frontiers in Neuroscience

ORIGINAL RESEARCH
published: 23 June 2021
doi: 10.3389/fnins.2021.690044

Possibilistic Clustering-Promoting Semi-Supervised Learning for EEG-Based Emotion Recognition

Yufang Dan[1†], Jianwen Tao[1†], Jianjing Fu[2] and Di Zhou[3]*

[1] Institute of Artificial Intelligence Application, Ningbo Polytechnic, Ningbo, China, [2] School of Media Engineering, Communication University of Zhejiang, Hangzhou, China, [3] Dazhou Industrial Technological Institute of Intelligent Manufacturing, Sichuan University of Arts and Science, Dazhou, China

The purpose of the latest brain computer interface is to perform accurate emotion recognition through the customization of their recognizers to each subject. In the field of machine learning, graph-based semi-supervised learning (GSSL) has attracted more and more attention due to its intuitive and good learning performance for emotion recognition. However, the existing GSSL methods are sensitive or not robust enough to noise or outlier electroencephalogram (EEG)-based data since each individual subject may present noise or outlier EEG patterns in the same scenario. To address the problem, in this paper, we invent a Possibilistic Clustering-Promoting semi-supervised learning method for EEG-based Emotion Recognition. Specifically, it constrains each instance to have the same label membership value with its local weighted mean to improve the reliability of the recognition method. In addition, a regularization term about fuzzy entropy is introduced into the objective function, and the generalization ability of membership function is enhanced by increasing the amount of sample discrimination information, which improves the robustness of the method to noise and the outlier. A large number of experimental results on the three real datasets (i.e., DEAP, SEED, and SEED-IV) show that the proposed method improves the reliability and robustness of the EEG-based emotion recognition.

Keywords: semi-supervised classification, membership function, electroencephalogram, emotion recognition, fuzzy entropy

OPEN ACCESS

Edited by:
Yuanpeng Zhang,
Nantong University, China

Reviewed by:
Yue Zhao,
Harbin Institute of Technology, China
Jin Cui,
Northwest University, China

***Correspondence:**
Jianjing Fu
fjjmsn@163.com

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *02 April 2021*
Accepted: *28 April 2021*
Published: *23 June 2021*

Citation:
Dan Y, Tao J, Fu J and Zhou D (2021) Possibilistic Clustering-Promoting Semi-Supervised Learning for EEG-Based Emotion Recognition. Front. Neurosci. 15:690044. doi: 10.3389/fnins.2021.690044

INTRODUCTION

Emotion is embodied by human beings: we are born with an innate understanding of emotion (Dolan, 2002; Zhang et al., 2016, 2019b). The complexity of emotion leads to different people's understanding of emotion. Therefore, it is more difficult for machines to accurately understand emotion. As one of the hottest research topics in the field of affective computing, emotion recognition has received extensive attention from the field of pattern recognition and brain neural research (Kim et al., 2013; Mühl et al., 2014; Chu et al., 2017). In this work, we focus on emotional speculation through changes in the body. Basically, the representative internal changes of the body include blood pressure, magneto encephalogram, electroencephalogram (EEG), heart rate, respiratory rate (Mühl et al., 2014), and so on. The EEG-based traditional emotion recognition system usually has two parts: feature extraction and recognizer training (Lan et al., 2019;

Zhang et al., 2020b). Jenke et al. (2014) made a comprehensive review on EEG feature extraction methods. In order to solve the recognition problem, many EEG-based emotion recognition methods have been provided recently (Musha et al., 1997; Kim et al., 2013). An ideal emotion-based brain–computer interface (BCI) can detect the emotional state through spontaneous EEG signals without explicit input from the user (Zhang et al., 2019b) and make a corresponding response to different emotional state. This kind of BCI may enhance the consumer experience in the time of an interactive session. Therefore, different approaches (Zhang et al., 2016, 2017) have been designed to recognize various emotion signals from brain waves. The latest affective BCIs (aBCI) have taken machine learning algorithms and depend on a few features with discriminative information (Jenke et al., 2014; Mühl et al., 2014). A representation of how aBCI exemplification operates is described here. When recording EEG signals, in order to generate a desired target emotion signal, it is necessary to provide users with affective stimulation of specific emotions. In the training/calibration session, the required features and corresponding emotion labels are extracted from EEG signals to train the classifier. In an ongoing BCI session, the feature extractor receives the real-time EEG data, sending the extracted features to the classifier for real-time affection classification. In this paradigm (Mühl et al., 2014), many researchers have reported a pleasing classification performance.

While effective machine learning and deep learning require a large amount of labeled data, sufficient labeled data are often difficult to obtain in real applications. Although manually labeled instances can make up for the lack of labeled instance to a certain extent, this process is time-consuming and laborious. Then, the semi-supervised learning (SSL; Zhou et al., 2003, 2014; Zhu et al., 2003; Chapelle et al., 2006; Zhu, 2008; Zhu and Goldberg, 2009; Gao et al., 2010; Zhou and Li, 2010; Zhao and Zhou, 2018; Tao et al., 2019; Wang Q.-W. et al., 2019; Wang T.-Z. et al., 2019; Zhang et al., 2019c) technique was proposed, which learns a model from a small amount of labeled instances and a large amount of unlabeled instances and solves the problem of insufficient labeled instance (i.e., poor generalization of the model obtained by supervised learning and inaccurate models obtained by unsupervised learning). Tu and Sun (2013) proposed a semi-supervised feature extraction method for EEG classification. Wu and Deng (2018) and Tao et al. (2015, 2016, 2017) proposed a semi-supervised classification framework based on collaborative training and differential evolution to improve the impact of random initial values of input layer parameters of neural networks on classification. Zu et al. (2019) explored to invent a semi-supervised classification method for large-scale remote sensing images based on low-rank block maps, and the reseults have been used to effectively improve classification performance, as graph-based semi-supervised learning (GSSL; Li and Zhou, 2011; Liu et al., 2012; Wang et al., 2012), with its intuitiveness and good learning performance, has been extensively studied. GSSL has two different types of inference, namely, transductive inference (Zhou et al., 2003; Zhu et al., 2003; Wang and Zhang, 2008; Wang et al., 2017) and inductive inference (Belkin et al., 2006; Gao et al., 2010; Nie et al., 2010). The transductive inference assumes that the unlabeled data in the learning process is exactly the test data, and it does not have a good prediction effect on the out-of-sample, for example, LGC (Zhou et al., 2003), GFHF (Zhu et al., 2003), LNP (Wang and Zhang, 2008), and ACA-S3VM (Wang et al., 2017), etc. The inductive inference puts all the instances together in the assumption learning process to find their commonalities and then gets a model. The important point to note is that the test instance does not exist in the training dataset. The Manifold Regularization (MR; Belkin et al., 2006) is a very common inductive GSSL inference, such as, GLSSVM (Gao et al., 2010), FME/U (Nie et al., 2010), etc., and the FME/U was proposed by Nie et al. (2010) generalized the MR framework induction.

Generally, GSSL inference requires certain assumptions. While the GSSL inference models have been employed on EEG datasets due to its effectiveness and intuitiveness, limited effort has been made on improving its performance by the clustering assumption. One of the most common assumptions is the clustering hypothesis: "Similar instances should share the same class label" (Chapelle et al., 2006; Zhu and Goldberg, 2009; Xue et al., 2011; Zhou et al., 2014; Wang Q.-W. et al., 2019; Wang T.-Z. et al., 2019). The assumption has an implicit assumption each instance should clearly belong to a certain class. We call this kind of classification hard classification. However, in real emotion recognition applications, it is difficult to strictly employ this assumption. For example, the same emotion will be understood as different emotion recognition by different subject at different/same scenario.

In order to solve the hard classification problem based on the traditional clustering assumption, Wang et al. (2012) and Zhang et al. (2019a) proposed a new semi-supervised classification method based on modified cluster assumption (SSCCM), which is a soft classification method based on clustering assumption. It constrained the similar instances that share the same label membership. Each instance could belong to multiple class labels and have corresponding membership values, which made good use of the fuzzy clustering assumption (Krishnapuram and Keller, 1993). However, its constraint condition made the total membership of each instance for different labels be 1, which may cause the label membership of some noises to be the same as the label membership of some normal instance, even for one or more classes. The label membership value of the noise may be greater than the normal instance, that is, the correlation is greater, which will cause misrecognition due to its constraint.

Toward the problem of the SSCCM method, we further develop a Possibilistic Clustering Promoting semi supervised learning for EEG-based Emotion Recognition (PCP-ER). The main idea of the method is threefold. First, each instance and its local weighted mean LWM (Bottou and Vapnik, 1992; Atkeson et al., 1997; Xue and Chen, 2007) share the similar memberships. Then, the recognition results obtained by the decision function and membership function are used to verify each other for enhancing the reliability of semi-supervised classification learning method. Finally, a regularization term about fuzzy entropy is added to increase the amount of sample discrimination information. We then obtain a membership function with stronger generalization ability, thereby overcoming

the interference of noise and outlier on the recognition result, and further improving the robustness of the recognition method. In sum, the main contributions of this paper as follows:

(1) A possibilistic clustering promoting semi supervised learning for EEG-based emotion recognition (called as PCP-ER shortly) is proposed;
(2) This method introduces a regularization term about fuzzy entropy to obtain a label membership function with more generalization to overcome the influence of noise and outlier and improve the robustness of the method;
(3) A serial of experiments performed on real-world EEG datasets (i.e., DEAP, SEED, and SEED-IV) to verify the robust effectiveness and recognition reliability of the proposed framework.

The rest of paper is organized as follows. We design our framework PCP-ER in section "Proposed Framework" followed by its corresponding optimal algorithm in section "Optimization." The Algorithm is explained in section "Algorithm description." Section "Discussion" gives algorithm analysis including the reliability, convergence and generalization error bound. Experimental results and analysis on three real-world EEG datasets (i.e., DEAP, SEED, and SEED-IV) are reported in section "Experiment." Finally, we draw a conclusion in section "Conclusion."

PROPOSED FRAMEWORK

Problem Statement

In real classification applications, there are some examples of EEG-based semi-supervised clustering methods where it is difficult to assign an instance explicitly belongs to only one class, such as those boundary instances. Since the hard clustering assumption implicitly constrained each instance that has a clear label assignment, the distribution of real data cannot fully be reflected, and the distribution of these boundary instances may be changed. Therefore, when a semi-supervised classification method adopts this assumption, the predictions on those boundary instances are not good. Wang et al. (2012) and Zhang et al. (2020a) proposed the classification method with modified clustering assumption to a certain extent improved the performance of the classification method based on the hard clustering assumption. Each instance will have a label membership value of a different class rather than only belonging to one class. It can reduce the "misleading" classification impact of those boundary instances. **Figure 1** gives an example, in which the data in both Class1 and Class2 with a question mark are unlabeled data, the rest of other instances are labeled data, and the dashed line is the middle dividing line of the two classes. $x1$ can be regarded as a boundary point or an outlier point. $x2$ is certainly more like an instance of Class1 and class2 than $x1$. However, following the SSCCM, the instance is closer to one class, the membership value about this class is larger and *vice versa*. Therefore, the membership values of $x2$ belonging to class1 and class2 are 0.5 and 0.5, respectively. The membership values of instance $x1$ belonging to class1 and class2 are 0.6 and 0.4, respectively. The membership values of $x1$ belonging to class1, which is larger than that of $x2$, making $x1$ more likely to be a normal instance and $x2$ an outlier. The mainly reason is the constraint term that the sum of membership from different classes of a single instance is always 1 in SSCCM, even if it is a boundary point or an outlier such as $x1$.

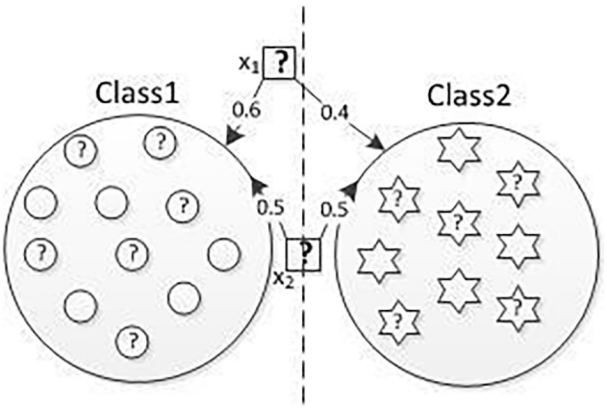

FIGURE 1 | Problem description.

In order to overcome the influence of noise and outlier data on classifiers, we propose a PCP-ER.

Formulation

In order to ensure the classification method of clustering assumption has better classification reliability and robustness, the PCP-ER method achieves the following three goals: (1) any instance should have similar label membership to its corresponding LWM; (2) the decision function and the membership function can mutually verify the classification results of a test instance and have convergence; and (3) we should try to overcome the influence caused by noise and outlier. The classification method proposed in this paper will calculate the LWM of each instance by Euclidean distance and will then obtain the decision function and label membership function through the objective function based on the square loss function with an alternating iterative strategy and utilize the fuzzy entropy to overcome the influence of noise and outlier, thereby improving the robustness of the method. Finally, an optimized classifier model with double verification is constructed by a decision function $f(x)$ and a membership function $w(x)$.

Let dataset $X = \{x_1, x_2, ..., x_i, x_{i+1}, ..., x_n\}$, where $X_l = \{x_i\}_{i=1}^{l}$ is the labeled data with the corresponding labels $Y_l = \{y_1, y_2, ..., y_l\}^T \in \mathbb{R}^{l\times M}$, and n is the total number of instances, $l \ll n$. $X_u = \{x_j\}_{j=l+1}^{n}$ is the unlabeled data, where $x_i \in R^d$ is the $i-th$ instance with d dimensions. The LWM of each x_i (i.e., $\widehat{x}$) is defined as

$$\hat{x}_i = \frac{\sum_{x_j \in Ne(x_i)} W_{ij} x_j}{\sum_{x_j \in Ne(x_i)} W_{ij}}, \tag{1}$$

Here, $Ne(x_i)$ is composed by k nearest neighbors of x_i, each of them is measured by the Euclidean Distance. $G = (X, W)$

denotes a undirected weight graph, where $W \in \mathbb{R}^{n\times n}$ is a weight matrix, $W_{ji} = W_{ij} \geq 0$, and the element of W is measured by

$$W_{ij} = \begin{cases} \exp(-\gamma\|x_i - x_j\|^2), & x_i \text{ is the nearest neighbor of } x_j \\ 0 & \text{otherwise} \end{cases},$$

where γ controls the local scope of the Gaussian kernel function. The larger γ is, the smaller the local scope (i.e., the width) is and *vice versa*. When γ is fixed, W_{ij} decreases monotonically with the increase in the distance between x_i and x_j. Therefore, the clustering problem is transformed into a graph problem. $\widehat{X}_l = \{\widehat{x}\}_{i=1}^{l}$ and $\widehat{X}_u = \{\widehat{x}\}_{i=l+1}^{n}$ are the LWM of l labeled data and $(n-l)$ unlabeled data, respectively. $\{c_m\}_{m=1}^{M}$ is the coded representation of Mclasses. If x_i belongs to the $m-$th class, then $y_i = c_m$, the label and the category encoding are encoded according to one of the M categories, that is, the both of the label and the category coding is a vector of dimension M so PCP-ER can be directly applied to multi-class classification tasks. Let $y_i \in \mathbb{R}^{1\times M}$ and $c_m \in \mathbb{R}^{1\times M}$. If x_i belongs to the $m-$th class, then the $m-$th element of y_i is designated as 1, that is, $y_{im} = 1, m = 1, 2, ..., M$, and the other elements of y_i are 0. $y_{io} = 0, o = 1, 2, ..., M$, and $o \neq m$; and the $m-$th element of c_m is set to 1, i.e., $c_{mm} = 1, m = 1, 2, ..., M$. The rest of the elements in c_m are 0, that is, $c_{mo} = 0, o = 1, 2, ..., M$, and $o \neq m$. Except for the decision function $f(x)$, this method also needs to define a membership function $w(x)$, $w(x_i) \in \mathbb{R}^M$ for any instance x_i, and $w_m(x_i)$ is the membership of x_i belonging to the $m-$th class. Finally, through the improved classification method, each instance is constrained to share the same membership vector with its corresponding LWM according to the local learning principle (Bottou and Vapnik, 1992; Atkeson et al., 1997). The optimization problem of PCP-ER is formulated as

$$\begin{aligned} \min_{f, w_m(x_i)} & \sum_{m=1}^{M}\sum_{i=1}^{n} w_m(x_i)^b \|f(x_i) - c_m\|^2 \\ & + \lambda_s \sum_{m=1}^{M}\sum_{i=1}^{n} w_m(x_i)^b \|f(\hat{x}_i) - c_m\|^2 + \lambda\|f\|_{\mathcal{H}}^2 \\ & - C\sum_{m=1}^{M}\sum_{i=1}^{n}(-w_m(x_i)^b \ln w_m(x_i)^b + w_m(x_i)^b) \\ s.t. \quad & 0 \leq w_m(x_i) \leq 1, \quad m = 1, \ldots, M \end{aligned} \tag{2}$$

where λ_s, λ, and C are regularization parameters corresponding to each term in the objective function, and the parameter b is an exponent on label membership. b is used to control the uncertainty of instances belonging to multiple classes. Specifically, when $b = 1$, the value of each label membership $w_m(x_i)$ is taken from {0, 1}, which will cause PCP-ER to degenerate to the original clustering classification. That is, each instance belongs to only one class. When $b = \infty$, the label membership of all instances on all classes will be equal. In order to avoid the occurrence of trivial solutions, given $b = 2$ in the subsequent derivation process of this paper, the detailed proof process of the value of b has been given in Krishnapuram and Keller (1993). The first term of the objective function in Eq. (2) describes the minimization of the loss by a squared loss function; the second term describes the consistency between the predictions of each instance and its corresponding LWM by adjusting the parameter λ_s; the third term is a regularization term, which is used to prevent the model from over-fitting, and the complexity of the model is controlled by adjusting the parameter λ. The last term describes how to adjust the influence of noise on the model through the fuzzy entropy (Kosko, 1986; Krishnapuram and Keller, 1996), and $\sum_{m=1}^{M}\sum_{i=1}^{n}(-w_m(x_i)^b \ln w_m(x_i)^b + w_m(x_i)^b)$ calculates the fuzzy entropy. The larger the fuzzy entropy, the greater the amount of discriminative information of the sample. The model has better generalization ability. $-C\sum_{m=1}^{M}\sum_{i=1}^{n}(-w_m(x_i)^b \ln w_m(x_i)^b + w_m(x_i)^b)$ is a monotonically decreasing function about $w_m(x_i)$, and it needs to adjust the balance parameter C and force $w_m(x_i)$ to be as large as possible for avoiding trivial solutions. In addition, it can also make noise data have smaller different labels membership. Therefore, fuzzy entropy controls noise and outlier, making the method more robust, and its Robustness has been analyzed and proved in detail in (Krishnapuram and Keller, 1996).

For labeled instances, the label membership function is defined as

$$w_m(x_i) = \begin{cases} 1, \text{ if } x_i \in X_m, & \begin{matrix} i = 1, 2, ...l, \\ m = 1, 2, ...M, \end{matrix} \\ 0, \text{ else} \end{cases} \tag{3}$$

where X_m is a subset with instances belonging to the $m-$th class. The Eq. (2) can be rewritten as

$$\begin{aligned} \min_{f, w_m(x_j)} & \sum_{i=1}^{l}\|f(x_i) - y_i\|^2 + \lambda_s\sum_{i=1}^{l}\|f(\hat{x}_i) - y_i\|^2 \\ & + \sum_{m=1}^{M}\sum_{j=l+1}^{n} w_m(x_j)^2 \|f(x_j) - c_m\|^2 \\ & + \lambda_s\sum_{m=1}^{M}\sum_{j=l+1}^{n} w_k(x_j)^2 \|f(\hat{x}_j) - c_m\|^2 + \lambda\|f\|_{\mathcal{H}}^2 \\ & + C\sum_{m=1}^{M}\sum_{j=l+1}^{n}(w_m(x_j)^2 \ln w_m(x_j)^2 - w_m(x_j)^2) \\ s.t. \quad & 0 \leq w_m(x_j) \leq 1, \quad m = 1, \ldots, M, \; j = l+1, \ldots, n \end{aligned} \tag{4}$$

According to PCP-ER, each instance has a membership vector about all classes and each instance, and its corresponding LWM share the same membership vector.

It should be noted that, in Eq. (2), we adopt a square loss function. However, other classification loss functions can also be used to develop different semi-supervised classification methods based on the possibility clustering assumption. Compared with the Eq. (3) in SSCCM (Wang et al., 2012), the Eq. (2) relaxes the constraint that the sum of the label membership on all classes is 1, and it employs the fuzzy entropy to overcome the influence of noise and outlier for obtaining the more robust model.

OPTIMIZATION

The optimization problem of PCP-ER is a non-convex problem with regard to $f(x)$. We adopt an alternating iterative strategy to achieve the optimization of the decision function $f(x)$ and the label membership function $w(x)$ in this paper, and each iteration has a closed-form solution.

Fixed $w(x)$ firstly to optimize $f(x)$. Since the sixth term in Eq. (4) has no calculation for $f(x)$, we can get

$$\begin{aligned}\min_f & \sum_{i=1}^{l}\|f(x_i)-y_i\|^2+\lambda_s\sum_{i=1}^{l}\|f(\widehat{x_i})-y_i\|^2 \\ & +\sum_{m=1}^{M}\sum_{j=l+1}^{n}w_m(x_j)^2\|f(x_j)-c_m\|^2 \\ & +\lambda_s\sum_{m=1}^{M}\sum_{j=l+1}^{n}w_m(x_j)^2\|f(\widehat{x_j})-c_m\|^2+\lambda\|f\|_H^2 \\ s.t.\ & 0\le w_m(x_j)^2\le 1,\quad m=1,...,M,\quad j=l+1,...,n\end{aligned} \tag{5}$$

According to the Representer Theorem, the minimization problem of Eq. (5) exists in the Reproducing kernel Hilbert space (RKHS), and its solution form can be written as $f(x)=\sum_{i=1}^{n}a_iK(x_i,x)$ (Belkin et al., 2006). The minimization problem is simplified to optimize the finite-dimensional space of coefficient α_i. In this paper, Mercer kernel function $K=(K_{ij})\in\mathbb{R}^{(l+u)\times(l+u)}$, $K_{ij}=K(x_i,x_j)$, $K_i\in\mathbb{R}^{(l+u)\times l}$, $K_u\in\mathbb{R}^{(l+u)\times u}$, $K=[K_l\ K_u]=\begin{bmatrix}K_{ll}&K_{lu}\\K_{lu}^T&K_{uu}\end{bmatrix}$, $\overline{K}=[\overline{K}_l\ \overline{K}_u]=\begin{bmatrix}\overline{K}_{ll}&\overline{K}_{lu}\\\overline{K}_{lu}^T&\overline{K}_{uu}\end{bmatrix}$.

Theorem 1. The best solution of the original optimization problem of the Eq. (5) is $f(x)=\sum_{i=1}^{n}a_iK(x_i,x)$, where

$$\begin{aligned}\alpha=&(Y_lK_l^T+\hat{V}LJ^TK_u^T+\lambda_sY_l\bar{K}_l^T+\lambda_s\hat{V}LJ^T\bar{K}_u^T)\\&(K_lK_l^T+\hat{V}K_uJJ^TK_u^T+\lambda_s\bar{K}_l\bar{K}_l^T+\lambda_s\hat{V}\bar{K}_uJJ^T\bar{K}_u^T+\lambda K)^{-1}\end{aligned}$$

$\alpha_i\in\mathbb{R}^{M\times1}$, $\alpha=[\alpha_1,\alpha_2,...,\alpha_n]\in\mathbb{R}^{M\times n}$ is Lagrange multiplier matrix. $Y_l=(y_1,y_2,...,y_l)\in\mathbb{R}^{M\times l}$, $y_i\in\mathbb{R}^{M\times1}$, $i=1,2,...,l$, $J=[\underbrace{I_u,...,I_u}_{M}]\in\mathbb{R}^{u\times(M\times u)}$, $I_u\in\mathbb{R}^{u\times u}$ is an identity matrix, $L=[L_1,...,L_M]\in\mathbb{R}^{M\times(M\times u)}$, $L_m\in\mathbb{R}^{M\times u}$ is a matrix with all-one vector in the $m-$th row, the other being all-zero vectors. Let $V=[v(x_1)\dots v(x_u)]\in\mathbb{R}^{M\times u}$, $v(x_i)\in\mathbb{R}^{M\times1}$ refers to the membership values of each unlabeled instance on the M classes. $\widehat{V}$ is a diagonal matrix with each element on the diagonal correspond to the squared values of the elements in the corresponding row in the matrix V.

Proof. Like the Eq. (1), each LWM of x_i in the kernel space can be rewritten as. Then $K_{lu}=\langle\phi(X_l),\phi(X_u)\rangle_{\mathcal{H}}$, $K_{ll}=\langle\phi(X_l),\phi(X_l)\rangle_{\mathcal{H}}$, and $K_{uu}=\langle\phi(X_u),\phi(X_u)\rangle_{\mathcal{H}}$, where $\overline{K}_{ll}=\left\langle\phi(X_l),\widehat{\phi(X_l)}\right\rangle_{\mathcal{H}}$, $\overline{K}_{lu}=\left\langle\phi(X_l),\widehat{\phi(X_u)}\right\rangle_{\mathcal{H}}$, $\overline{K}_{ul}=\left\langle\phi(X_u),\widehat{\phi(X_l)}\right\rangle_{\mathcal{H}}$, and $\overline{K}_{uu}=\left\langle\phi(X_u),\widehat{\phi(X_u)}\right\rangle_{\mathcal{H}}$, each element $\overline{K}_{ij}$ can be formulated as

$$\begin{aligned}\bar{K}_{ij}&=\langle\phi(x_i),\widehat{\phi(x_j)}\rangle_{\mathcal{H}}=\left\langle\phi(x_i),\frac{\sum_{x_s\in Ne(x_j)}W_{sj}\phi(x_s)}{\sum_{x_s\in Ne(x_j)}W_{sj}}\right\rangle_{\mathcal{H}}\\&=\frac{\sum_{x_s\in Ne(x_j)}W_{sj}\langle\phi(x_i),\phi(x_s)\rangle_{\mathcal{H}}}{\sum_{x_s\in Ne(x_j)}W_{sj}}\\&=\frac{\sum_{x_s\in Ne(x_j)}W_{sj}K_{is}}{\sum_{x_s\in Ne(x_j)}W_{sj}}\end{aligned} \tag{6}$$

By the F-norm with $\|X\|_F^2=tr(X^TX)$, all matrices and $f(x)=\sum_{i=1}^{n}a_iK(x_i,x)$ can be substituted into Eq. (5), we have

$$\begin{aligned}\min_\alpha F_1=&tr\left((\alpha K_l-Y_l)(\alpha K_l-Y_l)^T\right)\\&+\lambda_str\left((\alpha\bar{K}_l-Y_l)(\alpha\bar{K}_l-Y_l)^T\right)\\&+tr\left((\alpha K_u\mathrm{J}\text{-}\mathrm{L})^T\hat{V}(\alpha K_u\mathrm{J}\text{-}\mathrm{L})\right)\\&+\lambda_str\left((\alpha\bar{K}_u\mathrm{J}\text{-}\mathrm{L})^T\hat{V}(\alpha\bar{K}_u\mathrm{J}\text{-}\mathrm{L})\right)+\lambda tr\left(\alpha K\alpha^T\right)\end{aligned} \tag{7}$$

Set the derivative of F_1 w.r.t. α to zero, we have

$$\begin{aligned}\partial F_1/\partial\alpha=&(\alpha K_l-Y_l)K_l^T+\hat{V}(\alpha K_uJ-L)J^TK_u^T\\&+\lambda_s(\alpha\bar{K}_l-Y)\bar{K}_l^T+\lambda_s\hat{V}(\alpha\bar{K}_uJ-L)\\&J^T\bar{K}_u^T+\lambda\alpha K=0.\end{aligned} \tag{8}$$

According to the Eq. (8), we can get the solution of α, Theorem 1 is proved.

By Fixing $f(x)$ to optimize $w(x)$ from Eq. (4), we can obtain

$$\begin{aligned}\min_{w_m(x_j)}F_2=&\sum_{m=1}^{M}\sum_{j=l+1}^{n}w_m(x_j)^2\|f(x_j)-c_m\|^2\\&+\lambda_s\sum_{m=1}^{M}\sum_{j=l+1}^{n}w_m(x_j)^2\|f(\hat{x}_j)-c_m\|^2\\&+C\sum_{m=1}^{M}\sum_{j=l+1}^{n}(w_m(x_j)^2\ln w_m(x_j)^2-w_m(x_j)^2)\\s.t.\quad&0\le w_m(x_j)\le1,\quad m=1,\dots,M,\\&j=l+1,\dots,n\end{aligned} \tag{9}$$

Theorem 2. The optimal solution of the original optimization problem of the Eq. (4) is

$$w_m(x)=\exp\left(\frac{-(\|f(x)-c_m\|^2+\|f(\hat{x})-c_m\|^2)}{2C}\right). \tag{10}$$

Proof. Set the derivative of F_2 w.r.t. $w_m(x_j)$ to zero, we have

$$\begin{aligned}\partial F_2/\partial w_m(x_j)=&2w_m(x_j)\|f(x_j)-c_m\|^2\\&+2\lambda_sw_m(x_j)\|f(\hat{x}_j)-c_m\|^2\\&+C[2(w_m(x_j)\log w_m(x_j)^2]=0\end{aligned} \tag{11}$$

The solution of $w_m(x_j)$ is

$$w_m(x_j)=\exp\left(\frac{-(\|f(x_j)-c_m\|^2+\|f(\hat{x}_j)-c_m\|^2)}{2C}\right). \tag{12}$$

Therefore, the label membership vector of any instance x can be derived from Eq. (10), and Theorem 2 is proved.

ALGORITHM DESCRIPTION

The optimization of PCP-ER adopts an alternating iterative strategy. PCP-ER belongs to the category of semi-supervised large boundary methods that directly seek large boundary separators. In fact, iterative learning processes are often used in various semi-supervised learning methods. The initial value of membership of an unlabeled instance can be obtained through several strategies, such as randomization strategies, fuzzy clustering techniques (such as FCM), or all zeros simply being set. Actually PCP-ER is start with labeled data to initialize the decision function $f(x)$ in this paper. When $|F(\alpha_m, w_m(x)) - F(\alpha_{m-1}, w_{m-1}(x))| < \varepsilon F(\alpha_{m-1}, w_{m-1}(x))$, the iteration terminates, $F(\alpha_m, w_m(x))$ is the value of the objective function at the $m-$th iteration, and ε is a iterative termination parameter.

ALGORITHM DESCRIPTION OF PCP-ER

Input: the labeled data X_l with the labels Y_l, the unlabeled data X_u, the regularization parameters λ, λ_s, C, the iterative termination parameter ε, and the iterative maximum times T.

Output: the decision function $f(x)$, the label membership function $w(x)$.

Procedure:

1. Initialize the label memberships of unlabeled data;
2. Obtain the initial α by Eq. (6);
3. Obtain the initial $w(x)$ by Eq. (11);
4. Calculate the $F(\alpha_0, w_0(x))$ of objective function

for $m = 1$ **to** T **do**

{

5.1 Update α by Eq. (6);

5.2 Update $w(x)$ by Eq. (11);

5.3 Update the objective function value $F(\alpha_m, w_m(x))$;

5.4 **if** $|F(\alpha_m, w_m(x)) - F(\alpha_{m-1}, w_{m-1}(x))| < \varepsilon F(\alpha_{m-1}, w_{m-1}(x))$

break; **return** $f(x)$ and $w(x)$;

endif

endfor

}

DISCUSSION

PCP-ER Reliability

It takes the decision function and label membership function to identify each other's predicted classification results in order to further enhance the reliability of PCP-ER. This leads to Theorem 3.

Theorem 3. The decision function and label membership function are adopted in PCP-ER to obtain predictions, and their predictions are usually consistent (actually consistent or indirectly consistent). If the two predictions are not consistent, the predict instance may be located near the decision boundary and these predictions may be unreliable.

Proof. Each instance can be predicted by the decision function $y^* = \arg\max_{m=1,\ldots,M} f_m(x)$ from Theorem 1 or the label membership function $y^* = \arg\max_{m=1,\ldots,M} w_m(x)$ from Theorem 2. In the case of $\forall j = 1, \ldots, M; j \neq m$, if $f(x)$ is used to predict x and $f_m(x) > f_j(x)$, then $x \in X_m$. If $w(x)$ is taken to predict x, $w_m(x) > w_j(x)$, the result of $x \in X_m$ can also be obtained. If λ_s is fixed, $f_m(x) + \lambda_s f_m(x) > f_j(x) + \lambda_s f_j(x)$ can also obtain the above consistent prediction result. When $\lambda_s = 0$, the predictions of $f(x)$ and $w(x)$ are consistently. When $\lambda_s \neq 0$, x, and $\widehat{x}$ share the same label from $f(x)$, that is, $\arg\max_{m=1,\ldots,M} f_m(x) = \arg\max_{m=1,\ldots,M} f_m(\widehat{x})$, the prediction results of $f(x)$ and $w(x)$ are also consistent. If $f_j(\hat{x}) - f_m(\hat{x}) < \left(f_m(x) - f_j(x)\right)/\lambda_s \forall j = 1, \ldots, M, j \neq m$, the prediction results of $f(x)$ and $w(x)$ are also consistent. If x is located near the decision boundary, the prediction is obviously different between x and $\hat{x}$, and it is possible that $\hat{x}$ and x are located in different class, then this prediction of x is unreliable.

Finally, three instances can be summarized:

(1) Intrinsic consistent instance, where instances x and $\hat{x}$ get the same label by $f(x)$, then the prediction results of $f(x)$ and $w(x)$ on x are consistent;
(2) Fake-consistent instance, where x is not an intrinsic consistent instances, but $f_j(\hat{x}) - f_m(\hat{x}) < \left(f_m(x) - f_j(x)\right)/\lambda_s$ and the prediction results of $f(x)$ and $w(x)$ on x are still consistent;
(3) Inconsistent instance, where the prediction results of $f(x)$ and $w(x)$ on x are not consistently.

Thus this theorem is proved.

Actually, only one function is needed to predict new instances, and if the memberships of some instances are expected to be obtained, the label membership function is preferred. If these two functions are taken to predict instances at the same time, their prediction inconsistency is used to detect those boundary instances that are difficult to classify, and we do some special processing on them, such as manual labeling, to improve classification reliability. The prediction of these two functions can verify each other, and the reliability of semi-supervised classification can be enhanced by checking their consistency.

PCP-ER Convergence

In order to prove the convergence of Algorithm 1, we have Theorem 4.

Theorem 4. The sequence $\{F(\alpha_m, w_m)\}$ obtained from the above algorithm is convergent.

Proof. Since the objective function $F(\alpha, w)$ is a biconvex function on (α, w) (Gorski and Pfeuffer, 2007). Fixed $w(x)$ and the objective function is a convex function on α, so the optimal α^* can be calculated by minimizing $F(\alpha, w_m)$ in Eq. (6) or optimizing Eq. (5) equivalently. Given $\alpha_{m+1} = \alpha^*$, we know that $F(\alpha_{m+1}, w_m) = F(\alpha^*, w_m) \leq F(\alpha_m, w_m)$. At this time, Fixed α_{m+1} and the objective function is

a convex function on w. Therefore, the optimal w can be obtained by minimizing $F(\alpha_{m+1}, w_m)$ in Eq. (10) or optimizing Eq. (9) equivalently. Given $w_{m+1} = w^*$, we know that $F(\alpha_{m+1}, w_{m+1}) = F(\alpha_{m+1}, w^*) \leq F(\alpha_{m+1}, w_m)$, it can be inferred that$F(\alpha_{m+1}, w_{m+1}) \leq F(\alpha_{m+1}, w_m) \leq F(\alpha_m, w_m)$, $\forall m \in N$, $\{F(\alpha_m, w_m)\}$ is monotonically decreasing. Since the objective function is non-negative, it has a lower bound. Thus the theorem is proved.

Generalization Error Bound of PCP-ER

In statistical learning theory, the VC dimension (Vapnik Chervonenkis dimension; Vapnik, 1995) provided a generalization error bound method that can analyze machine learning (Bishop, 2006). Therefore, this paper selects the VC dimension method to analyze the generalization error bound of PCP-ER.

Theorem 5. (PCP-ER generalization error bound) Let H be the RKHS. The generalization error bound of the learning function $f^{\Phi} \in H$ that satisfies the following Eq. (13) at the probability $1 - \delta$ $(0 < \delta < 1)$:

$$\begin{aligned} R\left(f^{\Phi}\right) &\leq \varepsilon + \frac{1}{l}\sum_{i=1}^{l} L(f^{\phi}(x_i), y_i) + \\ &\quad \frac{1}{n-l}\sum_{m=1}^{M}\sum_{i=l+1}^{n} w_m^2(x_i)\, L(f^{\phi}(x_i), c_m), \\ &\leq \varepsilon + \frac{1}{n}\sum_{i=1}^{n} L(f^{\phi}(x_i), y_i) \end{aligned} \tag{13}$$

where $R\left(f^{\Phi}\right)$ is the expected error, it is the PCP-ER generalization error; the right side of the inequality is the upper bound of the generalization error, $\varepsilon = \sqrt{\frac{1}{2n}\left(\ln d + \ln\frac{1}{\delta}\right)}$ is a constant, d is the number of functions in the hypothesis space, n is the number of samples, and δ is the probability of a function occurrence in the hypothesis space. Since $\frac{1}{n}\sum_{i=1}^{n} L(f(x_i), y_i)$ is the empirical error of the traditional semi-supervised classification method. According to the design idea of PCP-ER, $\frac{1}{l}\sum_{i=1}^{l} L(f(x_i), y_i)$ is the empirical risk of labeled data and $\frac{1}{n-l}\sum_{m=1}^{M}\sum_{i=l+1}^{n} w_m^2(x_i)\, L(f(x_i), c_m)$ the empirical risk of unlabeled data in (14), and $0 \leq w_m(x_i) \leq 1$, $m = 1, \ldots, M$; $i = l+1, \ldots, n$. It follows that the second inequality in (13) is true. Compared with the empirical error of traditional semi-supervised classification methods, PCP-ER has a smaller generalization error bound and a better generalization.

By analysis of Theorem 5, the generalization error of this method can be adjusted and controlled by the membership function $w_m(x)$, which makes it possible to obtain better generalization.

EXPERIMENT

This section will compare the PCP-ER method with the latest semi-supervised recognition method, the hard PCP-ER method, the PCP-ER method joint features by multiple kernel functions (called as MKPCP-ER for short), the PCP-ER method with the features from deep learning (called as DLPCP-ER for short), respectively. The emotion recognition results of all these methods are compared on the real EEG-based datasets [i.e., DEAP (Koelstra et al., 2012), SEED (Zheng and Lu, 2015), and SEED-IV (Zheng et al., 2019)]. It aims to study the following three issues:

(1) How does PCP-ER compare to the latest semi-supervised classification methods?
(2) How does PCP-ER compare to MKPCP-ER, DLPCP-ER and hard PCP-ER?
(3) How does the regularization parameter$\lambda_s$affect the intrinsic consistence of PCP-ER?

Datasets Description

There are a few existing EEG datasets that can be used for affective states investigation. In this paper, we use three publicly available datasets: DEAP (Koelstra et al., 2012), SEED (Zheng and Lu, 2015), and SEED-IV (Zheng et al., 2019).

The DEAP dataset contains 32 experimental subjects. While recording physiological signals, each subject needs to watch 40 1-min music videos as emotional stimuli. The resulting dataset includes 32-channel EEG signals, 4-channel electroencephalogram, 4-channel electromyogram, respiration, plethysmograph, Galvanic Skin Response, and body temperature. Each subject recorded 40 EEG trials, each trial corresponding to an emotion caused by a music video. After watching each video, subjects need to immediately evaluate their truly-felt emotion from five dimensions: valence (related to pleasantness level), arousal (related to excitation level), dominance (related to control power), liking (Related to preference), and familiarity (related to stimulating knowledge). The rating ranges from 1 (weakest) to 9 (strongest) except for familiarity, which is rated from 1 to 5. The EEG signals were recorded by a Biosemi Active Two device at a sampling rate of 512 Hz and down-sampled to 128 Hz.

The original SEED dataset contains 15 experimental subjects. The movie clips are intended to elicit three emotions—positive, neutral, and negative emotions—and five movie clips are assigned to each emotion. All subject need to experience three EEG recording sessions, with two consecutive recording experiments separated by 2 weeks. In each experiment, each subject watch 15 movie clips, each of which was about 4 min long, to induce the desired emotions. The same 15 movie clips were used in all three experiments. The result dataset contains 15 EEG trials for each subject in each experiment and 5 trials for each emotion. EEG signals (EEG signals) are recorded by 62-channel[1] ESI NeuroScan equipment, with a sampling rate of 1,000 Hz and down-sampling to 200 Hz.

For the SEED-IV dataset, a total of 168 movie clips in the material pool containing four emotions (happy, sad, fear, and neutral), 44 participants (22 women, college students) were asked to evaluate their emotional state when watching the

[1]The 62 EEG channels include AF3, AF4, C1, C2, C3, C4, C5, C6, CB1, CB2, CP1, CP2, CP3, CP4, CP5, CP6, CPZ, CZ, F1, F2, F3, F4, F5, F6, F7, F8, FC1, FC2, FC3, FC4, FC5, FC6, FCZ, FP1, FP2, FPZ, FT7, FT8, FZ, O1, O2, OZ, P1, P2, P3, P4, P5, P6, P7, P8, PO3, PO4, PO5, PO6, PO7, PO8, POZ, PZ, T7, T8, TP7, and TP8.

movie clips [Scoring in the two dimensions of valence and arousal; -5 ∼ 5)]. The valence scale ranges from sad to happy. Arousal is measured from calm to excited. Finally, 72 movie clips with the highest recognition are carefully selected from the material pool for the four emotion-evoking experiments. Each movie clip is about 2 min long. The experiment contains 15 experimental subjects. In order to investigate the stability of the model over time, each experimental subject needs to make a total of three experimental records in different periods, to avoid repetition, each movie clip is used for only one trial by the same subject during the three experimental periods. Each experimental for one subject contains 24 trials (each 6 trials correspond to one emotion). The resulting dataset consists of 45 experiments for all subjects. A 62-channel ESI NeuroScan System was used to record the experimenter's EEG signal. The sampling rate was 1,000 Hz, and the sample was down-sampled to 200 Hz. SMI eye-tracking glasses records the eyes movement.

Baselines

This experimental part compares the PCP-ER method in this paper with LapSVM (Belkin et al., 2006), LapRLS (Belkin et al., 2006), TSVM (Joachims, 1999)[2], meanS3VM (Li et al., 2009)[3], and SSCCM (Wang et al., 2012)—five newest semi-supervised classification methods.

LapSVM is the Laplacian Support Vector Machine. This method takes the manifold hypothesis for semi-supervised classification. The loss function is the hinge loss function. According to the Laplacian graph, it searches a maximum-face decision function the entire data distribution.

LapRLS is the Laplacian regularized least squares. The method also uses the manifold hypothesis for semi-supervised classification, but the loss function is the least square loss function.

TSVM is the Transduced Support Vector Machine. This method uses the clustering assumption, in order to find an interface on labeled and unlabeled data, so as to guide the classification boundary through low-density regions.

MeanS3VM is a type of semi-supervised SVM based on the mean value of unlabeled data. The clustering assumption is also adopted, which actually contains two implementation methods (Krishnapuram and Keller, 1996), namely the meanS3VM-iter method based on alternating optimization and the meanS3VM-mkl method based on multiple kernel learning.

Semi-supervised classification method based on modified cluster assumption is a new semi-supervised classification method based on modified clustering assumption. The clustering assumption is also used, its purpose is to find a membership function and a decision function on labeled and unlabeled data, so that similar instances should share similar label membership, and one instance can belong to multiple Classes.

Besides, we then also compare the PCP-ER method with the hard PCP-ER, which only uses the hard clustering assumption. It is supposed in hard PCP-ER that each instance clearly belongs to only one class, which can be formulated in Eq. (14), where each $c_i \in \mathbb{R}^M$ is an one-hot vector.

$$\begin{aligned} \min_{f,y_j} \ & \sum_{i=1}^{l} \|f(x_i) - y_i\|^2 + \lambda_s \sum_{i=1}^{l} \|f(\hat{x}_i) - y_i\|^2 \\ & + \sum_{j=l+1}^{n} \|f(x_j) - y_j\|^2 + \lambda_s \sum_{j=l+1}^{n} \|f(\hat{x}_j) - y_j\|^2 . \\ & + \lambda \|f\|_{\mathcal{H}}^2 \\ s.t. \quad & y_j \in \{c_1, ..., c_M\}, \quad j = l+1, \ldots, n \end{aligned} \tag{14}$$

The objective function in Eq. (14) can also be readily solved by the same strategy as adopted in our PCP-ER.

Experimental Setting

In this section, we will give the experimental settings of the three datasets on the compared methods.

The DEAP dataset is a three binary classification (including valence, arousal, and dominance). One experimental subject contributed 40 samples. Then, the training set includes $40 \times 32 = 1{,}240$ samples from 32 experimental subject, and the test set contains 40 test samples from test subjects. In the DEAP training set, there are three settings: the first one contains 10 labeled instances, the second one contains 50 labeled instances, and the third one contains 100 labeled instances. Furthermore, each setting is related to 12 subsets of labeled data and average performance results on unlabeled data. As with other machine learning algorithms, these state-of-the-art semi-supervised classification algorithms require certain hyper-parameters to be set. For some hyper-parameters, we set them to the default values suggested by their authors. **Table 1** gives details of the other hyper-parameters used in this experiment. Here, you need to use a linear kernel and an RBF kernel. When 10 labeled instances are provided, the width parameter in the RBF kernel is set to the average distance between the instances, and when there are 50 labeled instances, leave-one-subject-out cross-validation is to be taken over the labeled data. When there are 100 labeled instances, 10-fold cross-validation is used on the labeled data, and finally a better result is selected between these two kernels.

The SEED dataset is a three-category dataset, including negative, neutral, and positive. One subject contributes 45 samples. The training set includes $45 \times 14 = 630$ samples from 14 subjects, a total of 210 samples per session, and 45 samples from the test subject. The training data contains 10 labeled and rest unlabeled instances. The linear kernel and a leave-one-subject-out cross-validation method a taken during this process to evaluate emotion recognition accuracy. As with other machine learning algorithms, these state-of-the-arts algorithms that require certain hyper-parameters to be set. For some hyper-parameters, we set them to the default values suggested by their authors. **Table 2** gives details of the hyper-parameters used in this experiment. C is used here to describe the divergence of the dataset. Different C values can be obtained according to different datasets, and $\bar{x} = \frac{\sum_{j=1}^{n} x_j}{n}$.

[2]The codes of LapSVM, LapRLS and TSVM are available from https://www.csie.ntu.edu.tw/~cjlin/libsvm/

[3]The code is downloaded from http://www.lamda.nju.edu.cn/CH.Data.ashx

TABLE 1 | Details of hyper-parameters on SEED dataset.

Method	Hyper-parameters
LapSVM	$C_1 = 1, C_2 = 0.1$
LapRLS	See above
TSVM	See above
MeanS3VM-iter	See above
MeanS3VM-mkl	See above
SSCCM	$\lambda = 0.1,\ \lambda_s = 0.1,\ m = 5,\ \varepsilon = 10^{-3}, C = \frac{\sum_{j=1}^{n} \|x_j - \bar{x}\|}{n}$,
PCP-ER	See above
Hard PCP-ER	See above

TABLE 2 | Details of hyper-parameters on SEED-IV dataset.

Method	Hyper-parameters
LapSVM	$C_1 = 1, C_2 = 0.1$
LapRLS	See above
TSVM	See above
MeanS3VM-iter	See above
MeanS3VM-mkl	See above
SSCCM	$\lambda = 0.1,\ \lambda_s = 0.1,\ m = 5,\ \varepsilon = 10^{-3}, C = \frac{\sum_{j=1}^{n} \|x_j - \bar{x}\|}{n}$,
PCP-ER	See above
Hard PCP-ER	See above

TABLE 3 | Details of hyper-parameters on DEAP dataset.

Method	Hyper-parameters
LapSVM	$C_1 = 100, C_2 = 0.1$
LapRLS	See above
TSVM	See above
MeanS3VM-iter	See above
MeanS3VM-mkl	See above
SSCCM	$\lambda = 1, \lambda_s = 0.1, m = 5, \varepsilon = 10^{-3}$
Hard PCP-ER	See above
PCP-ER	See above

The SEED-IV dataset is a four-category dataset: happy, sad, neutral, and fear. One subject contributes 72 samples. The training set includes $72 \times 14 = 1{,}008$ samples from 14 subjects; there are three sessions, a total of 336 samples per session, and 72 samples from the test subject. The training data contains 10 labeled instances, and the others are unlabeled instances. This process uses a linear kernel and a leave-one-out cross-validation method to evaluate classification accuracy. As with other machine learning algorithms, these state-of-the-arts algorithms that require certain hyper-parameters to be set. For some hyper-parameters, we set them to the default values suggested by their authors. **Table 3** gives details of the hyper-parameters used in this experiment. C is used here to describe the divergence of the data set. Different C values can be obtained according to different datasets, and $\bar{x} = \frac{\sum_{j=1}^{n} x_j}{n}$.

TABLE 4 | Performance about PCP-ER and the latest methods on DEAP dataset.

Methods	DEAP		
	10 Labels	**50 Labels**	**100 Labels**
LapSVM	49.22	53.05	63.22
LapRLS	50.06	57.49	63.46
TSVM	44.70	47.66	52.49
MeanS3VM-iter	49.83	53.43	59.17
MeanS3VM-mkl	52.11	58.47	60.54
SSCCM	**56.54**	61.31	63.33
PCP-ER	55.7	**62.40**	**66.71**
Consis. rate	0.9827	0.9943	1.00

Experimental Results and Analysis

Specifically, in the following tables (i.e., **Tables 4**, **5**) of experimental results, the bold values in each column indicate the best accuracy in all these tables, and the bold values in last column from **Table 5** indicate the best average performance results achieved by the compared methods. The last Avg. column shows the average performance of each method on all data sets. In **Tables 4**, **5**, the consistency rate of PCP-ER on different settings of dataset is given in the last row. From the results of **Tables 4**, **5**, we can draw the following conclusions.

Results on DEAP

The experimental comparison among PCP-ER and the six latest methods on DEAP with different number of labeled samples. The experimental results are shown in **Table 4**. The consistency rate gradually approaches to 1 which increases as the labeled samples increase. Since the SSCCM is used to deal with normal data, it obtains the best performance with 10 labeled samples. The possible reason is that the samples are normal. However, the SSCCM method is slightly better than the PCP-ER method in this case, and the PCP-ER method has the best performance in most cases with the increase of the number of labeled samples. It shows that the clustering hypothesis with fuzzy entropy can overcome the influence of noise and outliers in semi-supervised classification.

Results on SEED and SEED-IV

We performed an experimental comparison among PCP-ER and the six latest methods on each session with 10 labels of SEED and SEED-IV datasets. The experimental results are shown in **Table 5**. The average performance of the proposed method is the best. In addition, the consistency rate is close to 1. Although the consistency rate on Session 2 of SEED-IV dataset isn't good, its value is as high as 0.988. It is worth noting that only method MeanS3VM-mkl performs slightly better than PCP-ER method on session 2 of SEED dataset, the possible reason is that the MeanS3VM-mkl employed multiple kernels and enriched features of dataset. Nevertheless, the performance of the PCP-ER method is also closely followed, and the proposed method has the best performance of all the other datasets. It shows that the clustering hypothesis with fuzzy entropy can overcome the influence of noise in semi-supervised emotion recognition on SEED and SEED-IV.

TABLE 5 | Performance comparison among PCP-ER and the latest methods on SEED and SEED-IV datasets.

Methods	Seed				Seed-IV			
	Session1	Session2	Session3	Avg.	Session1	Session2	Session3	Avg.
LapSVM	52.26	49.46	58.39	53.37	58.00	51.88	56.33	56.33
LapRLS	52.08	50.55	57.16	53.26	57.29	51.04	55.72	55.72
TSVM	49.83	47.29	53.44	50.19	55.73	45.20	50.60	50.6
MeanS3VM-iter	55.27	49.71	58.21	54.40	58.08	48.95	56.49	56.49
MeanS3VM-mkl	60.02	**51.20**	58.78	56.67	60.29	51.68	57.25	57.25
SSCCM	61.78	50.68	60.11	57.52	59.38	51.17	61.78	61.78
PCP-ER	**62.13**	50.04	**60.48**	**57.55**	**62.45**	**52.30**	**62.36**	**62.36**
Consis. rate	0.991	0.99	1.00	0.994	1.00	0.988	0.999	0.999

Multiple-Kernel Learning

We further evaluate the effectiveness of our method with different kernel functions (called as MKPCP-ER for short) to present instances from each dataset. Given the empirical kernel mapping set $\{\phi_k\}_{k=1}^{\mho}$, each mapping X_a into $\mho$ different kernel spaces, each X corresponds to each dataset [i.e., DEAP (Koelstra et al., 2012), SEED (Zheng and Lu, 2015), and SEED-IV (Zheng et al., 2019)], and we can integrate them orthogonally to the final space by concatenation, i.e., $\widetilde{\phi}(x_i) = [\phi_1(x_i)^T, \phi_2(x_i)^T, ..., \phi_{\mho}(x_i)^T]^T \in \mathbb{R}^{\mho n_a}$, for $x_i \in X_a$, $n_a (a = 1, 2, 3)$ training samples in each dataset. The final kernel matrix in this new space is defined as $K_{new} = [\widetilde{K}_1; \widetilde{K}_2; ...; \widetilde{K}_{\mho}]$, where $\widetilde{K}_i$ is the kernel matrix in the $i-$thfeature space about $i-$th dataset. Therefore, besides the above mentioned Gaussian kernel, we additionally employ another three types of kernels in MKPCP-ER: Laplacian kernel $K_{ij} = \exp\left(-\sqrt{\sigma}\|x_i - x_j\|\right)$, inverse square distance kernel $K_{ij} = 1/\left(1 + \sigma\|x_i - x_j\|^2\right)$, and inverse distance kernel $K_{ij} = 1/\left(1 + \sqrt{\sigma}\|x_i - x_j\|\right)$.

It can be clearly seen from **Figure 2** that MKPCP-ER is obviously better than PCP-ER in terms of mean accuracies in all cases, which justifies that the multi-kernel trick can improve the quality of semi-supervised emotion recognition on each dataset.

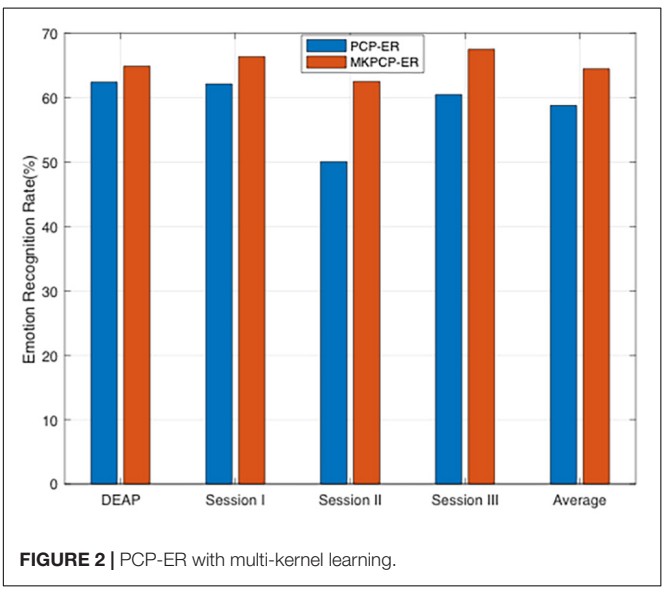

FIGURE 2 | PCP-ER with multi-kernel learning.

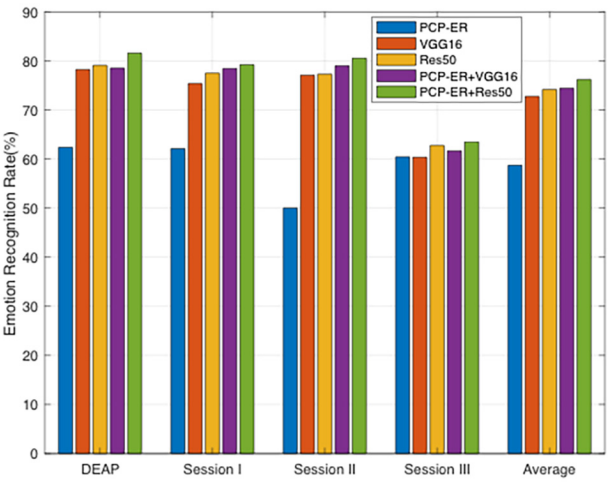

FIGURE 3 | Emotion recognition accuracies (%) of different methods using deeply extracted features.

Deep Features Learning

In the past decades, deep learning attracts more and more attention due to its powerful representation ability and dramatic improvement over the traditional shallow methods. The EEG emotion recognition method based on deep learning has also been widely used and has achieved better recognition effect than traditional methods. For example, in Zheng and Lu (2015) proposed a deep belief network for EEG emotion classification; in Song et al. (2018) and Zhong et al. (2020), authors used graphics to model multi-channel EEG features and then classified EEG emotion on this basis; Li et al. (2018) proposed a new neural network model, which uses time information for EEG emotion

TABLE 6 | Performance comparison between hard PCP-ER and PCP-ER on DEAP dataset.

Methods	DEAP			
	10 Labels	50 Labels	100 Labels	Avg.
hard PCP-ER	53.27	58.44	62.69	58.1
PCP-ER	**55.7**	**62.40**	**66.71**	**61.1**

TABLE 7 | Performance comparison between hard PCP-ER and PCP-ER on SEED and SEED-IV datasets.

Methods	SEED				SEED-IV			
	Session1	Session2	Session3	Avg.	Session1	Session2	Session3	Avg.
hard PCP-ER	60.57	47.32	56.78	54.89	60.38	49.08	61.66	57.04
PCP-ER	**62.13**	**50.04**	**60.48**	**57.55**	**62.45**	**52.30**	**62.36**	**59.03**

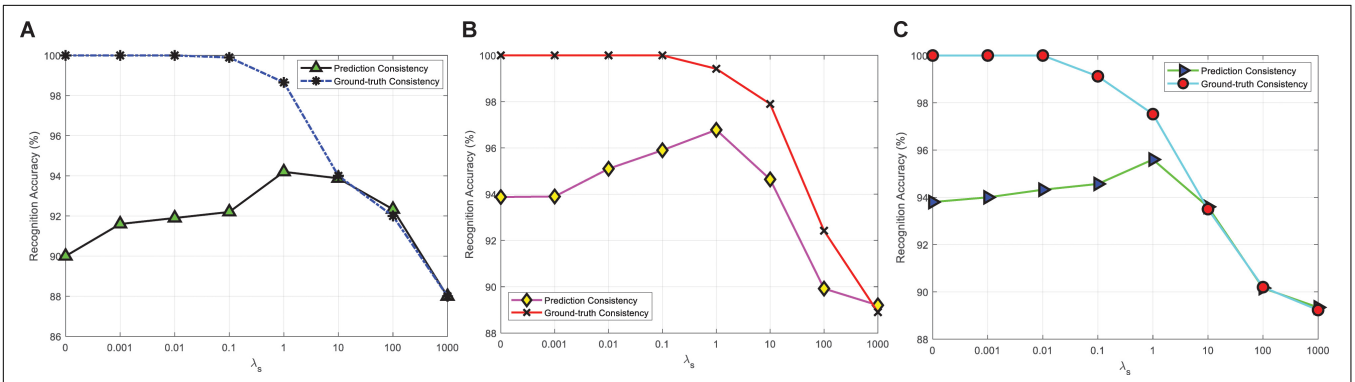

FIGURE 4 | Prediction consistency and ground-truth consistency between $f(x)$ and $w(x)$ with different λ_S values on **(A)** DEAP; **(B)** SEED; and **(C)** SEED-IV.

recognition task. We therefore additionally compare our PCP-ER method with the recently proposed deep transfer learning models VGG16 (Simonyan and Zisserman, 2014) and ResNet50 (He et al., 2016; it is the same as Res50 in the following) for emotion recognition using deeply extracted features. In our PCP-ER, we can tackle the problem of deep emotion recognition with two steps: firstly, a higher-level feature extraction is learnt in an unsupervised fashion from all available datasets using the popular deep architectures (e.g., VGG16 or Res50); secondly, our PCP-ER is trained on the transformed data of all datasets and then used to classify test dataset. For fair comparison, however, we follow the experimental setup in Zhou et al. (2018) and Zhu et al. (2017). Specifically, we first fine-tune pre-trained deep models (e.g., VGG16, and Res50) by using the labeled samples in the dataset, and then use these fine-tuned CNN models to extract the features from EEG in dataset. Finally, we perform emotion recognition using PCP-ER on these deeply extracted features. In the context of our experiments, we denote our methods with different deep models as PCP-ER+VGG16, PCP-ER+Res50, respectively. As for VGG16 and Res50, we use their released source codes and fine-tune the pre-trained deep models, respectively.

All experimental results are reported in **Figure 3**. As can be seen from this plot, the deep learning methods are originally proposed to learn each dataset features, while our proposed method aims to improve the anti-interference ability, namely, their methods focus on feature learning, while our work focuses on emotion recognition. So our proposed method can be used to further improve the recognition accuracies by employing the features extracted by deep models, i.e., VGG16 and Res50, rather than the original EEG patterns. This indicates that the recognition-level constraint can preserve all discriminative structures of labeled samples for the guidance of unlabeled samples recognition, which demonstrates the effectiveness of PCP-ER framework. From the plot bars of **Figure 3**, it can be observed that PCP-ER+VGG16 consistently outperforms VGG16, while PCP-ER+Res50 is consistently better than Res50, which demonstrates that our PCP-ER method is complementary to the two deep learning methods VGG16 and Res50 by exploiting more features from dataset.

Comparison Between PCP-ER and Hard PCP-ER

Tables 6, **7** show the performance results of PCP-ER and hard PCP-ER on DEAP, SEED, and SEED-IV datasets, respectively. Specifically, in the following tables (i.e., **Tables 6**, **7**) of experimental results, the bold values in each column indicate the best accuracy, and the bold values in last column of different dataset indicate the best average performance results achieved by the compared hard PCP-ER and PCP-ER method. From the overall observation of the results in **Tables 6**, **7**, when there are more training samples or more labeled samples, the PCP-ER and hard PCP-ER methods are better. Moreover, the performance and average performance of PCP-ER are better than hard PCP-ER. It demonstrates that the PCP-ER method based on membership degree is effective and robust on EEG-based emotion recognition.

Consistency Analysis

Figure 4 shows the experimental results of PCP-ER actual consistency rates corresponding to different λ_s values {0, 0.001, 0.01, 0.1, 1, and 101001000} on DEAP, SEED, SEED-IV three datasets. In **Figures 4A,C**, when λ_s is small enough, the prediction consistency rate can reach 100%, and then the consistency rate gradually decreases with the increase of λ_s. Since the indirect consistency instance becomes the inconsistent instance, it finally becomes equal to the ground-truth consistency rate. In **Figure 4B**, when λ_s is in the range of 1 to 1000, the prediction consistency rate and the ground-truth consistency rate are not equal, although the trend of change is the same with

Figures 4A,C, and the possible reason is that there are the least training samples on SEED dataset. In addition, in **Figures 4A–C**, with the increase of λ_s to 1, the ground-truth consistency rate also increases. When λ_s continues to increase, the ground-truth consistency rate begins to decrease. It may be that when λ_s is far less than or greater than 1, the PCP-ER will focus more on the emotion recognition of samples or LWM samples and not on its prediction consistency.

CONCLUSION

The existing GSSL methods construct undirected weight grapy, which is sensitive or not enough robust to noise or outlier from EEG patterns. At the end, we proposed PCP-ER for EEG-based affective recognition. By adding the regularization term of fuzzy entropy, the amount of discrimination information of samples is increased, and a more generalized emotion recognizer is obtained through learning to overcome the negative effects of noise and outliers and improve the robustness of the method. Experimental results on the three real datasets DEAP, SEED, and SEED-IV show that the proposed method improves more the reliability and robustness of emotion recognition than competing algorithms. Moreover, both PCP-ER with multi-kernel and depth features extraction obtained better performance than PCP-ER. These tricks can improve the quality of the PCP-ER method on each dataset. However, in the process of optimization, how to obtain an effective combined multi-kernel function or kernel space and how to analyze and demonstrate the consistency of the proposed method in theory are issues worthy of further discussion.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://epileptologie-bonn.de/cms/upload/workgroup/lehnertz/eegdata.html.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The ethics committee waived the requirement of written informed consent for participation.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

This work was supported partly by two High level talents introduction research projects of Ningbo Polytechnic (under grant nos. RC201808 and RC201702), and partly by Foundation of Zhejiang Educational Committee (under grant no. Y201941140), and partly by Zhejiang basic public welfare research program (under grant no. LGG20F020013).

REFERENCES

Atkeson, C.-G., Moore, A.-W., and Schaal, S. (1997). Locally weighted learning. *Artif. Intell. Rev.* 11, 11–73. doi: 10.1007/978-94-017-2053-3_2

Belkin, M., Niyogi, P., and Sindhwani, V. (2006). Manifold regularization: a geometric framework for learning from examples. *J. Mach. Learn. Res.* 7, 2399–2434.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York, NY: Springer.

Bottou, L., and Vapnik, V. (1992). Local learning algorithms. *Neural Comput.* 4, 888–900. doi: 10.1162/neco.1992.4.6.888

Chapelle, O., Bernhard, S., and Alexander, Z. (2006). *Semi-Supervised Learning*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/9780262033589.001.0001

Chu, W. S., De la Torre, F., and Cohn, J. F. (2017). Selective transfer machine for personalized facial action unit detection. *IEEE Trans. Patt. Analys. Mach. Intellig.* 39, 529–545. doi: 10.1109/TPAMI.2016.2547397

Dolan, R. J. (2002). Emotion, cognition, and behavior. *Science* 298, 1191–1194. doi: 10.1126/science.1076358

Gao, J., Wang, S.-T., and Deng, Z.-H. (2010). Global and local preserving based semi-supervised support vector machine. *Acta Electron. Sin.* 38, 1626–1633.

Gorski, J., and Pfeuffer, F. (2007). Biconvex sets and optimization with biconvex functions: a survey and extensions. *Math. Methods Operation Res.* 66, 373–407. doi: 10.1007/s00186-007-0161-1

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, 770–778. doi: 10.1109/CVPR.2016.90

Jenke, R., Peer, A., and Buss, M. (2014). Feature extraction and selection for emotion recognition from EEG. *IEEE Trans. Affect. Comput.* 5, 327–339. doi: 10.1109/taffc.2014.2339834

Joachims, T. (1999). "Transductive inference for text classification using support vector machines," in *Proceedings of the 16th International Conference on Machine Learning*, Bled, 200–209.

Kim, M.-K., Kim, M., Oh, E., and Kim, S.-P. (2013). A review on the computational methods for emotional state estimation from the human EEG. 2013:573734. doi: 10.1155/2013/573734

Koelstra, S., Mühl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., et al. (2012). DEAP: a database for emotion analysis using physiological signals. *IEEE Trans. Affect. Comput.* 3, 18–31. doi: 10.1109/t-affc.2011.15

Kosko, B. (1986). Fuzzy entropy and conditioning. *Inform. Sci.* 40, 165–174. doi: 10.1016/0020-0255(86)90006-x

Krishnapuram, R., and Keller, J.-M. (1993). A possibilistic approach to clustering. *IEEE Trans. Fuzzy Syst.* 1, 98–110. doi: 10.1109/91.227387

Krishnapuram, R., and Keller, J.-M. (1996). The possibilistic c-means algorithm: insights and recommendations. *IEEE Trans. Fuzzy Syst.* 4, 385–393. doi: 10.1109/91.531779

Lan, Z., Sourina, O., Wang, L., Scherer, R., and Müller-Putz, G. R. (2019). Domain adaptation techniques for eeg-based emotion recognition: a comparative study on two public datasets. *IEEE Trans. Cogn. Dev. Syst.* 11, 85–94. doi: 10.1109/tcds.2018.2826840

Li, Y., Zheng, W., Cui, Z., Zhang, T., and Zong, Y. (2018). "A novel neural network model based on cerebral hemispheric asymmetry for EEG emotion recognition," in *27th International Joint Conference on Artificial Intelligence (IJCAI)*, New York, NY.

Li, Y.-F., and Zhou, Z.-H. (2011). "Improving semi-supervised support vector machines through unlabeled instances selection," in *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, San Francisco, CA, 386–391.

Li, Y.-F., Kwok, J.-T., and Zhou, Z.-H. (2009). "Semi-supervised learning using label mean," in *Proceedings of the 26th International Conference on Machine Learning*, Montreal, QC, 633–640. doi: 10.1145/1553374.1553456

Liu, W., Wang, J., and Chang, S.-F. (2012). Robust and scalable graph-based semi-supervised learning. *Proc. IEEE* 100, 2624–2638. doi: 10.1109/JPROC.2012.2197809

Mühl, C., Allison, B., Nijholt, A., and Chanel, G. (2014). A survey of affective brain computer interfaces: principles, state-of-the-art, and challenges. *Brain Comput. Interf.* 1, 66–84. doi: 10.1080/2326263x.2014.912881

Musha, T., Terasaki, Y., Haque, H. A., and Ivamitsky, G. A. (1997). Feature extraction from EEGs associated with emotions. *Artific. Life Robot.* 1, 15–19. doi: 10.1007/bf02471106

Nie, F., Xu, D., Tsang, I., and Zhang, C. (2010). Flexible manifold embedding: a framework for semi-supervised and unsupervised di- mension reduction. *IEEE Trans. Image Process.* 19, 1921–1932. doi: 10.1109/tip.2010.2044958

Simonyan, K., and Zisserman, A. (2014). "Very deep convolutional networks for large-scale image recognition," in *Proceedings of the International Conference on Learning Representations (ICLR)*, New Orleans, LA, 1–14.

Song, T., Zheng, W., Song, P., and Cui, Z. (2018). EEG emotion recognition using dynamical graph convolutional neural networks. *Proc. IEEE Trans. Affect. Comput.* 11, 532–541. doi: 10.1109/taffc.2018.2817622

Tao, J. W., Song, D., Wen, S., and Hu, W. (2017). Robust multi-source adaptation visual classification using supervised low-rank representation. *Patt. Recogn.* 61, 47–65. doi: 10.1016/j.patcog.2016.07.006

Tao, J., Wen, S., and Hu, W. (2015). L1-norm locally linear representation regularization multi-source adaptation learning. *Neural Netw.* 69, 80–98. doi: 10.1016/j.neunet.2015.01.009

Tao, J., Wen, S., and Hu, W. (2016). Multi-source adaptation learning with global and local regularization by exploiting joint kernel sparse representation. *Knowl. Based Syst.* 98, 76–94. doi: 10.1016/j.knosys.2016.01.021

Tao, J., Zhou, D., Liu, F., and Zhu, B. (2019). Latent multi-feature co-regression for visual recognition by discriminatively leveraging multi-source models. *Patt. Recogn.* 87, 296–316. doi: 10.1016/j.patcog.2018.10.023

Tu, W., and Sun, S. (2013). Semi-supervised feature extraction for EEG classification. *Patt. Analys. Appl. Paa* 16, 213–222. doi: 10.1007/s10044-012-0298-2

Vapnik, V.-N. (1995). *The Nature of Statistical Learning Theory*. New York, NY: Springer, 69–83. doi: 10.1007/978-1-4757-2440-0

Wang, F., and Zhang, C. (2008). Label propagation through linear neighborhoods. *IEEE Trans. Knowl. Data Eng.* 20, 55–67. doi: 10.1109/tkde.2007.190672

Wang, Q.-W., Li, Y.-F., and Zhou, Z.-H. (2019). "Partial label learning with unlabeled data," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI'19)*, Macao, 3755–3761. doi: 10.24963/ijcai.2019/521

Wang, T.-Z., Huang, S.-J., and Zhou, Z.-H. (2019). "Towards identifying causal relation between instances and labels," in *Proceedings of the 19th SIAM International Conference on Data Mining (SDM'19)*, Calgary, 289–297. doi: 10.1137/1.9781611975673.33

Wang, Y.-Y., Chen, S.-C., and Zhou, Z.-H. (2012). New semi-supervised classification method based on modified cluster assumption. *IEEE Trans. Neural Netw. Learn. Syst.* 23, 689–702. doi: 10.1109/tnnls.2012.2186825

Wang, Y.-Y., Meng, Y., Fu, Z.-Y., and Xue, H. (2017). Toward-s safe semi-supervised classification: adjusted cluster assumption via clustering. *Neural Process. Lett.* 46, 1–12.

Wu, M.-S., and Deng, X.-G. (2018). Semi-supervised pattern classification method based on Tri-DE-ELM. *Comput. Eng. Appl.* 54, 109–114.

Xue, H., and Chen, S.-C. (2007). "Alternative robust local embedding," in *Proceedings of the International Conference Wavelet Analysis Pattern Recognition*, Paris.

Xue, H., Chen, S.-C., and Yang, Q. (2011). Structural regularized support vector machine: a framework for structural large margin classifier. *IEEE Trans. Neural Netw.* 22, 573–587. doi: 10.1109/tnn.2011.2108315

Zhang, Y., Chung, F. L., and Wang, S. (2019a). Takagi-sugeno-kang fuzzy systems with dynamic rule weights. *J. Intellig. Fuzzy Syst.* 37, 8535–8550. doi: 10.3233/jifs-182561

Zhang, Y., Dong, J., Zhu, J., and Wu, C. (2019b). Common and special knowledge-driven TSK fuzzy system and its modeling and application for epileptic EEG signals recognition. *IEEE Access* 7, 127600–127614. doi: 10.1109/access.2019.2937657

Zhang, Y., Li, J., Zhou, X., Zhou, T., Zhang, M., Ren, J., et al. (2019c). A view-reduction based multi-view TSK fuzzy system and its application for textile color classification. *J. Ambient Intellig. Hum. Comput.* 1–11. doi: 10.1007/s12652-019-01495-9

Zhang, Y., Tian, F., Wu, H., Xingyun, G., Xiaofeng, Z., Lemin, T., et al. (2017). Brain MRI tissue classification based fuzzy clustering with competitive learning. *J. Med. Imag. Health Inform.* 7, 1654–1659. doi: 10.1166/jmihi.2017.2181

Zhang, Y., Wang, L., Wu, H., Xingyun, G., Yao, D., Dong, J., et al. (2016). A clustering method based on fast exemplar finding and its application on brain magnetic resonance images segmentation. *J. Med. Imag. Health Inform.* 6, 1337–1344. doi: 10.1166/jmihi.2016.1923

Zhang, Y., Chung, F., and Wang, S. (2020a). Clustering by transmission learning from data density to label manifold with statistical diffusion. *Knowl. Based Syst.* 193:105330. doi: 10.1016/j.knosys.2019.105330

Zhang, Y., Wang, S., Xia, K., Jinag, Y., Qian, P., For the Alzheimer's Disease Neuroimaging Initiative, et al. (2020b). Alzheimer's disease multiclass diagnosis via multimodal neuroimaging embedding feature selection and fusion. *Inform. Fusion* 66, 170–183. doi: 10.1016/j.inffus.2020.09.002

Zhao, P., and Zhou, Z.-H. (2018). "Label distribution learning by optimal transport," in *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI'18)*, New Orleans, LA, 4506–4513.

Zheng, W.-L., and Lu, B.-L. (2015). Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Trans. Autonom. Ment. Dev.* 7, 162–175. doi: 10.1109/tamd.2015.2431497

Zheng, W.-L., Liu, W., Lu, Y.-F., Lu, B.-L., and Cichocki, A. (2019). EmotionMeter: a multimodal framework for recognizing human emotions. *IEEE Trans. Cybernet.* 49, 1110–1122. doi: 10.1109/tcyb.2018.2797176

Zhong, P., Wang, D., and Miao, C. (2020). EEG-based emotion recognition using regularized graph neural networks. *IEEE Trans. Affect. Comput.* 2020:1. doi: 10.1109/taffc.2020.2994159

Zhou, D., Bousquet, O., Lal, T., Weston, J., and Schlkopf, B. (2003). "Learning with local and global consistency," in *Proceedings of the Advances in Neural Information Processing Systems (NIPS '03)*, (Vancouver: Curran Associates Inc), 321–328.

Zhou, S.-H., Liu, X.-W., Zhu, C.-Z., Liu, Q., and Yin, J. (2014). "Spectral clustering-based local and global structure preservation for feature selection," in *Proceedings of 2014 International Joint Conference on Neural Networks*, Beijing, 550–557.

Zhou, X., Jin, K., Shang, Y., and Guo, G. (2018). Visually interpretable representation learning for depression recognition from facial Im-ages. *IEEE Trans. Affect. Comput.* 11, 542–552. doi: 10.1109/TAFFC.2018.2828819

Zhou, Z.-H., and Li, M. (2010). Semi-supervised learning by disagreement. *Knowl. Inform. Syst.* 24, 415–439. doi: 10.1007/s10115-009-0209-z

Zhu, X., Ghahramani, Z., and Lafferty, J. (2003). "Semi-supervised learning using Gaussian fields and harmonic functions," in *Proceedings of the 20th International Conference on Machine Learning*, (California: AAAI Press), 912–919.

Zhu, X.-J. (2008). *Semi-Supervised Learning Literature Survey. Computer Science TR 1530*. Madison: University of Wisconsin Madiso.

Zhu, X.-J., and Goldberg, A. (2009). *Introduction to Semi-Supervised Learning*. San Rafael, CA: Morgan & Claypool.

Zhu, Y., Shang, Y., Shao, Z., and Guo, G. (2017). "Automated depression diagnosis based on deep networks to encode facial appearance and dynamics," in *Proceedings of the IEEE Transactions on Affective Computing*, Cambridge. doi: 10.1109/TAFFC.2017.2650899

Zu, B.-B., Xia, K.-W., Wen, L., Niu, W.-J., and Jiang, X.-Q. (2019). Semi-supervised classification application of remote sensing image based on block low rank images. *J. Front. Comput. Sci. Technol.* 13, 1217–1226.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.