EEG-Based Emotion Recognition with Prototype-Based Data Representation

Yixin Wang[1,2], Shuang Qiu[1], Chen Zhao[4], Weijie Yang[4], Jinpeng Li[1,2], Xuelin Ma[1,2], Huiguang He[1,2,3,*]

***Abstract*— Emotions play an important role in human communication, and EEG signals are widely used for emotion recognition. Despite the extensive research of EEG in recent year, it is still challenging to interpret EEG signals effectively due to the massive noises in EEG signals. In this paper, we propose an effective emotion recognition framework, which contains two main parts: the representation network and the prototype selection algorithm. Through our proposed representation network, samples from the same kind of emotion state are more close to each other in high-level representation, and then, we selected the prototypes from the clustering set in feature space match the following testing samples. This method takes advantage of the powerful representation ability of deep learning and learns a better describable feature space rather than learn the classifier explicitly. The experiments on SEED dataset achieves a high accuracy of 93.29% and outperforms a set of baseline methods and the recent deep learning emotion classification approaches. These experimental results demonstrate the effectiveness of our proposed emotion recognition framework.**

I. INTRODUCTION

Emotions play a vital part in people's daily life, and their effect involves many aspects, such as human interaction, decision-making, perception of the world around us and so on [1]. Recently, interest was shown in making emotional connection between the human-being and the computer. The field of Affective Computing (AC) [2] has emerged to fill the gap between the high-level emotional signals and the low-level features of raw digital data. Emotion recognition is the primary process of affective computing, including detection and model the human emotional state.

There are various approaches to identify an affective state. On the one hand, several of them are non-physiological, like facial expressions [3], voice [4], body language and posture [5]. On the other hand, there are some physiological measurements which can catch the participants' underlying responses expressed at the time of stimulation [6]. They can also be separated into two parts according to the signals' sources, one is from the autonomic nervous systems, such as Galvanic Skin Response (GSR), Electromyography (EMG), Heart Rate (HR), and Respiration Rate (RR). Other is from the central nervous systems, for instance, Electroencephalography (EEG), functional Magnetic Resonance Imaging (fMRI), and etc [7]. Although EEG has a poor spatial resolution and need many electrodes placed at the specified location on the head, it provides great time resolution. The usage of EEG is noninvasive, fast, and inexpensive, making it a preferred method in detecting and evaluating emotion [8].

In the field of EEG-based emotion recognition, the delta, theta, alpha, beta and gamma bands are always being mentioned [9] [10], and then the features are extracted from these bands to use in the classification process. The most used methods were the Fourier Transform, such as the Short-time Fourier Transform (STFT), Power Spectral Density (PSD) [11]. Also the entropy method such as Differential Entropy (DE) [12] was widely used. However, EEG signals are very complex non-stationary, because of the variation of users' neural activity and the random noise during recording. Emotion recognition based on EEG is very challenging and remains many meaningful problems in different levels.

There are a large number of classifiers' families that are commonly used: bayesian, support vector machines, decision trees [8]. For example, Zheng and Lu [13] proposed selecting 12 channel electrodes DE features in SVM for positive, negative and neutral 3-classes problem, which provided 86.65% on average. Zheng proposed a discriminative Graph regularized Extreme Learning Machine (GELM) based on the idea that similar samples should share similar properties, which obtained the state-of-art accuracy of 91.07% [14]. With the help of the powerful deep learning [15] method, the explorations in emotional EEG recognition has been spread. From using the fundamental network (including deep belief networks (DBNs) [13], convolutional neural networks

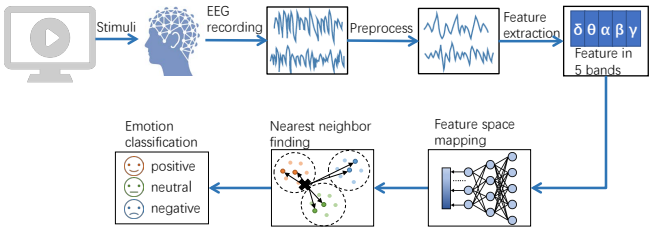

Fig. 1. The flow chart of EEG emotion classification with prototype-based data representation.

*This work was supported in part by the National Natural Science Foundation of China under Grant 81701785, 91520202, in part by the Strategic Priority Research Program of CAS, Grant No.XDB32040000, in part by the CAS Scientific Equipment Development Project under Grant YJKYYQ20170050, in part by the Beijing Municipal Science and Technology Commission under Grant Z181100008918010 and in part by the Youth Innovation Promotion Association CAS.

[1] Research Center for Brain-inspired Intelligence, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Science, Beijing, China

[2] University of Chinese Academy of Sciences, Beijing, China

[3] Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Science, Beijing, China

[4] School of Computer and Information Engineering, Beijing Technology and Business University, Beijing, China

* Corresponding author huiguang.he@ia.ac.cn

978-1-5386-1311-5/19/$31.00 ©2019 IEEE

(CNNs) [16] and recurrent neural networks (RNNs) [17]) to decoding the inner construction of emotional EEG signals, the deep learning has been becoming a promising orientation in our field. For instance, Li organized DE features extracted from 62 channels as 2-D maps to train the hierarchical convolutional neural network (HCNN) and achieved 88.2% at the Gamma wave band in three classes problem [18]. Zhang used a spatial-temporal RNN (STRNN) to integrate the feature learning from both spatial and temporal information, which can get the accuracy of 89.5% with DE features [19].

In this paper, we propose a brand-new method to solve the EEG recognition from representation learning point of view. Our method is composed of two main parts in the training stage. One is the representation network based on deep neural networks (DNNs), which make the distance minimum for same emotion samples, maximum for different emotion samples in feature space by weak supervision. The other is the prototype selection algorithm, intended to find a representative subset of the training set. In testing stage, we map the original testing samples to our feature space, and the prediction for test samples are the same category as thier the nearest prototypes. This method combines powerful representation ability of deep learning with the traditional metric learning. It's also a novel attempt to learn a better describable distribution rather than learn the classifier explicitly. We show the entire flow chart in the Fig 1.

In Section 2, we will introduce the preliminary knowledge and model construction. Section 3 presents experiment settings, data pre-processing, feature extraction and classifiers training details. Following the experiment results and discussion are described in Section 4. Conclusions and future work are in Section 5. The experiment result demonstrates that our method can significantly improve the emotion classification performance on a benchmark dataset—SEED.

II. THE PROPOSED METHOD

The framework of our method is shown in Fig 2. More details are demonstrated as follows.

A. Problem Statement

Given the training set $\{(x_i, y_i)\}_{i=1}^{N}$, where $x_i \in \mathbb{R}^D, y_i \in \{0, 1, ..., c-1\}$. Though the category labels are supplied, we only use binary labels between pairwise samples to learn the representation network. The goal of the representation network is to obtain a representative feature embedding space $f(x_i; \Theta) \in \mathbb{R}^M$, where Θ is the parameter of the model.

Let $f(\boldsymbol{X}) = \begin{bmatrix} f(x_1)^{\mathrm{T}} \\ f(x_2)^{\mathrm{T}} \\ \cdots \\ f(x_N)^{\mathrm{T}} \end{bmatrix} \in \mathbb{R}^{N \times M}$ be the representation data matrix of N training samples in $\mathbb{R}^M$, and the pairwise dissimilarity matrix $\boldsymbol{D} = \{d_{ij}\}_{i=1,...,N}^{j=1,...,N}$ can be calculated by some predefined metric function between each samples. In our proposed method, we choose Euclidean distance as the measure criterion, i.e., $d_{ij} = \|f(x_i) - f(x_j)\|_2^2$. Given dissimilarity matrix $\boldsymbol{D}$, we focus on the problem of selecting a few data points in the training set, called prototypes, that can describe all the samples.

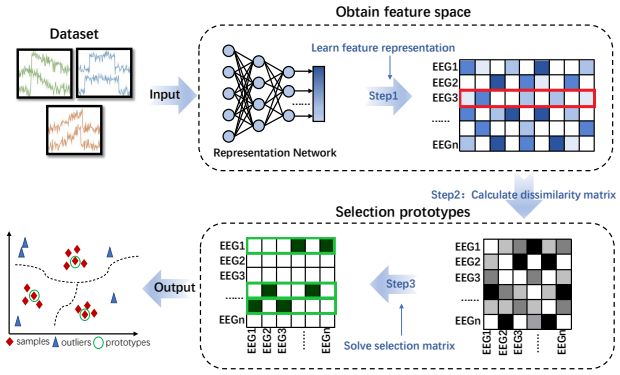

Fig. 2. The training stage framework of our proposed method.

Using the effective prototype selection method—Dissimilarity-based Sparse Subset Section (DS3) algorithm [20], we assume series of unknown variables z_{ij} connected with dissimilarities values d_{ij}, where $z_{ij} \in \{0, 1\}$. When the value is one, we select x_i as the representation of x_j, and is zero otherwise.

After getting the selection matrix, we can use these prototypes to match the testing set for classification. The prototypes are expressed as p_{ij}, where $i \in \{1, 2, \ldots, C\}$, means the index of classes and $j \in \{1, 2, \ldots, K_i\}$ means the index of prototypes in each class. Note that we can not ensure the number of prototypes in each class is equivalent.

In the training stage, we learn the deep dissimilarity metric learning network and the selection indicator, finally output the feature mapping $f(\cdot; \Theta)$ and the prototypes $\{p_{ij}\}$. In the testing stage, we map the original testing samples to the cluster feature space and find the nearest prototype according to Euclidean distance. The testing samples belong to the same category as their matching prototypes.

B. Problem Formulation

To select the most discriminative subset, we take advantage of the promising deep representation learning approach at first. And then, we solve the selection matrix $\boldsymbol{Z}$ as a row-sparsity regularized trace minimization problem. Combined with two key points, we come up with the following optimization goal:

$$\min_{\Theta, \{z_{ij}\}} \sum_{j=1}^{N} \sum_{i=1}^{N} \Phi_{dSim}(f(x_i; \Theta) f(x_j; \Theta)) z_{ij} + \lambda \sum_{i=1}^{N} \|\boldsymbol{z_i}\|_p$$
$$s.t. \qquad \sum_{i=1}^{N} z_{ij} = 1, \forall j; \;\; z_{ij} \geqslant 0, \forall i, j. \tag{1}$$

where the first term in the objective function is the total cost of representing $\boldsymbol{X}$ via prototypes, and the second term is to limit the number of prototypes. The regularization parameter λ sets the trade-off between the two terms. For smaller λ, we are focused on how to represent the original training set

better and get more prototypes. In extreme cases, every point becomes the prototype of itself. For larger λ, we put more emphasis on the row sparsity of $\boldsymbol{Z}$ and get fewer prototypes.

As shown in (1), $f(\cdot;\Theta)$ denote a mapping from high-dimensional observations to a low-dimensional feature space, and Θ is the set of parameters. $\Phi_{dSim}(f(x_i;\Theta), f(x_j;\Theta))$ indicates how well $f(x_i;\Theta)$ represents $f(x_j;\Theta)$. $\boldsymbol{Z} = \{z_{ij}\}_{i=1,\ldots,N}^{j=1,\ldots,N}$ is a selection matrix. We use the sum of l_p-norms of rows of $\boldsymbol{Z}$ instead of counting the number of nonzero rows of $\boldsymbol{Z}$. To ensure that each x_j is represented by one prototype, we must constrain $\sum_{i=1}^{N} z_{ij} = 1$. In addition, we give the relaxation condition $z_{ij} \in [0,1]$, and we set $p = \infty$, there we can typically consider that $\{z_{ij}\}$ are in $\{0,1\}$.

Once we solve the selection matrix, we can find the indices of nonzero rows of the solution $\boldsymbol{Z}$, and these can be regarded as indices of the chosen prototypes in the whole training set. The details about how to optimize Θ and $\boldsymbol{Z}$ are introduced in the following sections.

C. Representation Learning

In this section, we introduce a deep metric learning algorithm borrowed from Song [21], which can measure similarities between pairs of EEG samples and change their distribution. The main idea is to minimize the distance in the feature space for similar pairs, and maximize the distance for dissimilar pairs with the help of weak supervision.

Given a batch of D-dimensional embedding features $\boldsymbol{X} \in \mathbb{R}^{n\times D}$, and the column vector of squared norm of individual batch elements $\tilde{\boldsymbol{x}} = [\|f(\boldsymbol{x}_1)\|_2^2, \ldots, \|f(\boldsymbol{x}_n)\|_2^2]^{\mathrm{T}}$. We can efficiently compute the distance matrix by $\boldsymbol{D}^2 = \tilde{\boldsymbol{x}}\mathbf{1}^{\mathrm{T}} + \mathbf{1}\tilde{\boldsymbol{x}}^{\mathrm{T}} - 2\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$. The objective function of our representation network is defined as follows:

$$\begin{aligned}\tilde{J}_{i,j} = log(&\sum_{(i,k)\in\mathcal{N}} exp\{\alpha - D_{i,k}\} + \sum_{(j,l)\in\mathcal{N}} exp\{\alpha - D_{j,l}\}) \\ &+D_{i,j} \\ \tilde{J} = &\frac{1}{2\,|\mathcal{P}|}\sum_{(i,j)\in\mathcal{P}} max(0,\tilde{J}_{i,j})^2\end{aligned} \tag{2}$$

where $\mathcal{N}$ denotes the set of pairs of examples with different class labels, and $\mathcal{P}$ denotes the same class. α denotes a fixed margin constant. And the parameters of the representation network are optimized by mini-batch adaptive moment estimation algorithm:

$$\min_{\Theta} \tilde{J}(\boldsymbol{D}(f(\boldsymbol{X},\boldsymbol{Y};\Theta))) \tag{3}$$

D. Dissimilarity-based prototype selection

We can rewrite the optimization program (1) in the matrix form as:

$$\begin{aligned}&\min_{\boldsymbol{Z}} \; tr(\boldsymbol{D}^{\mathrm{T}}\boldsymbol{Z}) + \lambda\|\boldsymbol{Z}\|_{1,p} \\ &s.t. \qquad \mathbf{1}^{\mathrm{T}}\boldsymbol{Z} = \mathbf{1}^{\mathrm{T}},\; \boldsymbol{Z} \geqslant \mathbf{0}\end{aligned} \tag{4}$$

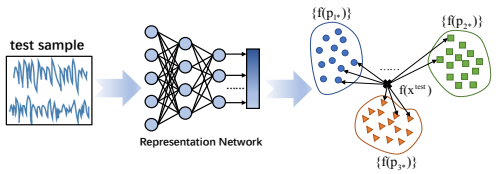

Fig. 3. The testing stage framework of our proposed method.

where $\sum_{i,j=1}^{N} d_{ij}z_{ij} = tr(\boldsymbol{D}^{\mathrm{T}}\boldsymbol{Z})$, $tr(\cdot)$ denotes the trace operator, $\mathbf{1} \in \mathbb{R}^N$ is a column vector with all 1.

The problem described in (4) can be solved by the Alternating Direction Method of Multipliers (ADMM) framework [22]. First, we introduce an auxiliary matrix $\boldsymbol{C} \in \mathbb{R}^{N\times N}$ and consider the following optimized objective

$$\begin{aligned}&\min_{\boldsymbol{Z},\boldsymbol{C}} \; tr(\boldsymbol{D}^{\mathrm{T}}\boldsymbol{C}) + \lambda\|\boldsymbol{Z}\|_{1,p} + \frac{\mu}{2}\|\boldsymbol{Z}-\boldsymbol{C}\|_F^2 \\ &s.t. \qquad \mathbf{1}^{\mathrm{T}}\boldsymbol{C} = \mathbf{1}^{\mathrm{T}},\; \boldsymbol{C} \geqslant \mathbf{0},\; \boldsymbol{Z} = \boldsymbol{C}\end{aligned} \tag{5}$$

where $\mu > 0$ is a parameter of penalty term. Since (1) is equivalent to (5), they have the same optimal solution for $\boldsymbol{Z}$. Then, we can augment the last equality constraint of (5) to the objective function by the Lagrange multiplier matrix $\boldsymbol{\Lambda} \in \mathbb{R}^{N\times N}$.

$$\begin{aligned}\mathcal{L} &= \lambda\|\boldsymbol{Z}\|_{1,p} + \frac{\mu}{2}\left\|\boldsymbol{Z} - (\boldsymbol{C} - \frac{\boldsymbol{\Lambda}}{\mu})\right\|_F^2 + h_1(\boldsymbol{C},\boldsymbol{\Lambda}) \\ &= \sum_{i=1}^{N}(\lambda\|\boldsymbol{Z}_{i*}\|_q + \frac{\mu}{2}\left\|\boldsymbol{Z}_{i*} - (\boldsymbol{C}_{i*} - \frac{\boldsymbol{\Lambda}_{i*}}{\mu})\right\|_2^2) + h_1(\boldsymbol{C},\boldsymbol{\Lambda})\end{aligned} \tag{6}$$

where $\boldsymbol{\Lambda}_{i*}$ denotes the i-th row of the matrix $\boldsymbol{\Lambda}$ and the term $h_1(\cdot)$ does not depend on $\boldsymbol{Z}$, we can rewrite the function as

$$\begin{aligned}\mathcal{L} &= \frac{\mu}{2}\left\|\boldsymbol{C} - (\boldsymbol{Z} + \frac{\boldsymbol{\Lambda}+\boldsymbol{D}}{\mu})\right\|_F^2 + h_2(\boldsymbol{Z},\boldsymbol{\Lambda}) \\ &= \sum_{i=1}^{N}\frac{\mu}{2}\left\|\boldsymbol{C}_{i*} - (\boldsymbol{Z}_{i*} + \frac{\boldsymbol{\Lambda}_{i*}+\boldsymbol{D}_{*i}}{\mu}\right\|_2^2) + h_2(\boldsymbol{Z},\boldsymbol{\Lambda})\end{aligned} \tag{7}$$

where the term $h_2(\cdot)$ does not depend on $\boldsymbol{C}$. The steps of the ADMM implementation consist of 1) initializing $\boldsymbol{Z}$, $\boldsymbol{C}$ and $\boldsymbol{\Lambda}$; 2) minimizing $\mathcal{L}$ with respect to $\boldsymbol{Z}$ with other variables fixed; 3) minimizing $\mathcal{L}$ with respect to $\boldsymbol{C}$ subject to the constrains $\{\mathbf{1}^{\mathrm{T}}\boldsymbol{C} = \mathbf{1}^{\mathrm{T}},\; \boldsymbol{C} \geqslant \mathbf{0}\}$ with other variables fixed; 4) updating the Lagrange multiplier matrix $\boldsymbol{\Lambda}$ with other variables fixed.

E. Nearest neighbor classifier for prediction

Given a test sample x, we first get the feature representation via the representation network $f(x;\Theta)$, then we compute the distance between the extracted feature and all the selected prototypes p_{ij}, and the classification operation can be formulated as follows:

$$x \in class \qquad arg \min_{i=1}^{C} g_i(x) \tag{8}$$

where $g_i(x)$ is the discriminant function for class i:

$$g_i(x) = -\min_{j=1}^{K_i} \|f(x;\Theta) - f(p_{ij};\Theta)\|_2 \tag{9}$$

and the process of test is shown in Fig 3.

III. MATERIALS

A. Experiment settings

In our paper, we use the public dataset – the SJTU emotion EEG dataset (SEED) [13]. There are three kinds of emotion states in our stimulating film clips, including positive, neutral and negative. The number of film clips is fifteen, and the duration of each film clip is about 4 minutes. There is a 15s hint before each clips and 10s feedback after each clip. The order of presentation is arranged so that two film clips targeting the same emotion are not shown consecutively. For the feedback, participants are told to report their emotional reactions to each film clip by completing the questionnaire immediately after watching each clip. EEG signals of 15 subjects were recorded while they were watching the emotional film clips.

EEG signals are recorded by an ESI NeuroScan system at a sampling rate of 1000 Hz from 62-channel electrode cap according to the international 10-20 system.

B. Data preprocessed and Feature extraction

The data was initially down-sampled to 200Hz. We extracted the EEG segments corresponding to the duration of each movie. The EEG data were visually checked and the recordings seriously contaminated by electromyography (EMG) and Electrooculography (EOG) were removed manually from the dataset. To further filter the noises, a bandpass frequency filter from 0.3-75Hz was applied finally.

The EEG of each channel was divided into 1s segments without overlapping. The total number of the segment is about 3400. Features are extracted on each EEG segment. Zheng has demonstrated that DE features are more suitable for EEG-based emotion recognition than other traditional features, including PSD, DASM, RASM, ASM and DCAU [14]. Therefore, we choose DE to characterize the EEG segments mainly.

According to five frequency bands: delta (1-3Hz); theta (4-7Hz); alpha (8-13Hz); beta (14-30Hz); and gamma (31-50Hz), in each frequency band, DE is equivalent to the logarithmic power spectral density for a fixed length EEG sequence. The differential entropy feature is defined as follows

$$\begin{aligned} h(X) = -\int_{\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} exp\frac{(x-\mu)^2}{2\sigma^2} log\frac{1}{\sqrt{2\pi\sigma^2}} \\ exp\frac{(x-\mu)^2}{2\sigma^2} dx = \frac{1}{2} log 2\pi\sigma^2 \end{aligned} \tag{10}$$

TABLE I

COMPARISONS ON EEG-BASED EMOTION DATASET SEED

Method	Feature	Frequency bands	Channels number	Accuracy (%)
SVM [13]	DE	$\delta, \theta, \alpha, \beta, \gamma$	62	83.99
	DE	$\delta, \theta, \alpha, \beta, \gamma$	12	86.65
DBN [13]	DE	$\delta, \theta, \alpha, \beta, \gamma$	62	86.08
HCNN [18]	DE	γ	62	88.20
STRNN [19]	DE	$\delta, \theta, \alpha, \beta, \gamma$	62	89.51
BDAE [23]	DE eye movement	$\delta, \theta, \alpha, \beta, \gamma$	62	91.01
GELM [14]	DE	$\delta, \theta, \alpha, \beta, \gamma$	62	91.07
Ours	DE	$\delta, \theta, \alpha, \beta, \gamma$	62	**93.28**

where X submits the Gauss distribution $N(\mu, \sigma^2)$, x is a variable, π and e are constants. Since each frequency band signal has 62 channels, we extracted differential entropy features with 310 dimensions for a 1s sample.

IV. RESULTS AND DISCUSSION

In this section, we design series of contrast experiments and present the results of our approaches on the SEED dataset.

A. Classification performance

We first compare our result with those of various existed algorithms which are also used the SEED dataset, including SVM [13], DBN [13], HCNN [18], STRNN [19], BDAE [23] and so on. In our experiment, the training data contains the first 9 sessions of data while the test data contains other 6 sessions of data, which is a widely used protocol in previous studies [13] [18] [19] [23].

The performance of our proposed model and the comparison classifiers using the SEED dataset are presented in Table 1. Our model with the DE feature achieve a high accuracy of 93.29%, outperforming the state-of-the-art methods. This is probably because our proposed method learns a discriminative data representation and selects the prototypes to match the testing samples, while other comparison methods directly learn a classifier. Benefiting from that, our method gains the outperformance.

In order to demonstrate the effectiveness of our method, we also evaluate the performance under different dataset partitions. Except for the widely-used partition way above, we reduce the proportion of the training data to the testing data gradually and design four other partition ways: (1) using the first three-quarters of data in each session of the first 9 sessions as training data and the remaining 6 sessions as testing data, (2) using the first 6 sessions as training data and the following 3 sessions as testing data, (3) using the first 6 sessions as training data and the following 6 sessions as testing data, (4) using the first 6 sessions as training data and the remaining 9 sessions as testing data.

We choose K nearest neighbor (kNN), logistic regression (LR), support vector machine (SVM) as baselines on account of their universality in emotion recognition [8]. Fig. 4 shows the mean accuracies with 5 different dataset partition ways. A

TABLE II

RESULT VARIATIONS WITH THE NUMBER OF SELECTED PROTOTYPES

	Number of prototype	Average accuracy(%)
Random selection	3	85.16 ± 9.14
No selection	2010	87.53 ± 6.4
Ours	50∼100	**93.29** ± **5.8*****††

*** Ours significantly outperformed to random selection. ($p<0.001$)
†† Ours significantly outperformed to no selection. ($p<0.01$)

Two-Way Repeated-Measure Analysis of Variance (ANOVA) was performed to analyze the effect of both dataset partition ways and classification method on the accuracy.It showed significant main effects of classification methods ($F_{(2.64)} = 13.11$, $p<<0.001$), and dataset partition ways ($F_{(2.40)} = 4.62$, $p<0.05$). All of these dataset partition experiments demonstrate that our method with the use of fewer training data can still retain the outperformance. We can see that our method significantly outperforms other common methods in (2)-(4) partition conditions, which achieves 5.02% higher mean accuracy than SVM, 6.03% for LR and 12.31% for kNN. Comparing the first partition way 3/4×9→6 with the widely used way 9→6, our method emerges a 5.63% decline, however SVM has no obvious changes, only 0.1%. It may due to the removed 1/4 session may contain some significant information for our method.

Our method first takes advantage of the network's representative ability. To show the performance of representation network using visualization methods, we choose one typical subject's data, compute the Pearson correlation coefficient [24] between the different emotion state samples and plot the correlation matrices in Fig 5. In Fig 5(a), we can see that the correlation matrix of the original DE feature is disordered and can not distinguish the emotion state directly. After the mapping from DE feature to the representative feature space, these three kinds of emotions are very distinct as Fig 5(b) shows. Fig 5(a)→Fig 5(b) is constrained by weak supervision, while we directly use the trained network for Fig 5(c)→Fig 5(d). Then, we used this trained network for testing samples. Fig 5(c) shows the original DE feature of testing samples. The testing samples went through this trained represent network, and then is shown in Fig 5(d). It has a good display. It implies that our representation network

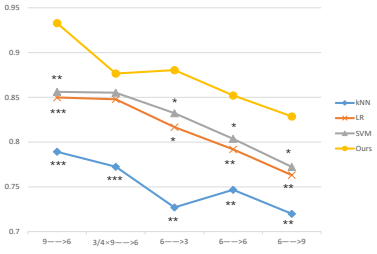

Fig. 4. The mean accuracies of different dataset partition in DE feature. p⟶q: train models on the first p sessions, and test on the following q sessions. (∗: $p<0.05$, ∗∗: $p<0.01$, ∗ ∗ ∗: $p<0.001$)

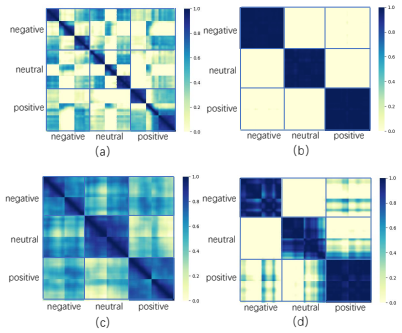

Fig. 5. Correlation matrices for one subject, which computed the self-correlations on three situations separately. (a) between the training data, (b) between the feature representation on training set, (c) between the testing data, (d) between the testing set via the trained representation network.

has extracted the emotion intrinsic information and had the primary ability to discriminate the EEG emotion signal.

To select the most representative subset as the prototypes is an important step in our model. We compare our model with two selection methods to demonstrate the effect of prototype selection. Firstly, we randomly select one sample from one class as the prototype of this class, which is called "Random selection". Thus, there are only three prototypes. Secondly we use all samples as prototypes, which is called "No selection", the number of prototypes is the same as that of the training set, which is 2010. Lastly, as for our method, we search the parameter λ, which is used to limit the number of prototypes in the range of $[0.1, 0.5, 1, 5]$ with a step of one. According to the parameter search for each subject, we can find that the number of selected prototypes, is range from 50 to 100 on average.

The results are presented in Table 2. A one-way Analysis of Variance (ANOVA) showed significant main effects of prototype selection methods ($F_{(3.22)} = 4.63$, $p<0.05$). Our prototype selection method is significantly outperforms random selection method ($p<0.001$) and no selection method ($p<0.01$). It indicates that the prototype selection can significantly improve the classification performance.

B. Feature Visualization

T-Distributed Stochastic Neighbor Embedding (t-SNE) [25] is a technique for dimensionality reduction that is particularly well suited for the visualization of high-dimensional datasets. In this paper, we choose t-sne as the dimensionality reduction approach to show the feature evolution along the cascaded stages. Here, we choose one typical subject's data.

As we can see in Fig 6(a), we input the high-dimensional original DE feature training data into the t-sne algorithm directly, we can intuitively see that the original data cannot be easily discriminated by linear classifiers. We put these original data into our representation network to learn a feature space, and plot the distributions of each class. The same class of training samples are preliminarily clustered in Fig 6(b). Then, we select the prototypes from the trained feature space, the prototype is marked by yellow sign, where there are more than 2000 dots specified training samples and

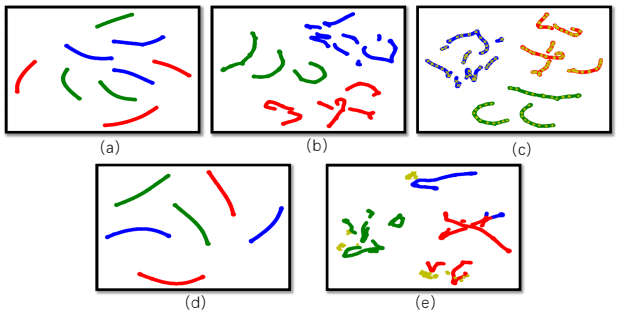

Fig. 6. The visualization of different stage in our methods. (a) is the training data (9 session), (b) is the feature space via the representation network, (c) is the selected prototypes in the feature space, (d) is the testing data (6 session) and (e) the test data and the selected prototypes in the feature space. (red/green/blue dots: the positive/neutral/negative emotion, yellow crosses/plus signs/asterisks: the selected prototypes of positive/neutral/negative emotion state.)

about 100 yellow signs.

In Fig 6(d), we first plot the high-dimensional original DE feature of testing data. And then we use the trained network to map the testing sample into our feature space, in addition, we also plot the selected prototypes together. We can see that the same kind of prototypes are more close to each other than testing samples. From Fig 6(d) $\rightarrow$ Fig 6(e), the testing samples become discriminative using these close prototypes in feature space, which get the accuracy of 89.52%. Note that the colors are marked according to prediction results in Fig 6(e), rather than the actual labels as Fig 6(a), Fig 6(b), Fig 6(c) and Fig 6(d) use.

V. CONCLUSION AND FUTURE WORK

In this paper, we proposed the combining the representation network and prototype selection framework to represent the distribution of the EEG samples and classify the three states of emotion, our method outperforms the state-of-the-art emotion classification approaches in the benchmark EEG emotion dataset–SEED dataset, which achieves a mean accuracy of 93.29%.

The visualizations of the embedding feature show the feature evolution along the cascaded stages. It indicates that our proposed methods can help to learn a more discriminative embedded feature space.

In future, we will focus on the following issues that we have not covered in this paper. First, we will explore the cross session or cross subject domain adaptation problem, so that we may try to use the prototype selection to finding the subset of the source data to represent the target data. Second, we will take advantage of the selected prototypes to cyclical refinement and get the better representation of the EEG samples. Besides, more experiments are needed in order to study the online performance.

REFERENCES

[1] S. Spence, "Descartes' error: Emotion, reason and the human brain," *BMJ*, vol. 310, no. 6988, p. 1213, 1995.

[2] J. Tao and T. Tan, "Affective computing: A review," in *International Conference on Affective computing and intelligent interaction*, pp. 981–995, Springer, 2005.

[3] P. Ekman, "Expression and the nature of emotion," *Approaches to emotion*, vol. 3, pp. 19–344, 1984.

[4] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[5] M. Coulson, "Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence," *Journal of nonverbal behavior*, vol. 28, no. 2, pp. 117–139, 2004.

[6] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Transactions on affective computing*, vol. 1, no. 1, pp. 18–37, 2010.

[7] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, "Emotion representation, analysis and synthesis in continuous space: A survey," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pp. 827–834, IEEE, 2011.

[8] S. M. Alarcao and M. J. Fonseca, "Emotions recognition using eeg signals: A survey," *IEEE Transactions on Affective Computing*, 2017.

[9] G. L. Ahern and G. E. Schwartz, "Differential lateralization for positive and negative emotion in the human brain: Eeg spectral analysis," *Neuropsychologia*, vol. 23, no. 6, pp. 745–755, 1985.

[10] E. Niedermeyer and F. L. da Silva, *Electroencephalography: basic principles, clinical applications, and related fields*. Lippincott Williams & Wilkins, 2005.

[11] C. A. Frantzidis, C. Bratsas, C. L. Papadelis, E. Konstantinidis, C. Pappas, and P. D. Bamidis, "Toward emotion aware computing: an integrated approach using multichannel neurophysiological recordings and affective visual stimuli," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 3, pp. 589–597, 2010.

[12] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu, "Differential entropy feature for eeg-based emotion classification," in *Neural Engineering (NER), 2013 6th International IEEE/EMBS Conference on*, pp. 81–84, IEEE, 2013.

[13] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks," *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 3, pp. 162–175, 2015.

[14] W.-L. Zheng, J.-Y. Zhu, and B.-L. Lu, "Identifying stable patterns over time for emotion recognition from eeg," *IEEE Transactions on Affective Computing*, 2017.

[15] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[16] S. Tripathi, S. Acharya, R. D. Sharma, S. Mittal, and S. Bhattacharya, "Using deep and convolutional neural networks for accurate emotion classification on deap dataset.," in *AAAI*, pp. 4746–4752, 2017.

[17] Z. Li, X. Tian, L. Shu, X. Xu, and B. Hu, "Emotion recognition from eeg using rasm and lstm," in *International Conference on Internet Multimedia Computing and Service*, pp. 310–318, Springer, 2017.

[18] J. Li, Z. Zhang, and H. He, "Hierarchical convolutional neural networks for eeg-based emotion recognition," *Cognitive Computation*, pp. 1–13, 2017.

[19] T. Zhang, W. Zheng, Z. Cui, Y. Zong, and Y. Li, "Spatial-temporal recurrent neural network for emotion recognition," *IEEE transactions on cybernetics*, no. 99, pp. 1–9, 2018.

[20] E. Elhamifar, G. Sapiro, and S. S. Sastry, "Dissimilarity-based sparse subset selection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 11, pp. 2182–2197, 2016.

[21] H. Oh Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4004–4012, 2016.

[22] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine learning*, vol. 3, no. 1, pp. 1–122, 2011.

[23] W. Liu, W.-L. Zheng, and B.-L. Lu, "Multimodal emotion recognition using multimodal deep learning," *arXiv preprint arXiv:1602.08225*, 2016.

[24] J. Lee Rodgers and W. A. Nicewander, "Thirteen ways to look at the correlation coefficient," *The American Statistician*, vol. 42, no. 1, pp. 59–66, 1988.

[25] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.