $S(i)$ for each point $\mathbf{x}_i$ as follows:

$$S(i) = \underset{s \in (E/l)}{\arg\max} \, d(\mathbf{x}_i, \mathbf{x}_s), \quad \forall \mathbf{x}_i \in \mathbf{X}_l, \tag{8}$$

where $\mathbf{X}_l = \{\mathbf{x}_i \mid E(\mathbf{x}_i = l)\}$ is the dataset among which the exemplar is l and $s \in (E/l)$ represents other exemplars in E except for l.

The EEM clustering model is a state-of-the-art achievement of exemplar-based clustering model and has been proved to be efficient and effective for numerous scenarios [16, 17, 30]. IEEM [30] is proposed to deal with link constraints by embedding a bound term in the target function. For dynamic data stream, Bi and Wang [16] also proposed an incremental EEM version DSC which processes data chunk by chunk. However, for incomplete EEG signals, these methods would not recognize epilepsy well.

3. Fast Enhanced Exemplar-Based Clustering Model

In this section, the proposed FEEC model will be stated and theoretically analyzed in detail. We first compress the exemplar list and reduce the pairwise similarity matrix, and then the target model is optimized by the enhanced α-expansion move framework.

3.1. Framework. As mentioned in the introduction section, we focus on the incomplete EEG signals which consist of most complete data and few incomplete data. To improve the efficiency of the EEM clustering model for these incomplete EEG signals, the proposed FEEC framework includes two stages, namely, compression stage and optimization stage. As shown in Figure 2, the compression stage compresses the potential exemplar list and the optimization stage determines the optimal exemplars from the potential exemplar list. Accordingly, the target function can be defined as follows:

$$\max_{E} \sum_{i=1}^{N} s(\mathbf{x}_i, \mathbf{x}_{E(\mathbf{x}_i)}) - \sum_{i=1}^{N} \eta_i(E), \tag{9}$$

where $\mathbf{X} = [\mathbf{X}_c, \mathbf{X}_l]$ is the input dataset consisting of most complete data $\mathbf{X}_c = \{\mathbf{x}_{c,1}, \mathbf{x}_{c,2}, \ldots, \mathbf{x}_{c,N_c}\}$ and few incomplete data $\mathbf{X}_l = \{\mathbf{x}_{l,1}, \mathbf{x}_{l,2}, \ldots, \mathbf{x}_{l,N_l}\}$. The total number of data is

defined as $N = N_c + N_l$, where N_c and N_l are the number of complete and incomplete data, respectively. Remember that we only consider the scenario that $N_c \gg N_l$ in this study. The second term in equation (9) guarantees the validity of the exemplar list; its definition is similar to that of $\boldsymbol{\delta}$ in equation (2). In the end, $E = \{E(\mathbf{x}_1), E(\mathbf{x}_2), \ldots, E(\mathbf{x}_N)\}$ represents the exemplar set in question.

In the compression stage, the number of potential exemplar list will be reduced by exemplar-based selection algorithm, namely, EEM method in this study. To be specific, we apply the EEM model on the most complete data to obtain the potential exemplars for these data. FEEC also pulls the few incomplete data into this potential exemplar list and then constructs compressed similarity matrix. Therefore, after compression, only the pairwise similarities between data and potential exemplars would be preserved. Considering that the FEEC method is built on the pairwise similarity matrix, the following clustering procedure would be applied on this compressed similarity matrix. Furthermore, the scale of similarity matrix is reduced from N^2 to Nc, where N and c are the number of data and potential exemplars, respectively.

In the optimization stage, only the similarity relationship between data and potential exemplars is considered. The new target function after compression is similar to that of other exemplar-based clustering model, like equations (1) and (5), so we take graph cuts and LBP into account. Nevertheless, graph cuts based optimization framework outperforms LBP structure [31]. So, the proposed FEEC utilizes the α-expansion move method to optimize the new target function. Moreover, along with EEM, FEEC also expands the expansion move space from a single data to the second optimal exemplar.

3.2. Compression Stage. In the compression stage, the target function of complete data can be defined as follows:

$$\min_{E_c} \sum_{i=1}^{N_c} d(\mathbf{x}_{c,i}, \mathbf{x}_{E_c(\mathbf{x}_{c,i})}) + \sum_{i=1}^{N_c} \sum_{\substack{j=1 \\ j>i}}^{N_c} \theta_{i,j}(E_c(\mathbf{x}_{c,i}), E_c(\mathbf{x}_{c,j})), \tag{10}$$

where $\mathbf{X}_c = \{\mathbf{x}_{c,1}, \mathbf{x}_{c,2}, \ldots, \mathbf{x}_{c,N_c}\} \in \mathbb{R}^{N_c * D}$ is the complete D-dimensional data and N_c is the number of these data. E_c is the potential exemplar list for most complete data, and the element among $E(\mathbf{x}_{c,i})$ is referred to the potential exemplar for each $\mathbf{x}_{c,i}$. The optimization framework of other exemplar-based clustering models, like EEM, can be utilized to solve
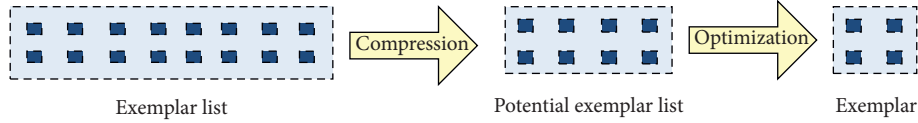
FIGURE 2: Framework of FEEC algorithm: compression stage and optimization stage.

equation (10). In this paper, we select the graph cuts algorithm instead of message-passing algorithm to compress the potential exemplar list. Thus, the potential exemplar list for complete data E_c can be determined, and the number of potential exemplars is defined as c_c.

The potential exemplar list after compression stage would be

$$E_{\text{new}} = [E_c, E_l], \tag{11}$$

where E_l is the exemplar set for the few incomplete data, which is the incomplete data itself actually. That is to say, $E_l(\mathbf{x}_{l,i}) = i$.

In this stage, we reduced the number of potential exemplars from N_c to c_c. In terms of the analysis in [13, 17, 30], the time complexity of this stage will be $O(N_c^2)$. Compared with the time complexity $O(N^2)$, if we apply exemplar-based clustering model directly considering the fact that $N < N_c$, the time complexity of this compression algorithm would be acceptable.

Therefore, on the basis of the new exemplar list after compression, we can construct the new similarity matrix $\mathbf{S}_{\text{new}} \in \mathbb{R}^{N \times c}$; the element relates to the distance, namely, $\mathbf{S}_{\text{new}}(i, j) = -\|\mathbf{x}_i - \mathbf{x}_{E_{\text{new}}(j)}\|^2$. The scale of the similarity matrix reduces from N^2 to Nc, where $c = c_c + N_l$ represents the number of potential exemplars.

3.3. Optimization Stage.
After compression, we define the new target function as follows:

$$\max_E \sum_{i=1}^{N} \sum_{j=1}^{c} \mathbf{S}_{\text{new}}(i, j) - \sum_{i=1}^{N} \eta_i(E), \tag{12}$$

where $\mathbf{S}_{\text{new}}$ is the new similarity matrix constructed after compression.

In this section, we construct an optimization framework for equation (12). The second term of equation (12) is set to guarantee the validity of the exemplar list; in order to utilize the graph cuts based method, this term should be pairwise [17]. So, $\eta_i(E)$ is modified as $\eta_{i,j}(E)$. Furthermore, similar to equation (5), we define $\eta_{i,j}(E)$ as follows:

$$\eta_{i,j}(E) = \begin{cases} M, & E(\mathbf{x}_i) = j, E(\mathbf{x}_j) \neq j, \text{ or } E(\mathbf{x}_j) = i, \quad E(\mathbf{x}_i) \neq i, \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

It has been proved that with the definition of $\eta_{i,j}(E)$, equation (12) can be optimized by the enhanced α-expansion method [30]. To improve the efficiency of framework, this method enlarges the expansion move to the second optimal exemplar.

Before optimization, we explain several symbols involved. First, we define $\mathbf{X}_e$ as those data with exemplar $\mathbf{x}_e$ and $\mathbf{x}_\alpha$ as the current potential exemplar. Then, the enhanced α-expansion move method considers the second optimal exemplar, which is defined as

$$S(\mathbf{x}_i) = \underset{s \in (E/\alpha)}{\text{argmax}}\, \mathbf{S}_{\text{new}}(\mathbf{x}_i, \mathbf{x}_s), \quad \forall \mathbf{x}_i \in \mathbf{X}_\alpha, \tag{14}$$

where (E/α) is the potential exemplar list except for α.

Apparently, this optimization method should consider two cases, namely, $\mathbf{x}_e$ is among exemplar list or not, as shown in Figures 3 and 4. To be specific, Figure 3 illustrates the case when $\mathbf{x}_\alpha$ is an exemplar, while Figure 4 shows the case when $\mathbf{x}_\alpha$ is not an exemplar. Remember that only when $\mathbf{x}_\alpha$ is a potential exemplar, $S(\mathbf{x}_i)$ works. We utilize the concepts of "energy reduction" because this method was first used to optimize the Markov random field (MRF) energy function.

In the situation shown in Figure 3, either $\forall \mathbf{x}_i \in \mathbf{X}_\alpha$ changes its exemplar to $S(\mathbf{x}_i)$ or nothing is changed. Therefore, the energy reduction $R1$ would be defined as

$$R1 = \max(0, R1_\alpha), \tag{15}$$

where $R1_\alpha$ is the energy reduction when $\forall \mathbf{x}_i \in \mathbf{X}_\alpha$ changes its exemplar to $S(\mathbf{x}_i)$ and is defined as

$$R1_\alpha = \sum_{\mathbf{x}_i \in \mathbf{X}_\alpha} \left(\mathbf{S}_{\text{new}}(\mathbf{x}_i, S(\mathbf{x}_i)) - \mathbf{S}_{\text{new}}(\mathbf{x}_i, \mathbf{x}_\alpha) \right). \tag{16}$$

On the other hand, as shown in Figure 4, a new exemplar $\mathbf{x}_\alpha$ should be considered. Whether to accept the new exemplar is decided by the energy reduction $R2$, which will be discussed next. First, we assume a new exemplar $\mathbf{x}_\alpha$ is accepted. In fact, the following procedure is similar to that shown in Figure 3. Specially, the remaining data would change its exemplar to either $\mathbf{x}_\alpha$ or $S(\mathbf{x}_i)$. For data in cluster $e \in E$, theoretical analysis proves that only when the exemplar $\mathbf{x}_e$ changes its exemplar, $\forall \mathbf{x}_i \in \mathbf{X}_e$ would change its exemplar as $S(\mathbf{x}_i)$. In this case, the energy reduction is defined as follows:

$$R2_e = \sum_{\mathbf{x}_i \in \mathbf{X}_e} \left(\mathbf{S}_{\text{new}}(\mathbf{x}_i, S(\mathbf{x}_i)) - \mathbf{S}_{\text{new}}(\mathbf{x}_i, \mathbf{x}_e) \right). \tag{17}$$

Otherwise, some data in cluster e may change their exemplar as $\mathbf{x}_\alpha$; we define these data as $\mathbf{X}_{e,\alpha}^{/e}$, and the corresponding energy reduction $R2_e$ is defined in the following equation:

$$R2_\alpha = \sum_{\mathbf{x}_i \in \mathbf{X}_{e,\alpha}^{/e}} \left(\mathbf{S}_{\text{new}}(\mathbf{x}_i, \mathbf{x}_\alpha) - \mathbf{S}_{\text{new}}(\mathbf{x}_i, \mathbf{x}_e) \right). \tag{18}$$

Then, the energy reduction $R2$ is defined as follows:
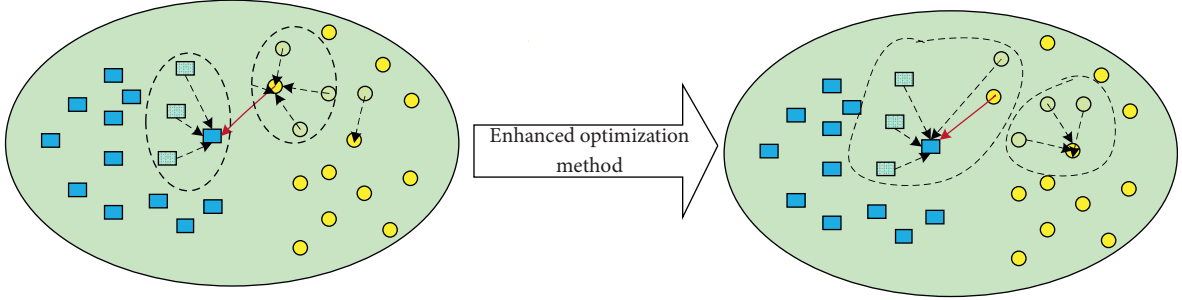
FIGURE 3: Case (i): $\mathbf{x}_\alpha$ is an exemplar.
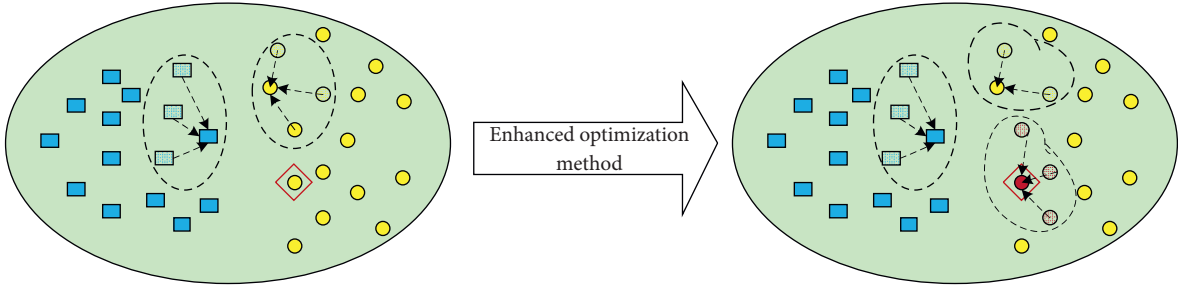
FIGURE 4: Case (ii): $\mathbf{x}_\alpha$ is not an exemplar.

$$R2 = \mathbf{S}_{\text{new}}\left(\mathbf{x}_\alpha, \mathbf{x}_\alpha\right) - \mathbf{S}_{\text{new}}\left(\mathbf{x}_\alpha, \mathbf{x}_{E_{\text{new}}(\alpha)}\right) + \max \sum_{e \in E} \{R2_e, R2_\alpha\}. \tag{19}$$

In sum, the new target function equation (12) is optimized, and the optimal exemplar list for the EEG signals is generated.

3.4. Time Complexity and Description.
The similarity relationship can be measured by Euclidean distance between data, defined as $d(\mathbf{x}_i, \mathbf{x}_j)$ in this study. The proposed algorithm FEEC consists of two stages, namely, compression stage and optimization stage. After compression, the scale of similarity matrix reduces from N^2 to Nc, so the optimization stage has the time complexity of $O(c^2)$. Therefore, the complexity of FEEC is considerably promising.

Based on the theoretical analysis above, the proposed FEEC for incomplete data can be summarized as Algorithm 1.

4. Experimental Study

To comprehensively evaluate the proposed algorithm FEEC, we have conducted several experiments based on both synthetic and real datasets. We also compare our new model with basic exemplar-based clustering model, namely, AP and EEM; to show these experimental results, we choose four performance indices in this section. In our experiments, all the algorithms were implemented using 2010a Matlab on a PC with 64 bit Microsoft Windows 10, an Intel(R) Core(TM) i7-4712MQ, and 8 GB memory.

4.1. Data Preparation.
We choose Aggregation [32], as shown in Figure 5, and the Bonn EEG signal datasets in this section. The Bonn dataset [9, 10] is from the University of Bonn, Germany (http://epileptologie-bonn.de/cms/upload/workgroup/lehnertz/eegdata.html). The EEG dataset contains five groups (A to E and each group contains 100 single channel EEG segments of 23.6s duration. The sampling rate of all the datasets was 173.6 Hz. Figure6 shows five healthy and epileptic EEG signals, and Table 2 lists detailed descriptions of these signals. Table 3 shows a brief description of these datasets. To construct the incomplete data scenario, we randomly choose 80% data as complete data and the remaining 20% as the incomplete data. We utilize KPCA to extract features from EEG signals in this section.

4.2. Performance Indices.
Here, we give the definitions of the three adopted performance indices ENERGY, NMI, and accuracy. Along with the description in [12, 16, 30, 33, 34], we call the result outputted by these involved models as cluster and the true labels as class.

4.2.1. ENERGY.
Since all the mentioned clustering algorithms are optimized, respectively, by the energy functions of the same type, we can compare them in terms of their energy values, defined as follows:

$$\text{ENERGY} = \sum_k \sum_i d\left(\mathbf{x}_k, \mathbf{x}_{k,i}\right), \tag{20}$$

where $\mathbf{x}_k$ denotes the kth exemplar, $\mathbf{x}_{k,i}$ is the ith data point in kth cluster, and $d(\mathbf{x}_k, \mathbf{x}_{k,i})$ is the Euclidean distance between $\mathbf{x}_k$ and $\mathbf{x}_{k,i}$ which can be seen as a measurement of energy.

Input: Given incomplete data $\mathbf{X} = [\mathbf{X}_c, \mathbf{X}_l] \in \mathbb{R}^{N \times D}, \mathbf{X}_c$
Output: Valid exemplar set E.
(1) **Compression**: Apply basic exemplar-based clustering algorithm to the complete data $\mathbf{X}_c$;
(2) Get the potential exemplar set E_{new} by equation (11) and construct the new similarity matrix $\mathbf{S}_{new} \in \mathbb{R}^{N \times c}$;
(3) Generate the expansion order o on the potential exemplar list E_{new}
(4) Let $t = 1$;
(5) **for** $e \in o$ **do**
(6)   **if** $e \in E$ **then**
(7)   compute $R1_\alpha, R1$ by equations (15) and (16);
(8)   **if** $R1_\alpha > 0$ **then**
(9)   for $\forall x_i \in \mathbf{X}_e$, set $E(x_i) = S(i)$;
(10)   **end**
(11)   **else**
(12)   compute $R2_e, R2_\alpha, R2$ by equations (17)–(19)
(13)   **if** $R2_\alpha > R2_e$ **then**
(14)   for $\forall x_i \in \mathbf{X}_{e,\alpha}^{/e}$, set $E(x_i) = \alpha$
(15)   **else**
(16)   for $\forall x_i \in \mathbf{X}_e$, set $E(x_i) = S(i)$
(17)   **else**
(18)   **if** $R2 > 0$ **then**
(19)   Accept the new exemplar α
(20)   **end**
(21)   **end**
(22) $t = t + 1$
(23) **end**
(24) Until converge.

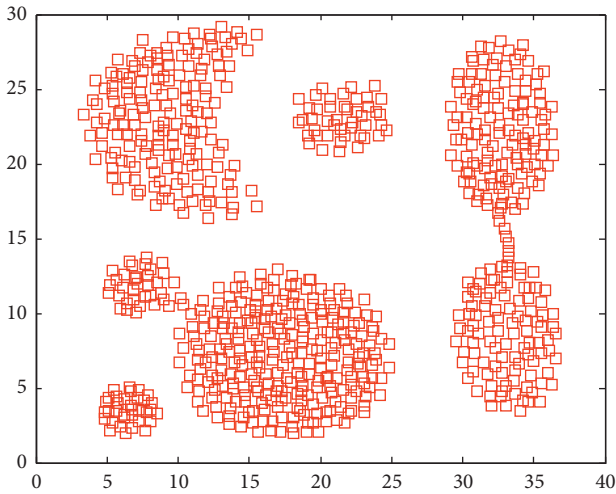ALGORITHM 1: Fast enhanced exemplar-based clustering algorithm.

FIGURE 5: Description of Aggregation data.

4.2.2. NMI. NMI has been widely used to evaluate the clustering quality as well, and its value can be calculated by the following equation:

$$\text{NMI} = \frac{\sum_{i=1}^{|C|} \sum_{j=1}^{|C|} N_{i,j} \ln\left(N N_{i,j} / N_i N_j\right)}{\sqrt{\left(\sum_{i=1}^{|C|} N_i \ln\left(N_i / N\right)\right)\left(\sum_{j=1}^{|C|} N_j \ln\left(N_j / N\right)\right)}}, \quad (21)$$

where $N_{i,j}$ is how clusters fit the classes, N_i is the number of data points in ith cluster, N_j is the number of data in jth class, and N is the total number of data points.

4.2.3. Accuracy. Accuracy is a more direct measure to reflect the effectiveness of clustering algorithms, which is defined as

$$\text{accuracy} = \frac{\sum_{i=1}^{N} \delta\left(c_i, \text{map}\left(\hat{c}_i\right)\right)}{N}, \quad (22)$$

where c_i is the real label of data points and $\hat{c}_i$ is the obtained clustering label. $\delta(i, j) = 1$, if $i = j$; $\delta(i, j) = 0$, otherwise. Function $\text{map}(\cdot)$ maps each obtained cluster to real class, and the optimized mapping function can be found in Hungarian algorithm.

The values of NMI and Acc range from 0 to 1, and the more it is close to 1, the more effective the clustering algorithm is. What is worth to mention is that we put % in the following relevant tables to show better precision. As to the performance index ENERGY, the smaller the value is, the better the clustering algorithm is.

4.3. Experimental Results and Discussion. The parameters involved FAP, AP, and EEM are in line with [13, 17, 29]. The preference $s(i, i)$ is set to be the median value of similarities between data. We run each algorithm over 10 runs under same parameters; the average results are shown in Table 4. Moreover, the detailed comparison in terms of the above 3 terms, NMI, accuracy, and ENERGY, are shown in Figures 7–12 and Table 4, respectively.

By analyzing Figures 7–11 and Table 4 in detail, we can conclude the following:
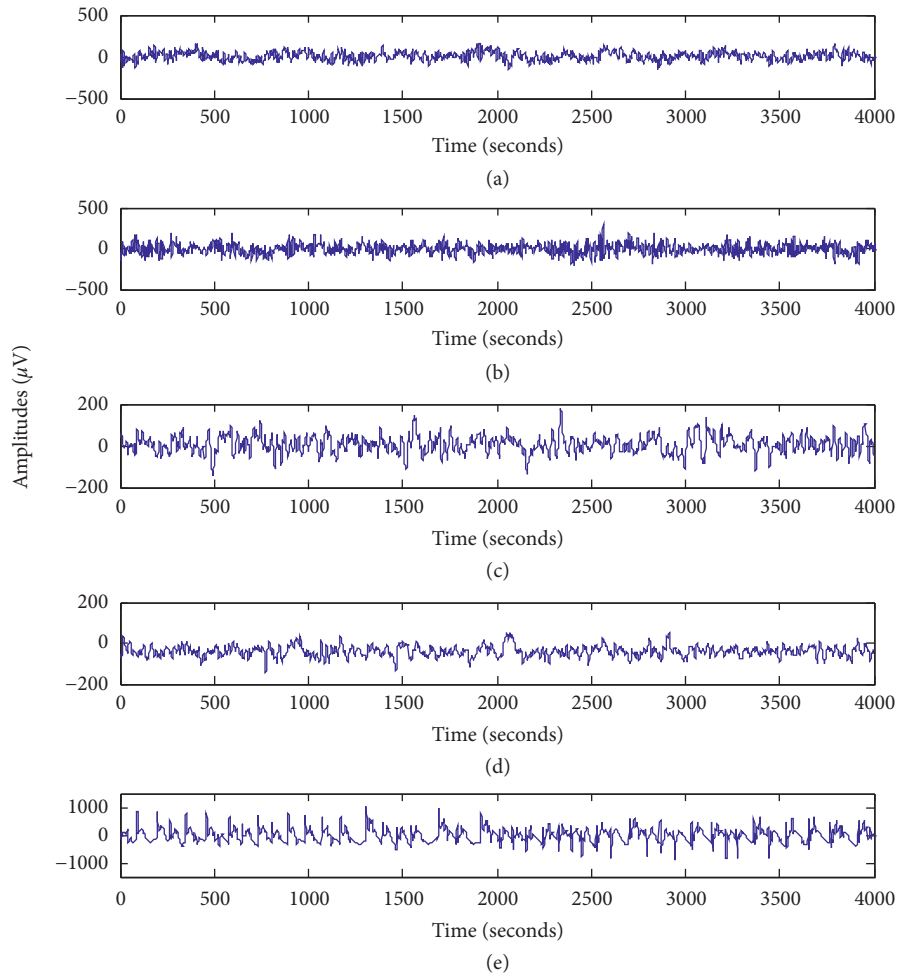
Figure 6: Typical EEG signals in groups A–E.

Table 2: Description of healthy and epileptic EEG data.

Subjects	Groups	Size of data	Descriptions
Healthy	A	100	Signals captured from volunteers with eyes open
	B	100	Signals captured from volunteers with eyes closed
Epileptic	C	100	Signals captured from volunteers during seizure silence intervals
	D	100	Signals captured from volunteers during seizure silence intervals
	E	100	Signals captured from volunteers during seizure activity

Table 3: Brief description of Aggregation and Bonn Epileptic EEG signals.

Datasets	Number of objects	Number of attributes	Number of clusters
Aggregation	788	2	7
Bonn	500	6	5

(1) The proposed algorithm FEEC can cluster data with 80% complete data and 20% incomplete data, and in most cases, the performance is very convincing. Specifically, for both Aggregation and epileptic EEG signals,

FEEC performs best, in terms of NMI, accuracy, and ENERGY.

(2) As to the computational time, compared with EEM, FEEC takes less time. Thus, with the assistance of

TABLE 4: Average experimental results over 10 runs on Aggregation and Bonn epileptic EEG signals.

Dataset	Algorithms	NMI	Accuracy	ENERGY	Time
Aggregation	FEEC	0.9636	0.9530	3522.38	9.54
	FAP	0.9017	0.9130	3609.79	9.65
	AP	0.8455	0.7016	4122.64	28.59
	EEM	0.8765	0.7208	3838.29	16.804
Bonn epileptic EEG signals	FEEC	0.9038	0.9786	1602.87	3.75
	FAP	0.8794	0.9445	1744.22	3.65
	AP	0.3324	0.4387	3018.12	3.97
	EEM	0.3728	0.4991	2050.51	6.01
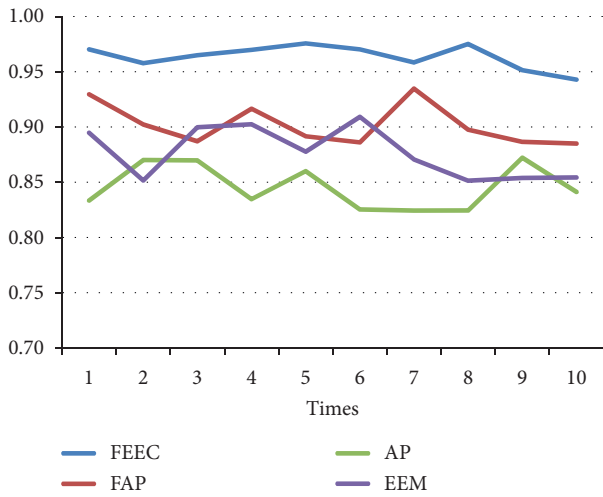
FIGURE 7: Comparison of NMI on Aggregation dataset.
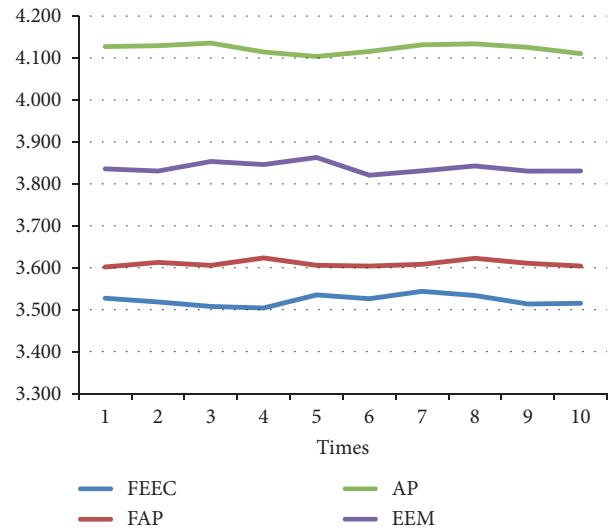
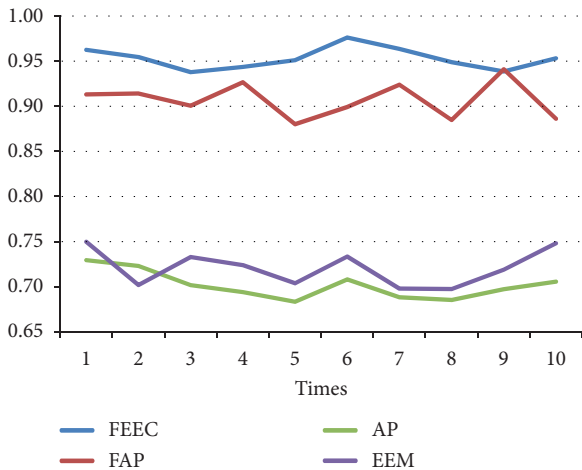FIGURE 9: Comparison of accuracy on Aggregation.

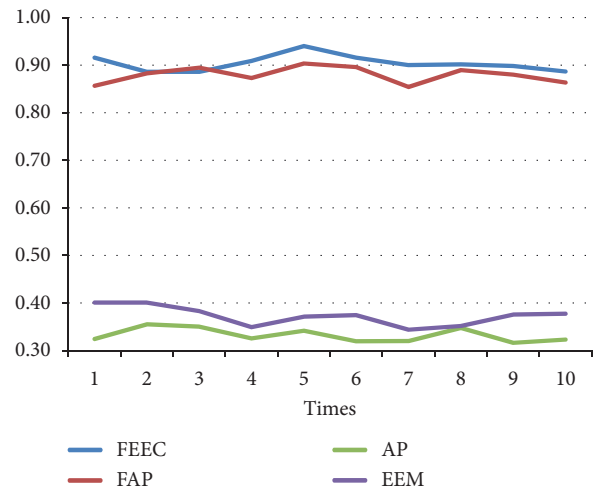FIGURE 8: Comparison of accuracy on Aggregation dataset.

FIGURE 10: Comparison of NMI on Bonn epileptic EEG signals.

compression stage, the time complexity of FEEC has been reduced, and the efficiency is improved as well. And FEEC has an equivalent computational time with FAP. Comparing other criteria of FEEC and FAP, it is worthwhile to spend more time.

(3) The proposed FEEC needs no more parameters except for preferences, while the performance of FAP relies much on k, which determines the number of nearest exemplars. Accordingly, in terms of the involved datasets in this section, FEEC would achieve satisfactory clustering results.
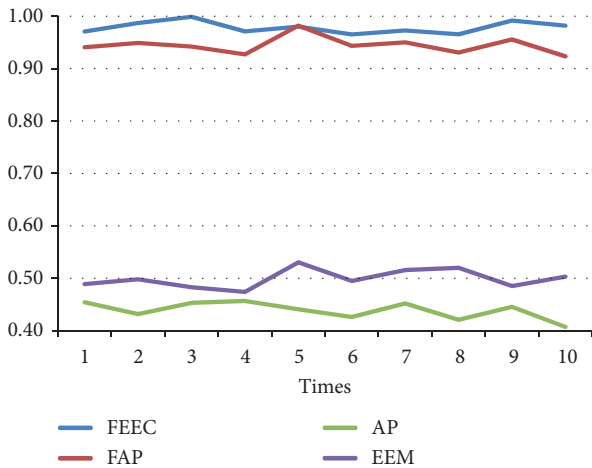
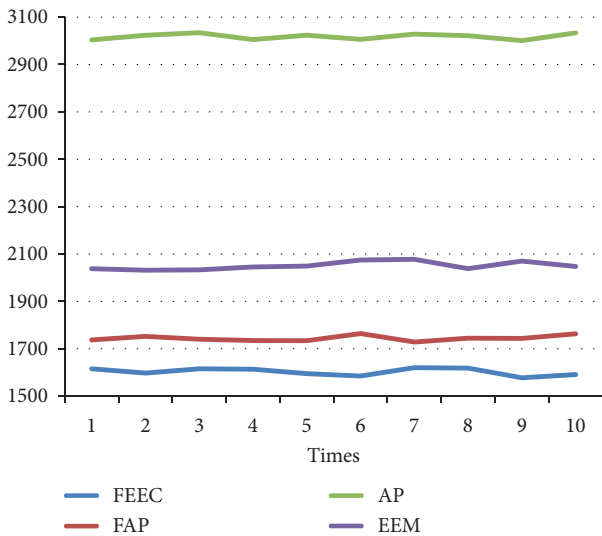FIGURE 11: Comparison of accuracy on Bonn epileptic EEG signals.

FIGURE 12: Comparison of ENERGY on Bonn epileptic EEG signals.

5. Conclusions

The diagnosis and treatment of epilepsy is always a significant direction for both machine learning and brain science. This paper newly proposes a fast exemplar-based clustering method for incomplete EEG signal. The FEEC method includes two stages, namely, compression and optimization. The performance of the proposed clustering algorithm is comprehensively verified by the experiments on two datasets.

Although most recognition methods of epilepsy are based on EEG signals at present, researchers also have to study on other neuroimaging modalities, such as cortical electroencephalography (ECoG), functional infrared optical imaging (fNIR), functional magnetic resonance imaging (fMRI), positron emission tomography (PET), and magnetoencephalography (MEG). Considering the fact that the brain activity is a nonlinear, networked, and unstable

complex system, we would focus on the multimodal clustering model for these neuroimaging modality signals in future.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

References

[1] H. Gao, H. Miao, L. Liu, J. Kai, and K. Zhao, "Automated quantitative verification for service-based system design: a visualization transform tool perspective," *International Journal of Software Engineering and Knowledge Engineering*, vol. 28, no. 10, pp. 1369–1397, 2018.

[2] H. Gao, W. Huang, and X. Yang, "Applying probabilistic model checking to path planning in an intelligent transportation system using mobility trajectories and their statistical data," *Intelligent Automation and Soft Computing*, vol. 25, no. 3, Article ID 5478C559, 2019.

[3] H. Gao, W. Huang, X. Yang, Y. Duan, and Y. Yin, "Toward service selection for workflow reconfiguration: an interface-based computing solution," *Future Generation Computer Systems*, vol. 87, pp. 298–311, 2018.

[4] H. Gao, W. Huang, Y. Duan et al., "Research on cost-driven services composition in an uncertain environment," *Journal of Internet Technology*, vol. 20, no. 3, pp. 755–769, 2019.

[5] L. Wang, K. C. L. Wong, H. Zhang, H. Liu, and P. Shi, "Noninvasive computational imaging of cardiac electrophysiology for 3-D infarct," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 4, pp. 1033–1043, 2011.

[6] Y. Xia, H. Zhang, L. Xu et al., "An automatic cardiac arrhythmia classification system with wearable electrocardiogram," *IEEE Access*, vol. 6, pp. 16529–16538, 2018.

[7] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[8] Y. Peng and B. L. Lu, "Immune clonal algorithm based feature selection for epileptic EEG signal classification," in *Proceedings of the 2012 11th International Conference on Information Science, Signal Processing and Their Applications (ISSPA)*, Montreal, Quebec, Canada, July 2012.

[9] Y. Jiang, D. Wu, Z. Deng et al., "Seizure classification from EEG signals using transfer learning, semi-supervised learning and TSK fuzzy system," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 12, pp. 2270–2284, 2017.

[10] Z. Jiang, F.-L. Chung, and S. Wang, "Recognition of multiclass epileptic EEG signals based on knowledge and label space

inductive transfer," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 4, pp. 630–642, 2019.

[11] Z. Iscan, Z. Dokur, and T. Demiralp, "Classification of electroencephalogram signals with combined time and frequency features," *Expert Systems with Applications*, vol. 38, no. 8, pp. 10499–10505, 2011.

[12] Y. Jiang, F.-. Chung, S. Wang, Z. Deng, J. Wang, and P Qian, "Collaborative fuzzy clustering from multiple weighted views," *IEEE Transactions on Cybernetics*, vol. 45, no. 4, pp. 688–701, 2015.

[13] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, 2017.

[14] V. Kolmogorov and C. Rother, "Comparison of energy minimization algorithms for highly connected graphs," in *Proceedings of the 9th European Conference on Computer Vision (ECCV)*, vol. 3952, Graz, Austria, May 2006.

[15] M. F. Tappen and W. T. Freeman, "Comparison of graph cuts with belief propagation for stereo, using identical MRF parameters," in *Proceedings of the Ninth IEEE International Conference on Computer Vision*, pp. 900–906, Nice, France, October 2003.

[16] A. Bi and S. Wang, "Incremental enhanced α-expansion move for large data: a probability regularization perspective," *International Journal of Machine Learning and Cybernetics*, vol. 8, no. 5, pp. 1615–1631, 2016.

[17] Y. Zheng and P. Chen, "Clustering based on enhanced α-expansion move," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 10, pp. 2206–2216, 2013.

[18] B. R. Greene, S. Faul, W. Marnane, G. Lightbody, I. Korotchikova, and G. Boylan, "A comparison of quantitative EEG features for neonatal seizure detection," *Clinical Neurophysiology*, vol. 119, no. 6, pp. 1248–1261, 2008.

[19] C. W. N. F. C. W. Fadzal, W. Mansor, L. Y. Khuan, and A. Zabidi, "Short-time fourier transform analysis of EEG signal from writing," in *Proceedings of the 2012 IEEE 8th International Colloquium on Signal Processing and its Applications*, Malacca, Malaysia, March 2012.

[20] E. A. Vivaldi and A. Bassi, "Frequency domain analysis of sleep EEG for visualization and automated state detection," in *Proceedings of the 2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, New York, NY, USA, September 2006.

[21] D. Hu, W. Li, and X. Chen, "Feature extraction of motor imagery eeg signals based on wavelet packet decomposition," in *Proceedings of the 2011 IEEE/ICME International Conference on Complex Medical Engineering*, pp. 694–697, Harbin, China, May 2011.

[22] Z. Zhang, H. Kawabata, and Z. Q. Liu, "EEG analysis using fast wavelet transform," in *Proceedings of the SMC 2000 Conference Proceedings: 2000 IEEE International Conference on Systems, Man and Cybernetics*, Nashville, TN, USA, October 2000.

[23] K. Murphy, Y. Weiss, and M. Jordan, "Loopy belief propagation for approximate inference: an empirical study," in *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, Bellevue, WA, USA, August 2013.

[24] K. Wang, J. Zhang, D. Li, X. Zhang, and T. Guo, Adaptive affinity propagation clustering," 2008, https://arxiv.org/abs/0805.1096.

[25] J. Vlasblom and S. J. Wodak, "Markov clustering versus affinity propagation for the partitioning of protein interaction graphs," vol. 10, no. 1, pp. 99-100, 2009.

[26] X. Mishra, R. Guan, L. Wang, Z. Pei, and Y. Liang, "An incremental affinity propagation algorithm and its applications for text clustering," in *Proceedings of the 2009 International Joint Conference on Neural Networks*, pp. 2914–2919, Atlanta, GA, USA, June 2009.

[27] L. Sun and C. Guo, "Incremental affinity propagation clustering based on message passing," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 11, pp. 2731–2744, 2014.

[28] I. Givoni and B. Frey, "Semi-supervised affinity propagation with instance-level constraints," *Journal of Machine Learning Research—Proceedings Track*, vol. 5, pp. 161–168, 2009.

[29] L. Sun, C. Guo, C. Liu, and H. Xiong, "Fast affinity propagation clustering based on incomplete similarity matrix," *Knowledge and Information Systems*, vol. 51, no. 3, pp. 941–963, 2017.

[30] A. Bi, F. Chung, S. Wang, Y. Jiang, and C. Huang, "Bayesian enhanced α-expansion move clustering with loose link constraints," *Neurocomputing*, vol. 194, pp. 288–300, 2016.

[31] S. Guha, A. Meyerson, N. Mishra, R. Motwani, and L. O'Callaghan, "Clustering data streams: theory and practice," *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 3, pp. 515–528, 2003.

[32] A. Gionis, H. Mannila, and P. Tsaparas, "Clustering aggregation," *ACM Transactions on Knowledge Discovery from Data*, vol. 1, no. 1, p. 4, 2007.

[33] Y. Jiang, Z. Deng, F.-L. Chung et al., "Recognition of epileptic EEG signals using a novel multiview TSK fuzzy system," *IEEE Transactions on Fuzzy Systems*, vol. 25, no. 1, pp. 3–20, 2017.

[34] Y. Jiang, F.-L. Chung, H. Ishibuchi, Z. Deng, and S. Wang, "Multitask TSK fuzzy system modeling by mining intertask common hidden structure," *IEEE Transactions on Cybernetics*, vol. 45, no. 3, pp. 548–561, 2015.