

Article

EEG-Based Emotion Recognition by Retargeted Semi-Supervised Regression with Robust Weights

Ziyuan Chen ¹, Shuzhe Duan ¹  and Yong Peng ^{2,3,*} ¹ Zhuoyue Honors College, Hangzhou Dianzi University, Hangzhou 310018, China² School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China³ Zhejiang Key Laboratory of Brain-Machine Collaborative Intelligence, Hangzhou 310018, China

* Correspondence: yongpeng@hdu.edu.cn

Abstract: The electroencephalogram (EEG) can objectively reflect the emotional state of human beings, and has attracted much attention in the academic circles in recent years. However, due to its weak, non-stationary, and low signal-to-noise properties, it is inclined to cause noise in the collected EEG data. In addition, EEG features extracted from different frequency bands and channels usually exhibit different levels of emotional expression abilities in emotion recognition tasks. In this paper, we fully consider the characteristics of EEG and propose a new model RSRRW (retargeted semi-supervised regression with robust weights). The advantages of the new model can be listed as follows. (1) The probability weight is added to each sample so that it could help effectively search noisy samples in the dataset, and lower the effect of them at the same time. (2) The distance between samples from different categories is much wider than before by extending the ϵ -dragging method to a semi-supervised paradigm. (3) Automatically discover the EEG emotional activation mode by adaptively measuring the contribution of sample features through feature weights. In the three cross-session emotion recognition tasks, the average accuracy of the RSRRW model is 81.51%, which can be seen in the experimental results on the SEED-IV dataset. In addition, with the support of the Friedman test and Nemenyi test, the classification of RSRRW model is much more accurate than that of other models.

Keywords: electroencephalogram (EEG); robustness; semi-supervised classification; emotion recognition; emotional activation pattern mining



Citation: Chen, Z.; Duan, S.; Peng, Y. EEG-Based Emotion Recognition by Retargeted Semi-Supervised Regression with Robust Weights. *Systems* **2022**, *10*, 236. <https://doi.org/10.3390/systems10060236>

Academic Editors: Enrico Vezzetti, Andrea Luigi Guerra, Gabriele Baronio, Domenico Speranza and Luca Ulrich

Received: 27 October 2022

Accepted: 25 November 2022

Published: 29 November 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

As a complex psychological state, emotion plays a key role in human cognition, including rational decision-making, perception, interpersonal communications and human intelligence [1]. Therefore, emotion recognition has attracted the attention of researchers from various disciplines. Usually, researchers investigate emotion recognition from the data sources of language, body movements, speech, and facial expressions. However, these patterns have certain drawbacks. (1) When subjects deliberately disguise their emotions, the performance of the method may be significantly affected by the deceptive data collected from the subjects based on the above data model. (2) The previous mode is impractical for people with physical disabilities (deafness, aphasia and etc.) Therefore, we need a more objective mode of emotional recognition. Firstly, considering that emotion is hard to measure as the result of spontaneousness, thus it can be observed by the accompanying physiological reactions in the central nervous system and periphery [2,3]. Secondly, EEG is a signal from the central nervous system, which has several advantages, such as large amounts of information, simple operation and low-cost [4]. So, together with the rapid development of non-stationary signal processing and analysis techniques, EEG-based emotion recognition has become a research hotspot [5].

In recent years, emotion recognition based on EEG signals has aroused the discussion of researchers. Li et al. [6] summarized preceding research results, including traditional machine learning works [7], deep learning based works [8], transfer learning based works [9,10] and ensemble learning based works [11]. Considering that it is weak and unstable, noise can interfere with the EEG data during the data acquisition process, therefore, the quality of different EEG samples may vary accordingly. In addition, due to the multi-channel and multi-rhythm properties of EEG data, features from different frequency bands and channels should have different correlations with emotional effects. However, existing studies either ignore both the issues of sample and the importance of feature or only consider one of them.

Therefore, we establish the Retargeted Semi-supervised Regression with Robust Weights model (RSRRW). Firstly, to improve the robustness of the model, we introduce the probability weight factor, which has clear physical meaning. When the weight is 0, it means that the corresponding sample is a noise point, otherwise, it is a normal point. If a certain sample deviates from the entirety too much, the RSRRW model will choose to skip absolutely over the error as well as its contribution to noise points, thus ensuring that the model does not skew toward these outliers. In order to maintain the good discriminative ability of the model, the ϵ -dragging method is introduced to widen the gap between different classes. Additionally, feature weight factor is taken into consideration for the purpose of exploring the extent to which different dimensions of feature weight contribute to the emotion recognition task.

Consequently, the main contributions of this work to available data can be summed up.

- A new factor, probability weight, is added to each sample in the model. With the help of probability weight, the model can differentiate noise in the samples and remove the negative effects of noise. At the same time, the value of this variable also has a clear physical meaning.
- Innovatively apply the ϵ -dragging technique to the semi-supervised paradigm. It is aimed at estimating the direction matrix by gradually optimizing the label of the unlabeled samples during the learning process, which effectively increases the margin between classes.
- Compared with similar models, RSRRW has higher recognition accuracy. With the help of the feature weight factor, RSRRW can discover EEG activation patterns relating to the task, and resultantly determine comparatively significant frequency bands and predominant leads of the EEG under the current task.

Notations. In this paper, the frequency bands of EEG will be denoted by *Delta*, *Theta*, *Alpha*, *Beta*, and *Gamma*. Greek letters, such as η , λ , γ , represent the model parameters (e.g., α represents the confidence level). Matrices and vectors are respectively identified by boldface uppercase and lowercase letters, respectively. The $\ell_{2,1}$ -norm of matrix $\mathbf{W} \in \mathbb{R}^{d \times c}$ is defined as $\|\mathbf{W}\|_{2,1} = \sum_{i=1}^d \sqrt{\sum_{j=1}^c w_{ij}^2} = \sum_{i=1}^d \|\mathbf{w}^i\|_2$. \mathbf{w}^i is the i -th row of \mathbf{W} and \mathbf{w}_j is the j -th column of \mathbf{W} . The \odot is a Hadamard product operator of matrices.

The main body of this article is organized as follows. Section 2 briefly reviews the background about EEG-based emotion recognition and some related techniques. In Section 3, we introduce the formulation and optimization of the RSRRW model in detail. In Section 4, we conduct experimental studies to illustrate the performance of RSRRW. Section 5 concludes the whole paper.

2. Related Work

2.1. EEG-Based Emotion Recognition

A typical EEG-based emotion recognition process includes three stages, data preprocessing, feature extraction and model training.

Since the EEG signals are easily disturbed during acquisition, in order to provide quality data, preprocessing is necessary in the experiment. EEG data preprocessing usually includes sampling and artifact removal. In general, the sampling frequency of EEG is

usually 128~1024 Hz [12]. The higher the sample rate, the more detail the EEG can capture in the data, but it also produces more noise. Noises, in other words, the causes of interference, mainly contain electrooculography, electrocardiogram, electromyography, vascular waves and so on. Common processing methods include principal components analysis, independent component analysis, and various filtering algorithms.

Feature extraction occupies an important position in the EEG emotion recognition task, and the way of feature extraction will directly have a great impact on the performance of emotion classification. EEG feature extraction methods can be classified into four categories, namely feature extraction based on time domain, feature extraction based on frequency domain, feature extraction based on time-frequency domain and feature extraction based on spatial domain. To be specific, the feature extraction methods based on time domain include histogram analysis method [13], Hjorth parameter [14], event-related potential [15] and so on. Time domain-based analysis methods often obtain information by analyzing the geometric features of EEG signals, and the rate of information loss is very low. The frequency domain-based method mainly performs feature extraction by transforming the EEG signal from the time domain to the frequency domain, and dividing the acquired spectrum into multiple sub-bands. Its general methods include power spectral density [16], higher-order crossover [17], differential entropy (DE) [18] and higher-order spectrum [19] and so on. EEG feature extraction based on time-frequency includes short-time fourier transform [20], wavelet transform [21], wavelet packet transform [22], Hilbert-Huang transform [23] and more. Feature extraction based on spatial domain is mainly to explore the correlation features between EEG signals at different locations. The methods include common spatial patterns [24], sub-band common spatial pattern [25], rational asymmetry [26], etc. Some review studies summarized the feature extraction methods in EEG-based emotion recognition [23,27].

As for the model training process, a lot of efforts were made in the past decades. Roughly, existing models can be categorized into linear and nonlinear models. In recent studies, deep learning has attracted much attention due to its powerful nonlinear learning ability. Thammasan et al. utilized the deep belief network (DBN) to classify the EEG emotions based on the handcrafted features extracted from EEG (fractal dimension, power spectral density, etc). [28]. Li et al. used the differential entropy features of EEG to build 2D-images, and used CNN network to complete the task of emotion recognition, and achieved good results [29]. On account of the previous study, some deep learning models further unify feature extraction and recognition to form an end-to-end learning model [30,31]. Although the above deep learning models have achieved relatively good results in EEG emotion recognition tasks, most of their models stay at the stage of black-box training modes, the outcomes are poorly interpretable, and the underlying mechanism is relatively abstract [32]. Furthermore, to improve the interpretability of the model, Peng et al. proposed a unified framework named GFIL for discovering feature importance in EEG emotion recognition tasks [33]. In [34], GFIL was extended to a semi-supervised graph learning framework. Recent advances in EEG-based emotion recognition can be found in [6].

2.2. Related Techniques

2.2.1. Emotion Models

According to the existing research, affective models can be roughly divided into two categories, one is the discrete affective state model, and the other is the dimension-space affective model. In the discrete model, the emotional space covers a limited number of basic emotions. The six basic emotions are generally recognized as anger, disgust, surprise, sadness, happiness and fear, based on which other emotions are mostly considered as their combinations [35,36]. In typical dimensional space models, emotion is considered to be distributed in two-dimension or three-dimension space, and the attributes of different dimensions provide evidence for locating positions of emotions. Among these spacial models, it is common for researchers to employ valence-arousal model (VA) [37] and

valence-arousal-dominance model (VAD) [38]. In VA, the valence axis is used to measure the positive-negative degree of emotions while the arousal axis is used for measuring the intensity of emotions, as exemplified in Figure 1a. Compared with VA, VAD in Figure 1b adds the dominance axis to measure the degree to which emotions can be controlled.

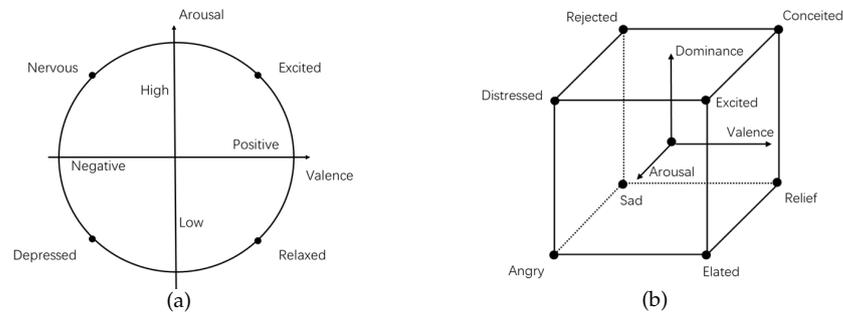


Figure 1. Emotion models VA (a) and VAD (b). (a) The VA model consists of the dimensions named valence and arousal, (b) The VAD model consists of the dimensions named valence, arousal and dominance.

2.2.2. Rescaled Least Squares Regression

Given a data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$ where d is the feature dimension, n is the number of samples. We use $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n] \in \mathbb{R}^{c \times n}$ to denote the binary indicator matrix. Specifically, if sample \mathbf{x}_i belongs to the j -th class and \mathbf{y}_i is the i -th column of \mathbf{Y} , then the j -th element of \mathbf{y}_i is 1 and all the others of \mathbf{y}_i are 0. Its mathematical form is

$$y_{ij} = \begin{cases} 1, & \mathbf{x}_j \text{ belongs to the } i\text{-th emotion;} \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

Least squares regression is a typical statistical analysis technique, which has high validity for data analysis. It has been favored in previous researches, and its semi-supervised general form is as follows

$$\begin{aligned} \min_{\mathbf{W}, \mathbf{b}, \mathbf{Y}_u} & \left\| \mathbf{W}^T \mathbf{X} + \mathbf{b} \mathbf{1}^T - \mathbf{Y} \right\|_2^2 + \gamma \|\mathbf{W}\|_F^2, \\ \text{s.t. } & \mathbf{Y}_u \geq \mathbf{0}, \mathbf{Y}_u^T \mathbf{1} = \mathbf{1}, \end{aligned} \quad (2)$$

where \mathbf{Y}_u are unlabeled samples.

In the EEG emotion recognition task, considering that EEG data has the natural characteristic of multi-channel, the importance of data from different channels can be of difference from each other in the classification task, however, objective function (2) fail to meet the requirements of differentiated feature contribution. Thus Chen et al. [39] constructed the following model,

$$\begin{aligned} \min_{\mathbf{w}, \mathbf{b}, \Theta, \mathbf{Y}_u} & \left\| \mathbf{W}^T \Theta \mathbf{X} + \mathbf{b} \mathbf{1}^T - \mathbf{Y} \right\|_2^2 + \lambda \|\mathbf{W}\|_F^2, \\ \text{s.t. } & \mathbf{Y}_u \geq \mathbf{0}, \mathbf{Y}_u^T \mathbf{1} = \mathbf{1}, \Theta = \text{diag}(\sqrt{\theta}), \theta \geq \mathbf{0}, \theta^T \mathbf{1} = \mathbf{1}, \end{aligned} \quad (3)$$

where Θ is a diagonal matrix, $\Theta_{jj} = \sqrt{\theta_j}$, θ_j describes the significance of the j -th feature. Performing a simple transformation of the objective function (3), we obtain

$$\begin{aligned} \min_{\tilde{\mathbf{W}}, \mathbf{b}, \mathbf{Y}_u} & \left\| \tilde{\mathbf{W}}^T \mathbf{X} + \mathbf{b} \mathbf{1}^T - \mathbf{Y} \right\|_2^2 + \lambda \|\tilde{\mathbf{W}}\|_{2,1}^2, \\ \text{s.t. } & \mathbf{Y}_u \geq \mathbf{0}, \mathbf{Y}_u^T \mathbf{1} = \mathbf{1}, \end{aligned} \quad (4)$$

where $\tilde{\mathbf{W}} = \Theta \mathbf{W}$.

2.2.3. Discriminative Least Squares Regression

The traditional LSR method is usually employed for data fitting. When it is used for data classification, the regression target y_i of sample x_i will be represented as a discrete value (such as +1 for the first category, −1 for the second category) or represented by one-hot encoding. In classification, data points from different classes are expected to be farther apart; however, traditional LSR cannot achieve the goal.

Therefore, Xiang et al. proposed the Discriminative Least Squares Regression (DLSR) method [40]. By introducing the ϵ -dragging technology, the class label information is embedded in the LSR framework and then the regression targets from different categories of samples move to the opposite direction, which can expand the distance between sample points from different classes. Specifically, for a positive slack variable ϵ_{ij} , if $y_{ij} = 1$, the output will become $1 + \epsilon_{ij}$; if $y_{ij} = 0$, the output will become $-\epsilon_{ij}$. This method can be naturally extended to multi-class classification task. The mathematical form of DLSR is,

$$\min_{\mathbf{W}, \mathbf{b}, \mathbf{M}} \left\| \mathbf{X}^T \mathbf{W} + \mathbf{1b}^T - \mathbf{Y} - \mathbf{B} \odot \mathbf{M} \right\|_F^2 + \lambda \|\mathbf{W}\|_F^2, \quad (5)$$

s.t. $\mathbf{M} \geq \mathbf{0}$.

The matrix \mathbf{B} is defined as,

$$B_{ij} = \begin{cases} +1, & y_{ij} = 1, \\ -1, & y_{ij} = 0, \end{cases} \quad (6)$$

which is used to determine the target dragging direction. Variable \mathbf{M} is the traction matrix and its elements are iteratively updated in model learning.

In the optimization process, ϵ -dragging constructs the loss function into two types of piecewise function, for the purpose of increasing the sample distance between different categories. Below is an example to explain it. Considering that we have a four-class classification problem, and the sample label vectors are in one-hot encoding. It is advisable to set its label of a sample from the first class as $[1, 0, 0, 0]^T$, and set the predicted value of the model under each category as $\mathbf{W}^T \mathbf{x} + \mathbf{b} = [\hat{y}_1, \hat{y}_2, \hat{y}_3, \hat{y}_4]^T$. For its category (the first category, #1), if its predicted value $\hat{y}_1 > 1$, ϵ will be updated as $\hat{y}_1 - 1$, so that the error of the sample under the category is 0. For the other categories (the second, third, fourth category, #2 #3 #4), if its predicted value is $\hat{y}_j < 0$ ($j \neq 1$), ϵ will be updated as $-\hat{y}_j$. Accordingly, it makes the error of this sample under other categories be 0. The detailed mathematical principle will be explained in the third step of optimization process in Section 3.2.

Table 1 graphically illustrates the purpose of the ϵ -dragging method. For a four-class dataset with eight sample points, the regression targets of the models (2) and (5) can be found respectively in the third and fourth columns in the Table 1.

Table 1. Illustration to the ϵ -dragging method.

Sample	Class	LSR Targets	DLSR Targets	Constraint
\mathbf{x}_1	1	$[1, 0, 0, 0]^T$	$[1 + \epsilon_{11}, -\epsilon_{12}, -\epsilon_{13}, -\epsilon_{14}]^T$	$\epsilon_{11}, \epsilon_{12}, \epsilon_{13}, \epsilon_{14} \geq 0$
\mathbf{x}_2	1	$[1, 0, 0, 0]^T$	$[1 + \epsilon_{21}, -\epsilon_{22}, -\epsilon_{23}, -\epsilon_{24}]^T$	$\epsilon_{21}, \epsilon_{22}, \epsilon_{23}, \epsilon_{24} \geq 0$
\mathbf{x}_3	2	$[0, 1, 0, 0]^T$	$[-\epsilon_{31}, 1 + \epsilon_{32}, -\epsilon_{33}, -\epsilon_{34}]^T$	$\epsilon_{31}, \epsilon_{32}, \epsilon_{33}, \epsilon_{34} \geq 0$
\mathbf{x}_4	2	$[0, 1, 0, 0]^T$	$[-\epsilon_{41}, 1 + \epsilon_{42}, -\epsilon_{43}, -\epsilon_{44}]^T$	$\epsilon_{41}, \epsilon_{42}, \epsilon_{43}, \epsilon_{44} \geq 0$
\mathbf{x}_5	3	$[0, 0, 1, 0]^T$	$[-\epsilon_{51}, -\epsilon_{52}, 1 + \epsilon_{53}, -\epsilon_{54}]^T$	$\epsilon_{51}, \epsilon_{52}, \epsilon_{53}, \epsilon_{54} \geq 0$
\mathbf{x}_6	3	$[0, 0, 1, 0]^T$	$[-\epsilon_{61}, -\epsilon_{62}, 1 + \epsilon_{63}, -\epsilon_{64}]^T$	$\epsilon_{61}, \epsilon_{62}, \epsilon_{63}, \epsilon_{64} \geq 0$
\mathbf{x}_7	4	$[0, 0, 0, 1]^T$	$[-\epsilon_{71}, -\epsilon_{72}, -\epsilon_{73}, 1 + \epsilon_{74}]^T$	$\epsilon_{71}, \epsilon_{72}, \epsilon_{73}, \epsilon_{74} \geq 0$
\mathbf{x}_8	4	$[0, 0, 0, 1]^T$	$[-\epsilon_{81}, -\epsilon_{82}, -\epsilon_{83}, 1 + \epsilon_{84}]^T$	$\epsilon_{81}, \epsilon_{82}, \epsilon_{83}, \epsilon_{84} \geq 0$

3. Method

In this section, we present the model formulation of RSRRW firstly and then its detailed optimization procedure.

In EEG-based semi-supervised emotion recognition problem, we use the matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$ to represent the EEG data. \mathbf{X} consists of two subsets, one is labeled samples $\mathbf{X}_l \in \mathbb{R}^{d \times n_l}$, corresponding to the labels $\mathbf{Y}_l \in \mathbb{R}^{c \times n_l}$. The other one is unlabeled samples $\mathbf{X}_u \in \mathbb{R}^{d \times n_u}$, correspondingly to the labels $\mathbf{Y}_u \in \mathbb{R}^{c \times n_u}$. d is the feature dimension of EEG samples, c is the number of emotion categories, n_l is the number of labeled samples, n_u is the number of unlabeled samples and $n = n_l + n_u$.

The target of this algorithm is to accurately predict \mathbf{Y}_u using the given data $\mathbf{X} = [\mathbf{X}_l, \mathbf{X}_u]$ and \mathbf{Y}_l .

3.1. Model Formulation

The RSRRW model framework is shown in Figure 2, in which the model part mainly includes four components, namely sample weight learning, feature weight learning, ϵ -dragging process, coefficient matrix learning. The functions of each part are summarized as follows. (1) Based on sample weight learning, add a probability weight to each sample point in the data. When the sample is a normal point, its probability weight is 1, otherwise, its probability weight is 0. (2) Based on feature weight learning, differentiate the contributions of features. (3) In the ϵ -dragging process, the regression targets of different categories are forced to move in opposite directions, expanding the distance between categories. (4) Coefficient matrix learning, using the least squares method to learn \mathbf{W} and \mathbf{b} .

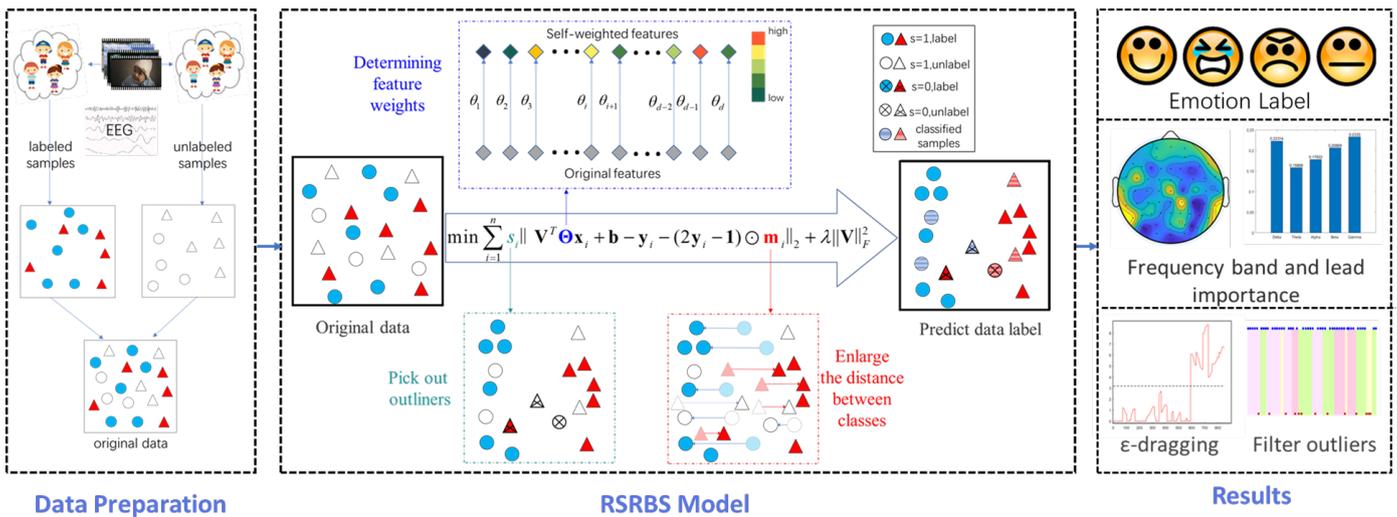


Figure 2. The general framework of RSRRW model.

The ϵ -dragging strategy was proposed in [40], and its mathematical form is shown in

$$\min_{\mathbf{W}, \mathbf{b}, \mathbf{M}} \left\| \mathbf{V}^T \mathbf{X} + \mathbf{b} \mathbf{1}^T - \mathbf{Y} - \mathbf{B} \odot \mathbf{M} \right\|_F^2 + \lambda \|\mathbf{V}\|_F^2, \quad (7)$$

$$\text{s.t. } \mathbf{M} \geq \mathbf{0},$$

where \mathbf{V} is original coefficient matrix.

Considering that EEG data is a multi-channel time series data, features from different frequency bands and leads have different contributions to specific tasks, and the objective function (7) cannot distinguish the importance of different features, which is defective. Therefore, we introduce the sample feature weight representation matrix Θ to describe the importance of sample features, and construct the following objective function,

$$\min_{\mathbf{W}, \mathbf{b}, \mathbf{M}} \left\| \mathbf{V}^T \Theta \mathbf{X} + \mathbf{b} \mathbf{1}^T - \mathbf{Y} - \mathbf{B} \odot \mathbf{M} \right\|_F^2 + \lambda \|\mathbf{V}\|_F^2, \quad (8)$$

$$\text{s.t. } \mathbf{M} \geq \mathbf{0}, \Theta = \text{diag}(\sqrt{\theta}), \theta \geq 0, \theta^T \mathbf{1} = 1.$$

Considering the weak and unstable characteristics of EEG data, the quality of different samples varies greatly. In order to improve the robustness of the model to samples, we introduce the sample probability weight \mathbf{s} , so that the model can automatically filter outliers during the training process. The optimized objective function is

$$\begin{aligned} & \min \sum_{i=1}^n s_i \left\| \mathbf{V}^T \Theta \mathbf{x}_i + \mathbf{b} - \mathbf{y}_i - \mathbf{d}_i \odot \mathbf{m}_i \right\|_2 + \lambda \|\mathbf{V}\|_F^2, \\ & \text{s.t. } \mathbf{s}^T \mathbf{1} = k, 0 \leq s_i \leq 1, \Theta = \text{diag}(\sqrt{\theta}), \theta \geq 0, \theta^T \mathbf{1} = 1, \mathbf{m}_i \geq \mathbf{0}. \end{aligned} \tag{9}$$

In model (9), \mathbf{s} is the sample weight, k is the number of normal points in the sample, \mathbf{d}_i is the i -th column of \mathbf{B} . However, model (9) can only be used in a supervised setting, here we extend it to a semi-supervised setting,

$$\begin{aligned} & \min \sum_{i=1}^n s_i \left\| \mathbf{V}^T \Theta \mathbf{x}_i + \mathbf{b} - \mathbf{y}_i - (2\mathbf{y}_i - \mathbf{1}) \odot \mathbf{m}_i \right\|_2 + \lambda \|\mathbf{V}\|_F^2, \\ & \text{s.t. } \mathbf{s}^T \mathbf{1} = k, 0 \leq s_i \leq 1, \Theta = \text{diag}(\sqrt{\theta}), \theta \geq 0, \theta^T \mathbf{1} = 1, \\ & \quad \mathbf{Y} = [\mathbf{Y}_l; \mathbf{Y}_u], \mathbf{Y}_u \geq \mathbf{0}, \mathbf{Y}_u^T \mathbf{1} = \mathbf{1}, \mathbf{m}_i \geq \mathbf{0}. \end{aligned} \tag{10}$$

By setting $\mathbf{W} = \Theta \mathbf{V}$, we have $\mathbf{V} = \Theta^{-1} \mathbf{W}$ and then the above objective function can be rewritten as

$$\begin{aligned} & \min \sum_{i=1}^n s_i \left\| \mathbf{W}^T \mathbf{x}_i + \mathbf{b} - \mathbf{y}_i - (2\mathbf{y}_i - \mathbf{1}) \odot \mathbf{m}_i \right\|_2 + \lambda \left\| \Theta^{-1} \mathbf{W} \right\|_F^2, \\ & \text{s.t. } \mathbf{s}^T \mathbf{1} = k, 0 \leq s_i \leq 1, \Theta = \text{diag}(\sqrt{\theta}), \theta \geq 0, \theta^T \mathbf{1} = 1, \\ & \quad \mathbf{Y} = [\mathbf{Y}_l; \mathbf{Y}_u], \mathbf{Y}_u \geq \mathbf{0}, \mathbf{Y}_u^T \mathbf{1} = \mathbf{1}, \mathbf{m}_i \geq \mathbf{0}. \end{aligned} \tag{11}$$

When $\mathbf{s}, \mathbf{W}, \mathbf{b}, \mathbf{M}$ are fixed, because $\Theta = \text{diag}(\sqrt{\theta})$ and $\theta^T \mathbf{1} = 1$, the second term of objective function (11) can be rewritten as

$$\min_{\theta \geq 0, \theta^T \mathbf{1} = 1} \sum_{j=1}^d \frac{\|\mathbf{w}^j\|_2^2}{\theta_j}. \tag{12}$$

According to the Lagrange multiplier method, we write the Lagrangian function of (12) about θ_j and set its derivative to θ_j to 0, then we can get

$$\theta_j = \frac{\|\mathbf{w}^j\|_2}{\sum_{l=1}^d \|\mathbf{w}^l\|_2}. \tag{13}$$

With the above solution to θ , objective function (12) is equivalent to

$$\min_{\theta \geq 0, \theta^T \mathbf{1} = 1} \|\mathbf{W}\|_{2,1}^2. \tag{14}$$

Now model (11) can be rewritten as

$$\begin{aligned} & \min \sum_{i=1}^n s_i \left\| \mathbf{W}^T \mathbf{x}_i + \mathbf{b} - \mathbf{y}_i - (2\mathbf{y}_i - \mathbf{1}) \odot \mathbf{m}_i \right\|_2 + \lambda \|\mathbf{W}\|_{2,1}^2, \\ & \text{s.t. } \mathbf{s}^T \mathbf{1} = k, 0 \leq s_i \leq 1, \mathbf{Y} = [\mathbf{Y}_l; \mathbf{Y}_u], \mathbf{Y}_u \geq \mathbf{0}, \mathbf{Y}_u^T \mathbf{1} = \mathbf{1}, \mathbf{m}_i \geq \mathbf{0}. \end{aligned} \tag{15}$$

3.2. Model Optimization

There are five variables in model (15) that need to be optimized, namely $\mathbf{W}, \mathbf{b}, \mathbf{Y}_u, \mathbf{s}, \mathbf{M}$. For this problem, we designed a joint iterative optimization algorithm, which divides

the above problem into four sub-problems to solve them separately and performs multiple iterations until convergence.

- Update \mathbf{W}, \mathbf{b} by fixing \mathbf{Y}_u, \mathbf{s} and \mathbf{M} .

Performing simple variable substitution on objective function (15), we can get

$$\min_{\mathbf{W}, \mathbf{b}} \sum_{i=1}^n s_i \mu(e_i) + \lambda \|\mathbf{W}\|_{2,1}^2, \quad (16)$$

where $e_i = \|\mathbf{W}^T \mathbf{x}_i + \mathbf{b} - \mathbf{z}_i\|_2^2$, \mathbf{z}_i is the i -th column of \mathbf{Z} , $\mathbf{Z} = \mathbf{Y} + (2\mathbf{Y} - \mathbf{1}\mathbf{1}^T) \odot \mathbf{M}$ and μ is the function of e , $\mu(e) = e^{\frac{1}{2}}$.

We introduce an effective algorithm [41] to solve objective function (16). The general problem is as follows

$$\min_{x \in \mathbb{C}} \sum_i h_i(g_i(x)) + f(x). \quad (17)$$

The form of x and $g(x)$ could be scalar, vector or matrix. Algorithm 1 describes the detailed procedures. By comparison of (16) and (17), we can find $\mu(e)$, $e_i(\mathbf{W}, \mathbf{b})$, $\lambda \|\mathbf{W}\|_{2,1}^2$ in problem (17) as $h(x)$, $g_i(x)$, and $f(x)$ in (17), respectively.

Algorithm 1 Solution of (17).

Input: Initialize $x \in \mathbb{C}$

Output: The optimal x

- 1: **while** not converged **do**
 - 2: Calculated $d_i = h'_i(g_i(x))$;
 - 3: Solve the following minimization problem $\min_{x \in \mathbb{C}} \sum_i \text{Tr}(d_i^T g_i(x)) + f(x)$;
 - 4: **end while**
-

Firstly, we need to calculate d_i . The formulation of d_i is

$$d_i = \mu'(e_i) = \frac{1}{2\|\mathbf{W}^T \mathbf{x}_i + \mathbf{b} - \mathbf{z}_i\|_2}. \quad (18)$$

Secondly, we need to solve

$$\min_{\mathbf{W}, \mathbf{b}} \sum_{i=1}^n s_i d_i \|\mathbf{W}^T \mathbf{x}_i + \mathbf{b} - \mathbf{z}_i\|_2^2 + \lambda \|\mathbf{W}\|_{2,1}^2. \quad (19)$$

To facilitate subsequent calculations, we rewrite (19) into matrix form

$$J = \text{Tr}\left((\mathbf{W}^T \mathbf{X} + \mathbf{b}\mathbf{1}^T - \mathbf{Z})\mathbf{\Lambda}(\mathbf{W}^T \mathbf{X} + \mathbf{b}\mathbf{1}^T - \mathbf{Z})^T\right) + \lambda \|\mathbf{W}\|_{2,1}^2, \quad (20)$$

where $\mathbf{\Lambda} = \mathbf{S}\mathbf{D}$, $\mathbf{S} = \text{diag}(s_1, s_2, \dots, s_n)$, $\mathbf{D} = \text{diag}(d_1, d_2, \dots, d_n)$. Considering that (19) is an unconstrained optimization problem, we use gradient descent to solve it. Take the partial derivative of J with respect to \mathbf{b} , we have

$$\frac{\partial J}{\partial \mathbf{b}} = 2\left(\mathbf{W}^T \mathbf{X} + \mathbf{b}\mathbf{1}^T - \mathbf{Z}\right)\mathbf{\Lambda}\mathbf{1}. \quad (21)$$

Let (21) take 0, we can get the optimization formula of \mathbf{b} as

$$\mathbf{b} = \frac{\mathbf{Z}\mathbf{\Lambda}\mathbf{1} - \mathbf{W}^T \mathbf{X}\mathbf{\Lambda}\mathbf{1}}{\mathbf{1}^T \mathbf{\Lambda}\mathbf{1}}. \quad (22)$$

Replacing \mathbf{b} in (20) with the result obtained in (22) and simplifying it, we can get

$$J = \text{Tr}\left((\mathbf{W}^T \mathbf{X} - \mathbf{Z})\mathbf{K}(\mathbf{W}^T \mathbf{X} - \mathbf{Z})^T\right) + \lambda \sum_{i=1}^d \sqrt{\|\mathbf{w}^i\| + \delta}. \quad (23)$$

where $\mathbf{K} = \left(\mathbf{I} - \frac{\Lambda \mathbf{1} \mathbf{1}^T}{\mathbf{1}^T \Lambda \mathbf{1}}\right) \Lambda \left(\mathbf{I} - \frac{\Lambda \mathbf{1} \mathbf{1}^T}{\mathbf{1}^T \Lambda \mathbf{1}}\right)^T$ and $\mathbf{K} = \mathbf{K}^T$. Similarly, (23) is also an unconstrained optimization problem, so we take the partial derivative of (23) in terms of \mathbf{W} , then we can get

$$\frac{\partial J}{\partial \mathbf{W}} = 2(\mathbf{X}\mathbf{K}\mathbf{X}^T \mathbf{W} - \mathbf{X}\mathbf{K}\mathbf{Z}^T + \lambda \mathbf{Q}\mathbf{W}), \quad (24)$$

where $\mathbf{Q} \in \mathbb{R}^{d \times d}$ is a diagonal matrix whose i -th diagonal elements is

$$q_{ii} = \frac{\sum_{i=1}^d \sqrt{\|\mathbf{w}^i\|_2^2 + \delta}}{\sqrt{\|\mathbf{w}^i\|_2^2 + \delta}}, \quad (25)$$

and δ is a fixed minimal constant value. By making the partial derivative value 0, the expression of \mathbf{W} is obtained

$$\mathbf{W} = (\mathbf{X}\mathbf{K}\mathbf{X}^T + \lambda \mathbf{Q})^{-1} \mathbf{X}\mathbf{K}\mathbf{Z}^T. \quad (26)$$

- Update \mathbf{s} by fixing $\mathbf{Y}_u, \mathbf{W}, \mathbf{b}$ and \mathbf{M} .

The corresponding objective function in terms of variable \mathbf{s} is

$$\min_{\mathbf{s}} \sum_{i=1}^n s_i e_i \text{ s.t. } \mathbf{s} \geq \mathbf{0}, \mathbf{s}^T \mathbf{1} = k. \quad (27)$$

where e_i measures the approximation error on sample \mathbf{x}_i by the ℓ_2 -norm. It is interesting that parameter k and the variable \mathbf{s} are closely related in model (27). To be specific, the value of k represents the number of elements of one in \mathbf{s} , corresponding to the number of normal samples involved in model training. By calculating and ranking the loss e_i for each sample, the optimal \mathbf{s} can be obtained, in which the weights of the first k samples with the smallest errors are all set as one and the remaining values are zeros. This regularity is depicted by the following Theorem 1.

Theorem 1. *The optimal \mathbf{s} in problem (27) is a binary vector, in which the corresponding weights of the first k samples with the smallest errors are one and the others are zero.*

Proof 1. Suppose there is another weight vector \mathbf{s}' that satisfies

$$\sum_{j=1}^k s'_j + \sum_{j=k+1}^n s'_j = k, \quad (28)$$

s.t.

$$\sum_{j=1}^n s'_j \mu(e^{(j)}) \leq \sum_{j=1}^n s'_j \mu(e^{(j)}). \quad (29)$$

Firstly, we sort each sample in ascending order of error, i.e.,

$$\mu(e^{(1)}) \leq \mu(e^{(2)}) \leq \dots \leq \mu(e^{(j)}) \leq \dots \leq \mu(e^{(n)}). \quad (30)$$

After a simple split on (29), we get

$$\sum_{j=1}^k s'_j \mu(e^{(j)}) + \sum_{j=k+1}^n s'_j \mu(e^{(j)}) \leq \sum_{j=1}^k \mu(e^{(j)}). \quad (31)$$

By moving the first term to the left of the inequality sign to the right, we have

$$\sum_{j=k+1}^n s'_j \mu(e^{(j)}) \leq \sum_{j=1}^k (1 - s'_j) \mu(e^{(j)}). \tag{32}$$

According to (30), we can get

$$\sum_{j=k+1}^n s'_j \mu(e^{(j)}) \geq (1 - s'_j) \mu(e^{(k+1)}). \tag{33}$$

After combining (32) and formula (33), we can get

$$\sum_{j=1}^k (1 - s'_j) \mu(e^{(k+1)}) \leq \sum_{j=1}^k (1 - s'_j) \mu(e^{(j)}). \tag{34}$$

When inequality signs in (30) are not taken equal at the same time, Equation (34) obviously does not hold, so the assumption is logically impossible. Therefore, the optimal \mathbf{s} in problem (27) is a binary vector, in which the corresponding weights of the first k samples with the smallest errors are one and the others are zero. \square

- Update \mathbf{M} by fixing $\mathbf{Y}_u, \mathbf{W}, \mathbf{b}$ and \mathbf{s} .

From problem (15), the objective function in terms of variable \mathbf{M} is

$$\begin{aligned} \min_{\mathbf{M} \geq 0} & \left\| \mathbf{W}^T \mathbf{X} + \mathbf{b} \mathbf{1}^T - \mathbf{Y} - (2\mathbf{Y} - \mathbf{1} \mathbf{1}^T) \odot \mathbf{M} \right\|_2 \\ & \triangleq \left\| \mathbf{P} - \mathbf{B} \odot \mathbf{M} \right\|_2, \end{aligned} \tag{35}$$

where $\mathbf{P} \triangleq \mathbf{W}^T \mathbf{X} + \mathbf{b} \mathbf{1}^T - \mathbf{Y}$ and $\mathbf{B} \triangleq 2\mathbf{Y} - \mathbf{1} \mathbf{1}^T$. It is easily verified that matrix \mathbf{B} is similar to the one defined in [40] in supervised learning. Considering that in the optimization process of (35), the optimization process of each element is independent of each other, therefore, for an element in the matrix, we can convert (35) into the following form

$$\min_{M_{ij} \geq 0} (P_{ij} - B_{ij} M_{ij})^2. \tag{36}$$

Obviously, the solution to the above optimization problem is

$$M_{ij} = \max\left(\frac{P_{ij}}{B_{ij}}, 0\right). \tag{37}$$

Then, the solution to (35) is

$$\mathbf{M} = \max(\mathbf{P} ./ \mathbf{B}, \mathbf{0}). \tag{38}$$

In the formula (38), we find that the ϵ -dragging method essentially encourages the predicted value \hat{y} to migrate towards $1 + \epsilon_{ij}$ or $-\epsilon_{ij}$ by constructing the loss function as a two-class piecewise function. Below we will use an example to detail the updating method during the optimization process. One point to note is that in the algorithm, we have implemented the ϵ -dragging method through matrix \mathbf{M} and matrix \mathbf{B} , and the elements in \mathbf{M} can be regarded as ϵ_{ij} . To maintain the consistence of symbols, the element is represented as M_{ij} .

For the sample \mathbf{x}_i , we set the emotional label as $\mathbf{y}_i = [1, 0, 0, 0]^T$. In the model (35), its predicted value is $\hat{\mathbf{y}}_i = \mathbf{W}^T \mathbf{x}_i + \mathbf{b} = [\hat{y}_1, \hat{y}_2, \hat{y}_3, \hat{y}_4]^T$. In order to distinguish it from \mathbf{b}

in the LSR, we set the i -th column of the direction matrix \mathbf{B} to be $\mathbf{B} = [\tilde{\mathbf{b}}_1, \tilde{\mathbf{b}}_2, \dots, \tilde{\mathbf{b}}_n]$. The direction vector of \mathbf{x}_i is $\tilde{\mathbf{b}}_i = [1, -1, -1, -1]^T$. The error generated by the sample is,

$$\begin{aligned} \text{error}_i^2 &= \left\| \hat{\mathbf{y}}_i - \mathbf{y}_i - \tilde{\mathbf{b}}_i \odot \mathbf{m}_i \right\|_2^2 \\ &= (\hat{y}_1 - (1 + M_{1,i}))^2 + (\hat{y}_2 + M_{2,i})^2 + (\hat{y}_3 + M_{3,i})^2 + (\hat{y}_4 + M_{4,i})^2. \end{aligned} \tag{39}$$

Based on (39), the error_i^2 consists of four items, which can be divided into two categories. One is the error generated by the predicted value of the category to which the sample belongs, corresponding to the first item in (39); the other is the error generated by the predicted value of other categories, corresponding to the second, third, and fourth items in (39). We can easily conclude the following findings.

There are two steps in the calculation of the error of the category of the sample (that is, the first item in (39)). Firstly, $M_{1,i}$ can be represented by the following piecewise function,

$$M_{1,i} = \begin{cases} \hat{y}_1 - 1 & \hat{y}_1 \geq 1 \\ 0 & \hat{y}_1 < 1 \end{cases} \tag{40}$$

Then the corresponding error is calculated as,

$$\text{error}_{1,i}^2 = \begin{cases} 0 & \hat{y}_1 \geq 1 \\ (\hat{y}_1 - 1)^2 & \hat{y}_1 < 1 \end{cases} \tag{41}$$

The formula (41) indicates that when the predicted value \hat{y}_1 of the category to which the sample belongs is greater than 1, the error generated by \hat{y}_1 is 0. However, in the traditional LSR, an error will still be generated at this time, and its value is $(\hat{y}_1 - 1)^2$. The ϵ -dragging method uses the no-negative slack variable $M_{1,i}$, offsetting the corresponding error. Thereby it encourages the predicted value \hat{y}_1 of the category to move in the direction $1 + \epsilon$.

Similarly, for the samples of the other categories (i.e., items 2, 3, and 4 in (39)), the analysis is presented below. The variable $M_{c_j,i}$ ($c_j \neq 1$) can be represented as

$$M_{c_j,i} = \begin{cases} 0 & \hat{y}_{c_j} \geq 0, \\ -\hat{y}_{c_j} & \hat{y}_{c_j} < 0, \end{cases} \tag{42}$$

Secondly, the corresponding error is calculated as,

$$\text{error}_{c_j,i}^2 = \begin{cases} \hat{y}_{c_j}^2 & \hat{y}_{c_j} \geq 0, \\ 0 & \hat{y}_{c_j} < 0. \end{cases} \tag{43}$$

The formula (43) indicates that when the predicted value \hat{y}_{c_j} of other categories is less than 0, the error generated by \hat{y}_{c_j} is 0. However, in LSR, an error will still be generated at this time, and its value is $(\hat{y}_{c_j} - 1)^2$. The ϵ -dragging method uses the no-negative slack variable $M_{c_j,i}$ to offset the corresponding error. Thereby it encourages the predicted value of other category \hat{y}_{c_j} to move in the direction $-\epsilon$.

- Update \mathbf{Y}_u by fixing $\mathbf{s}, \mathbf{W}, \mathbf{b}$ and \mathbf{M} .

Rows in \mathbf{Y}_u are independent of each other, and then we can optimize \mathbf{Y}_u in a column-wise manner. To be specific, for $i = l + 1, l + 2, \dots, l + u$, the corresponding objective function in terms of \mathbf{y}_i is

$$\begin{aligned} \min_{\mathbf{y}_i \geq 0, \mathbf{y}_i^T \mathbf{1} = 1} & \left\| \mathbf{W}^T \mathbf{x}_i + \mathbf{b} + \mathbf{m}_i - \mathbf{y}_i \odot (\mathbf{1} + 2\mathbf{m}_i) \right\|_2 \\ & \triangleq \left\| \mathbf{a}_i - \mathbf{y}_i \right\|_2, \end{aligned} \tag{44}$$

where $\mathbf{a} \triangleq (\mathbf{W}^T \mathbf{x}_i + \mathbf{b} + \mathbf{m}_i) ./ (\mathbf{1} + 2\mathbf{m}_i)$. To simplify the derivation, we optimize the equivalent form of problem (44) as

$$\min_{\mathbf{y}_i \geq 0, \mathbf{y}_i^T \mathbf{1} = 1} \frac{1}{2} \|\mathbf{a}_i - \mathbf{y}_i\|_2^2. \tag{45}$$

Problem (45) can be solved by the Lagrange multiplier method combined with KKT condition. The corresponding Lagrangian function is

$$\mathcal{L} = \frac{1}{2} \|\mathbf{y}_i - \mathbf{a}_i\|_2^2 - \eta (\mathbf{y}_i^T \mathbf{1} - 1) - \boldsymbol{\beta}^T \mathbf{y}_i, \tag{46}$$

where η and $\boldsymbol{\beta} \in \mathbb{R}^c$ are the Lagrangian multipliers respectively in scalar and vector forms. Assume that $\tilde{\mathbf{y}}_i$ is the optimal solution and the associated optimal Lagrangian multipliers are $\tilde{\eta}$ and $\tilde{\boldsymbol{\beta}}$. According to the KKT condition, we have

$$\begin{cases} \tilde{y}_j \geq 0, \\ \tilde{\beta}_j \geq 0, \\ \tilde{y}_j \tilde{\beta}_j = 0, \\ \tilde{y}_j - a_j - \tilde{\eta} - \tilde{\beta}_j = 0, \end{cases} \tag{47}$$

for each $j \in \{1, 2, \dots, c\}$. The last expression in (47) can be written in vector form as

$$\mathbf{y} - \mathbf{a} - \tilde{\eta} \mathbf{1} - \tilde{\boldsymbol{\beta}} = \mathbf{0}. \tag{48}$$

Because $\mathbf{y}^T \mathbf{1} = 1$, the parameter a in (48) can be written as

$$\tilde{\eta} = \frac{1 - \mathbf{1}^T \mathbf{a} - \mathbf{1}^T \tilde{\boldsymbol{\beta}}}{c}. \tag{49}$$

Substituting (49) into (48), we have

$$\tilde{\mathbf{y}} = \mathbf{a} - \frac{\mathbf{1}\mathbf{1}^T}{c} \mathbf{a} + \frac{1}{c} \mathbf{1} - \frac{\mathbf{1}^T \tilde{\boldsymbol{\beta}}}{c} \mathbf{1} + \tilde{\boldsymbol{\beta}}. \tag{50}$$

Denote $\bar{\boldsymbol{\beta}} = \frac{\mathbf{1}^T \tilde{\boldsymbol{\beta}}}{c}$ and $\mathbf{g} = \mathbf{a} - \frac{\mathbf{1}\mathbf{1}^T}{c} \mathbf{a} + \frac{1}{c} \mathbf{1}$, (50) can be rewritten as

$$\tilde{\mathbf{y}} = \mathbf{g} + \tilde{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}} \mathbf{1}. \tag{51}$$

Therefore we have

$$\tilde{y}_j = g_j + \tilde{\beta}_j - \bar{\beta}. \tag{52}$$

Through (47) and (52), we can get $g_j + \tilde{\beta}_j - \bar{\beta} = (g_j - \bar{\beta})_+$, where

$$(f(\cdot))_+ = \begin{cases} 0 & f(\cdot) < 0, \\ f(\cdot) & \text{otherwise.} \end{cases} \tag{53}$$

Therefore, (52) can be written as

$$\tilde{y}_j = (g_j - \bar{\beta})_+. \tag{54}$$

At this time, if the optimal $\bar{\beta}$ can be determined, according to (54), the optimal $\tilde{\mathbf{y}}$ can also be determined. Similar to (54), (52) can be rewritten as $\tilde{\beta}_j = \tilde{y}_j + \bar{\beta} - g_j$ such that $\tilde{\beta}_j = (\bar{\beta} - g_j)_+$. Therefore, $\bar{\beta}$ can be calculated as

$$\bar{\beta} = \frac{1}{c} \sum_{j=1}^c (\bar{\beta} - g_j)_+. \quad (55)$$

According to the constraint $\mathbf{y}^T \mathbf{1} = 1$ and (54), we can define the following function

$$f(\bar{\beta}) = \sum_{j=1}^c (g_j - \bar{\beta}) - 1, \quad (56)$$

and the optimal $\bar{\beta}$ should satisfy $f(\bar{\beta}) = 0$. When (56) equals 0, the optimal $\bar{\beta}$ can be calculated by the Newton method

$$\bar{\beta}_{k+1} = \bar{\beta}_k - \frac{f(\bar{\beta}_k)}{f'(\bar{\beta}_k)}. \quad (57)$$

Finally, the optimization procedure to sub-problem (44) is listed in Algorithm 2 and the procedure to solve problem (15) is summarized in Algorithm 3.

Algorithm 2 The algorithm to solve sub-problem (45).

Input: vector $\mathbf{a}_i \in \mathbb{R}^c$

Output: variable $\mathbf{x} \in \mathbb{R}^c$

- 1: Compute $\mathbf{g} = \mathbf{a}_i - \frac{\mathbf{1}\mathbf{1}^T}{c} + \frac{1}{c}\mathbf{1}$;
 - 2: Use Newton's method to obtain the root $\bar{\beta}$ of (56);
 - 3: Obtain the optimal solution by (54);
-

Algorithm 3 The optimization algorithm of RSRRW.

Input: Labeled EEG data $\mathbf{X}_l \in \mathbb{R}^{d \times n_l}$ and its corresponding label matrix $\mathbf{Y}_l \in \mathbb{R}^{c \times n_l}$, unlabeled EEG data $\mathbf{X}_u \in \mathbb{R}^{d \times n_u}$, parameters λ ;

Output: The estimated label $\mathbf{Y}_u \in \mathbb{R}^{c \times n_u}$.

- 1: Initialize \mathbf{W} randomly. Initialize each element of \mathbf{Y}_u as $\frac{1}{c}$. Initialize each element of \mathbf{M} as 0. Initialize each element of \mathbf{s} as 1;
 - 2: **while** not converged **do**
 - 3: Update \mathbf{b} via (22);
 - 4: Update \mathbf{W} via (26);
 - 5: Update \mathbf{s} by solve (27);
 - 6: Update \mathbf{M} via (38);
 - 7: Update \mathbf{Y}_u by solving (45) for each $i \binom{n_l+n_u}{n_l+1}$;
 - 8: **end while**
-

3.3. Discussion on RSRRW

In this section, we will discuss the variable \mathbf{s} and ϵ -dragging methods in RSRRW. In addition, the complexity of the RSRRW model will also be briefly analyzed.

- We know that $s_i \binom{n}{i=1}$ reflects the importance level of the i -th sample. Through proof in Section 3.2, we can know that \mathbf{s} is a binary vector. When sample i is a noise point, the loss caused by it is 0, that is, sample \mathbf{x}_i will not get involved in the training process. On the contrary, when the sample \mathbf{x}_i is a normal point, its weight is 1, meaning that it contributes to the training. Therefore, with the help of the sample weight \mathbf{s} , RSRRW can automatically find outliers in the data and remove them, thereby preventing the model from skewing towards outliers. Comparison experiments suggests that RSRRW

performs better than DLSR. At the same time, through the visualization of \mathbf{s} in the time dimension, we can intuitively observe the time distribution of outliers during the experiment, which can also provide some references for future experimental design. We will elaborate on this in Section 4.3 below.

- In [40], the ϵ -dragging method was applied to supervised learning tasks and \mathbf{B} was obtained by calculating the label of the training set samples, so as to obtain the dragging direction of each sample and achieve the purpose of expanding the distance between classes. Considering that in the EEG emotion recognition task, there are usually few labeled data samples, and the distribution of data from different sessions is quite different. Therefore, it may be hard to obtain ideal results by wholly relying on the supervised learning paradigm to complete the emotion recognition task. Therefore, we extend the ϵ -dragging method to the semi-supervised learning process and estimate the dragging direction of each sample through the unlabeled sample category probability obtained in the optimization process, realizing semi-supervised ϵ -dragging method. Ideally, if \mathbf{Y}_u can be estimated accurately, then we can get the exact dragging direction of each sample. In the following experiments, we will intuitively show the performance of ϵ -dragging method.
- We analyze the asymptotic time complexity of our model in this section. Firstly, in the optimization of \mathbf{b} , the asymptotic time complexity is $\mathcal{O}(n^2c + ndc + n^2c)$. Secondly, in the optimization of \mathbf{W} , the asymptotic time complexity of $(\mathbf{X}\mathbf{K}\mathbf{X}^T + \lambda\mathbf{Q})^{-1}$ is $\mathcal{O}(n^2d + d^2c + d^3)$, the asymptotic time complexity of $\mathbf{X}\mathbf{K}\mathbf{Z}^T$ is $\mathcal{O}(n^2d)$. To calculate \mathbf{K} , considering that it depends on the operation of the diagonal matrix and the unit matrix, so its time complexity is obviously lower than $\mathcal{O}(n^2d)$. Then, as for the improvement of \mathbf{s} , its asymptotic time complexity is $\mathcal{O}(n \log n)$. Finally, in the optimizing process of \mathbf{M} , its asymptotic time complexity is $\mathcal{O}(dcn)$. In the optimization process of \mathbf{Y}_u , since $\mathbf{W}^T\mathbf{X}$ has been completed in the process of updating \mathbf{M} , its asymptotic time complexity is $\mathcal{O}(uc)$, where u is the number of unlabeled samples. In general, $n \approx u > d \gg c$. Therefore, the overall asymptotic time complexity of the RSRRW algorithm model is $\mathcal{O}(n^2dt)$, where t is the number of iterations.

4. Experiments

4.1. Data Sets

The SEED-IV dataset contains EEG signals collected from subjects while they're watching movie clips. It selects 72 film clips with highly emotional contents as signals for emotional induction. A total of 15 healthy right-handed subjects (including 7 males and 8 females) participated in the experiment, and each one needed to take exams in three different sessions on different days. The four emotional states evoked by the video are happy, fearful, neutral, and sad. In each session, one affective state corresponds to six trials. The experimental process of each session is shown in Figure 3. There will be a 5 s start prompt before each video clip is played, and the subjects will get a 45 s self-assessment after the playback ends.

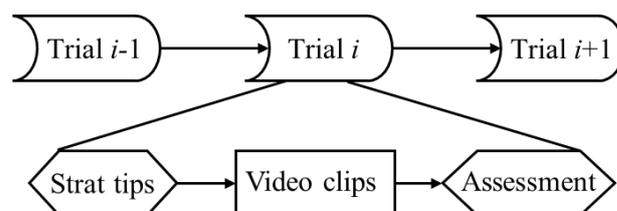


Figure 3. Experimental protocol for SEED-IV [42].

The EEG acquisition equipment used during the experiment included the ESI Neuroscan system and a 62-electrode cap that conforms to the international 10–20 placement standard. In the experiments mentioned in this paper, we used differential entropy (DE) features extracted from five frequency bands, namely *Delta* (1–3 Hz), *Theta* (4–7 Hz), *Alpha*

(8–13 Hz), *Beta* (14–30 Hz) and *Gamma* (31–50 Hz). The sample x_i is formed by concatenating the values of the 62 leads in each of the 5 frequency bands, resulting in a dimension of 310. More details can be found in [42].

4.2. Experimental Settings

During the experiments, we performed the cross-session emotion recognition task in chronological order. In other words, each subject is supposed to complete three tasks, including “session 1–session 2”, “session 1–session 3”, “session 2–session 3”. In the “session 1–session 2” task, the samples in session 1 are labeled data, and the samples in session 2 are unlabeled data. During the model training process, the samples from session 2 will be gradually labeled.

We compare the RSRRW algorithm with five closely related models, including the traditional semi-supervised linear square regression model (sLSR), semi-supervised support vector machine model (sSVM), the optimized rescaled linear square regression model (RLSR) [39], discriminative linear square regression model (DLSR) [40], and robust semi-supervised least squares regression model (RSLSR) [43]. RLSR is an optimization of sLSR, which adds a feature weight factor to describe feature importance on the basis of sLSR. DLSR is another optimization of sLSR. On the basis of LSR, it adds the ϵ -dragging method to increase the distance between categories. In sSVM, we use a linear kernel. RSLSR adds sample weights on the basis of sLSR, which can filter out outliers in samples. In the above five methods, there are hyperparameters that need to be adjusted (λ in sLSR, sSVM, RLSR, DLSR, RSLSR, RSRRW; k in RSRRW, RLSR). In this experiment, the adjustment range of λ is $\{2^{-10}, 2^{-9}, \dots, 2^{10}\}$. For the RSSL and RSRRW, the adjustment range of k is $\{[0.8a], [0.81a], \dots, a\}$, where a is the number of samples. The iteration stopping condition is that the number of iterations reaches 100 or the rate of change of the objective function $obj_\delta = \frac{obj^{(t+1)} - obj^{(t)}}{obj^{(t)}}$ is less than or equal to 1×10^{-5} . The typical number of iterations of RSRRW is around 70.

4.3. Results and Analysis

Tables 2–4 shows the recognition accuracy of the above models in the cross-session EEG emotional states recognition tasks. Among them, Table 2 shows the experimental results with session 1 serving as labeled data and session 2 as unlabeled data. Table 3 shows the experimental results with the labeled data in session 1 and the unlabeled data in session 3. Table 4 shows the experimental results with session 2 as labeled data and session 3 as unlabeled data. The highest accuracy in each group is marked in bold.

Table 2. Cross-session emotion recognition results (%) of session 1–session 2.

Subject	sLSR	sSVM	RLSR	DLSR	RSLSR	RSRRW
subject 1	60.22	72.72	75.60	62.50	77.04	76.20
subject 2	79.21	82.45	83.53	76.44	83.53	86.78
subject 3	65.26	71.39	77.88	79.21	77.88	78.00
subject 4	68.39	55.29	68.51	72.24	68.51	80.65
subject 5	61.78	72.60	54.33	66.83	56.85	73.44
subject 6	56.97	64.66	53.25	57.93	58.29	75.36
subject 7	73.80	70.19	80.89	79.33	82.33	89.90
subject 8	75.36	67.55	74.76	68.99	74.76	88.70
subject 9	60.82	68.63	62.38	63.70	70.55	70.91
subject 10	45.55	52.76	47.00	54.33	59.74	66.83
subject 11	54.33	54.69	59.74	52.40	61.30	68.15
subject 12	69.23	56.01	56.49	63.58	62.14	75.00
subject 13	63.10	62.38	58.77	71.15	60.58	73.80
subject 14	84.01	87.62	85.22	78.85	85.22	90.26
subject 15	72.60	92.79	98.56	97.36	98.56	98.56
Avg.	66.04	68.78	69.13	69.66	71.82	79.50

Table 3. Cross-session emotion recognition results (%) of session 1–session 3.

Subject	sLSR	sSVM	RLSR	DLSR	RLSLR	RSRRW
subject 1	73.84	81.02	80.78	81.63	83.09	92.58
subject 2	80.29	86.86	91.00	86.13	92.21	92.34
subject 3	41.12	53.65	57.06	37.47	61.68	63.38
subject 4	81.39	60.58	80.29	92.94	80.29	74.21
subject 5	68.86	79.93	72.51	82.60	74.33	82.73
subject 6	74.70	74.45	77.13	82.48	79.93	83.45
subject 7	61.31	84.91	80.66	82.12	87.23	92.34
subject 8	86.25	54.87	83.21	84.55	84.91	92.46
subject 9	65.33	54.87	53.77	57.91	63.75	66.79
subject 10	45.01	60.83	41.85	39.90	64.48	66.55
subject 11	61.19	54.74	71.65	63.38	73.72	80.54
subject 12	55.60	58.39	67.64	61.92	70.92	76.28
subject 13	47.57	52.31	60.95	55.47	63.99	71.41
subject 14	70.92	71.05	79.44	67.76	90.88	87.23
subject 15	72.87	80.54	93.07	85.04	95.01	96.47
Avg.	65.75	67.27	72.73	70.75	77.76	81.25

Table 4. Cross-session emotion recognition results (%) of session 2–session 3.

Subject	sLSR	sSVM	RLSR	DLSR	RLSLR	RSRRW
subject 1	56.20	70.92	71.41	71.90	70.92	82.24
subject 2	80.29	77.37	86.73	85.89	86.01	91.50
subject 3	59.85	65.09	70.81	69.83	71.78	78.59
subject 4	83.09	72.14	77.65	77.74	76.76	87.71
subject 5	73.84	69.49	71.90	77.37	78.95	82.70
subject 6	88.32	65.69	87.43	79.56	87.10	91.73
subject 7	82.60	87.83	88.59	83.45	88.93	93.80
subject 8	77.98	75.30	77.11	70.32	81.39	85.85
subject 9	50.24	51.19	49.78	42.09	59.85	69.95
subject 10	66.91	60.34	73.27	65.94	72.87	76.28
subject 11	47.93	61.68	49.36	52.55	52.31	60.71
subject 12	80.41	66.79	72.74	85.28	79.56	81.27
subject 13	54.50	56.33	53.01	46.47	67.03	79.44
subject 14	91.48	94.89	90.69	91.00	94.04	95.86
subject 15	88.08	78.71	92.88	90.39	94.89	98.91
Avg.	72.12	70.25	74.22	72.65	77.49	83.77

Through the observation of the above data, we found the following meaningful points.

- Although the EEG data from different sessions have significant differences in distribution, our RSRRW model still achieves high recognition accuracy. Specifically, the average accuracies of RSRRW in the three cross-session tasks are 79.50%, 81.25%, and 83.77%, respectively, which are 7.68%, 3.49%, and 6.28% higher than the runner-up.
- Compared with sLSR, the RLSLR model adds the sample weight factor s . From the average performance of the model, in the three cross session tasks, RLSLR is 5.78%, 12.01%, and 5.37% higher than sLSR, which shows that the method of dynamically screening samples by using the sample probability weight s has a good effect. That is, when the sample error is too large, we can discard these samples to ensure that the model will not shift to noise.
- Compared with sLSR, the RLSR model adds the feature weight factor Θ . From the average performance of the model, RLSR is 3.09%, 6.98%, and 2.10% higher than sLSR, respectively, indicating that the model performance has been improved to a certain extent through adaptive feature weight learning. That is, for the EEG emotion recognition task, the EEG contribution of different frequency bands and different leads will be different.

- Compared with sLSR, the DLSR model adds ϵ -dragging method. From the average performance of the model, DLSR is 3.62%, 5.00%, and 0.53% higher than sLSR, respectively. It can be found that the ϵ -dragging method can effectively improve the model performance.

In order to illustrate the performance advantages of the RSRRW model compared to other models as far as possible, we use Friedman test [44] and Nemenyi test method [45] for significance testing. Firstly, we put forward the null hypothesis H_0 that “the experimental results of each group come from the population with no significant difference in numerical value”. Obviously, the alternative hypothesis H_1 is “the experimental results of each group come from the population with obvious differences in numerical value”. Secondly, determine the number of models $k = 6$, and the groups of cross-subject EEG emotion state recognition tasks $N = 45$. Thirdly, we rank the accuracy under each cross session emotional state recognition task (the higher the recognition accuracy, the higher the ranking). After sorting, the average rank r_i of each model is calculated separately, and the result is shown in Table 5.

Table 5. The average rank of each model.

Model	RSRRW	RSLSR	RLSR	DLSR	sLSR	sSVM
r	1.22	2.77	3.96	3.98	4.60	4.48

r : Average rank.

Fourthly, we calculate the variable

$$\tau_{\mathcal{F}} = \frac{(N-1)\tau_{\chi^2}}{N(K-1) - \tau_{\chi^2}}, \quad (58)$$

where

$$\tau_{\chi^2} = \frac{12N}{K(K+1)} \left(\sum_{i=1}^K r_i^2 - \frac{K(K+1)^2}{4} \right). \quad (59)$$

In this experiment, considering that the number of models $k = 6$ and the number of datasets $N = 45$ are relatively large, it can be considered that $\tau_{\chi^2} \sim \chi^2(k-1)$. Calculated by MATLAB, $\tau_{\mathcal{F}} = 39.9503$, which is much larger than the critical value 2.2551 of the \mathcal{F} test when $k = 6, n = 45$. Therefore, it can be accepted that in the case of confidence level $\alpha = 0.05$, reject the null hypothesis H_0 and accept the alternative hypothesis H_1 (that is, the probability of error in the judgment of rejecting the null hypothesis is 5%).

Based on this, in order to further distinguish the performance of each model, we performed post hoc test detection based on the Nemenyi test on the above experimental results. First, we calculate the critical distance (CD) of average ranking difference under this set of data through

$$CD = q_{\alpha} \sqrt{\frac{K(K+1)}{6N}}, \quad (60)$$

where q_{α} represents the critical value in the Nemenyi test. In the case of $k = 6, \alpha = 0.05$, the critical value is $q_{\alpha} = 2.850$. By calculation, $CD = 1.1241$. According to the above experimental results, we have drawn Figure 4. In the figure, we use a vertical line to represent the critical region, and the midpoint of the line is the average rank of the model. If two lines do not overlap, then we can draw into the conclusion that these two model differ from each other significantly, otherwise, the former bears some resemblance to the latter. For example, the rank value of RSRRW is 1.22, and the rank value of RSLSR is 2.77. Their corresponding vertical lines do not overlap, so there is a significant difference between the two at the confidence level of 0.05. Therefore, RSRRW significantly outperforms RSLSR in the cross-session emotion recognition task.

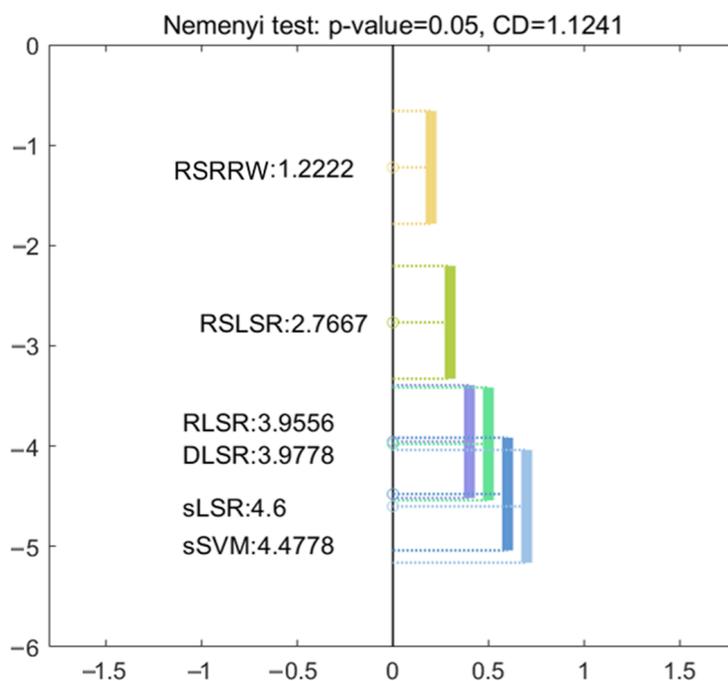


Figure 4. Nemenyi test result.

4.4. EEG Spatial-Frequency Activation Patterns Mining

For the feature weight Θ obtained in our experiment, we can determine the contribution of each frequency band and lead in emotion recognition according to the corresponding relationship between spectral features and EEG frequency bands and channels [33]. The average band importance for all 45 experiments is shown in Figure 5a, from which we can find that the *Gamma* band contributes the most in distinguishing different emotional states. Based on the consensus that brain regions may correlate differently with the occurrence of affective effects, we can also determine the importance of different EEG channels by feature weights Θ . As shown in Figure 5b, we show channel importance on the EEG topography and find that channels in the temporal and parietal lobes have higher importance. In Figure 6, we show the weight values of the top 10 channels. We believe that *P7, CP2, TP8, CZ, FC5, TP7, F8, T7, FPZ* and *P8* have high importance for the cross session emotion recognition task. The layout of the 10 key leads is shown in Figure 7.

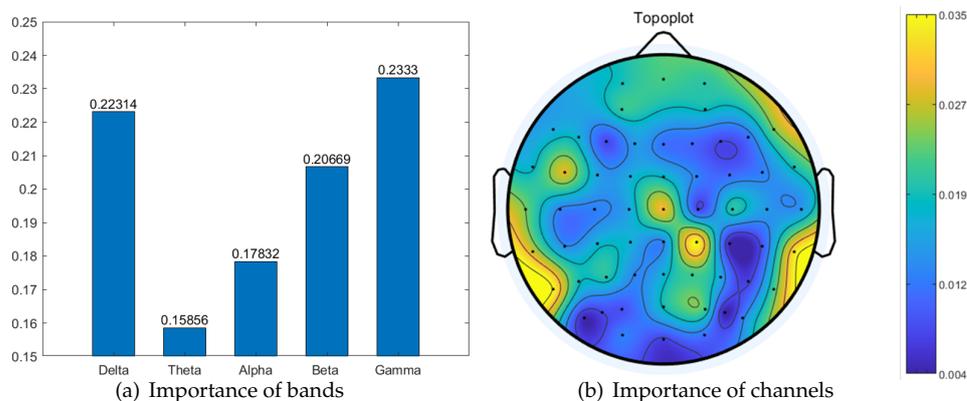


Figure 5. The average importance of EEG channels (a) and frequency bands (b) obtained by RSRRW.

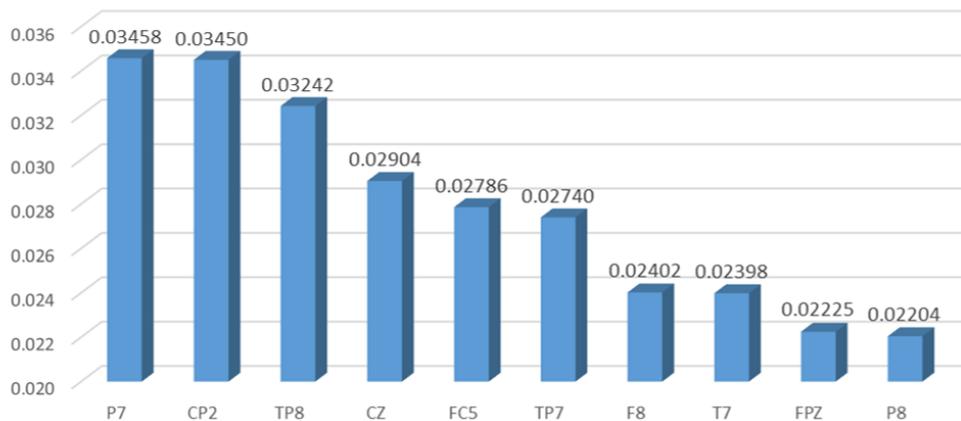


Figure 6. Top 10 EEG channels.

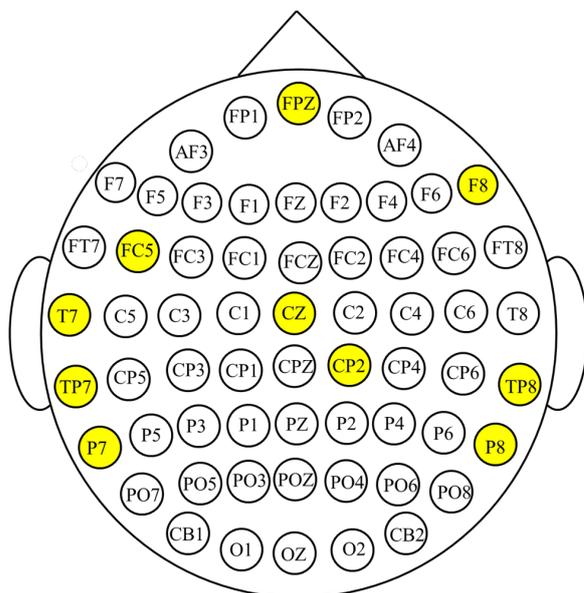


Figure 7. Top 10 EEG channels.

4.5. Effect of the Dragging-Matrix

For samples from different categories, we want to embed the labels of the samples into the objective function to achieve the purpose of increasing the distance between categories. Therefore, we introduce the ϵ -dragging method to drag the regression targets from different classes in opposite directions by matrix **M** and **B**. The dragging-matrix **M** can significantly expand the distance between categories, and each column can be regarded as a binary regressor. For samples grouped into class j , their corresponding labels become $1 + m_{ij}$, and for samples that do not belong to class j , their labels become $-m_{ij}$, where $m_{ij} \geq 0$.

To observe the performance of the ϵ -dragging method, we selected data from subject 15: session 1–session 2, subject 5: session 2–session 3 to plot the Y value of unlabeled samples. In Figure 8, we plotted certain columns of predicted labels Y for the above experiments. Then, modify the order of the samples, put the samples belonging to the same category together, so as to observe the distance between samples under each emotion category. Each color represents an affective state, with dashed lines for negative samples and solid lines for positive samples. It can be found that this method can significantly increase the distance between samples.

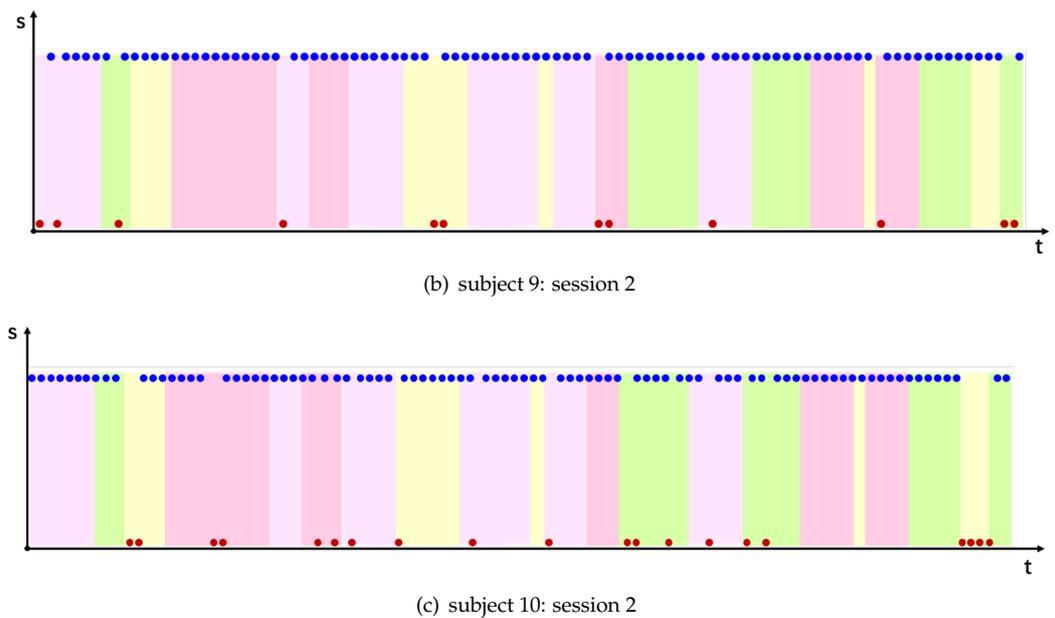


Figure 9. Visualization of sample weights s .

In the experimental paradigm of SEED-IV, each subject only had 45 s to evaluate and rest between trials. We believe that such short time cannot make some subjects completely recover from the previous emotional state. Therefore, in the follow-up EEG emotion experiment, the rest time of the subjects after each video clip should be appropriately extended to ensure that the subjects can get enough rest before the next experiment. Alongside this, it can also improve the quality of samples to a certain extent and ensure that the samples and labels have better consistence.

5. Conclusions

In this paper, we propose a model of Retargeted Semi-supervised Regression with Robust Weights Self-learning (RSRRW). We will summarize this paper from two aspects: Advantages and Results.

Advantages: (1) Compared with the DLSR method, the robustness of the model is fully considered, and binary weights are added to each sample to determine whether the sample is a noise point, which can ensure the performance of the model when some labels are inaccurate. Compared with the traditional LSR method, RSRRW introduces feature weights Θ to distinguish the weights of different features in the EEG emotion recognition task, thus obtaining the emotional activation mode. (2) Compared with the traditional LSR method, RSRRW introduces feature weight variable Θ to distinguish the weights of different features in the EEG emotion recognition task, thus obtaining the emotional activation mode. (3) RSRRW implements the ϵ -dragging method under the semi-supervised learning paradigm to expand the sample distance between different categories and improve the model performance. (4) The model has achieved good experimental results on the SEED-IV emotion recognition dataset, the average accuracies of RSRRW in the three cross-session tasks are 79.50%, 81.25%, and 83.77%, respectively, which are 7.68%, 3.49%, and 6.28% higher than the runner-up.

Results: (1) The RSRRW model greatly improves the accuracy of emotion recognition. (2) It is much more important to consider *Gamma* and *Delta* EEG frequency bands and the EEG channels locating within the temporal and (central) parietal lobes in emotion recognition tasks.

Author Contributions: Conceptualization, Y.P. and Z.C.; Data curation, Z.C. and S.D.; Investigation, Y.P.; Methodology, Z.C. and Y.P.; Software, Z.C. and Y.P.; Validation, Y.P. and S.D.; Writing—original draft preparation, Z.C. and Y.P.; Writing—review and editing, Z.C. and Y.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (61971173), Fundamental Research Funds for the Provincial Universities of Zhejiang (GK209907299001-008) and National Innovation Training Program for College Students (202210336051).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of Shanghai Jiao Tong University (protocol code 2017060).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Not applicable.

Acknowledgments: The authors also would like to thank the anonymous reviewers for their comments on this work.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Papero, D.; Frost, R.; Havstad, L.; Noone, R. Natural Systems Thinking and the Human Family. *Systems* **2018**, *6*, 19. [\[CrossRef\]](#)
- Shu, L.; Xie, J.; Yang, M.; Li, Z.; Li, Z.; Liao, D.; Xu, X.; Yang, X. A review of emotion recognition using physiological signals. *Sensors* **2018**, *18*, 2074. [\[CrossRef\]](#)
- Chang, C.; Chen, J.E. Multimodal EEG-fMRI: Advancing insight into large-scale human brain dynamics. *Curr. Opin. Biomed. Eng.* **2021**, *18*, 100279. [\[CrossRef\]](#)
- He, Z.; Li, Z.; Yang, F.; Wang, L.; Li, J.; Zhou, C.; Pan, J. Advances in multimodal emotion recognition based on brain-computer interfaces. *Brain Sci.* **2020**, *10*, 687. [\[CrossRef\]](#)
- MacNamara, A.; Joyner, K.; Klawohn, J. Event-related potential studies of emotion regulation: A review of recent progress and future directions. *Int. J. Psychophysiol.* **2022**, *176*, 73–88. [\[CrossRef\]](#) [\[PubMed\]](#)
- Li, X.; Zhang, Y.; Tiwari, P.; Song, D.; Hu, B.; Yang, M.; Zhao, Z.; Kumar, N.; Marttinen, P. EEG based Emotion Recognition: A Tutorial and Review. *ACM Comput. Surv.* **2022**. [\[CrossRef\]](#)
- Peng, Y.; Kong, W.; Qin, F.; Nie, F.; Fang, J.; Lu, B.L.; Cichocki, A. Self-weighted semi-supervised classification for joint EEG-based emotion recognition and affective activation patterns mining. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–11. [\[CrossRef\]](#)
- Lawhern, V.J.; Solon, A.J.; Waytowich, N.R.; Gordon, S.M.; Hung, C.P.; Lance, B.J. EEGNet: A compact convolutional neural network for EEG-based brain-computer interfaces. *J. Neural Eng.* **2018**, *15*, 056013. [\[CrossRef\]](#)
- Wang, F.; Zhang, W.; Xu, Z.; Ping, J.; Chu, H. A deep multi-source adaptation transfer network for cross-subject electroencephalogram emotion recognition. *Neural. Comput. Appl.* **2021**, *33*, 9061–9073. [\[CrossRef\]](#)
- Li, W.; Peng, Y. Transfer EEG Emotion Recognition by Combining Semi-Supervised Regression with Bipartite Graph Label Propagation. *Systems* **2022**, *10*, 111. [\[CrossRef\]](#)
- Chen, Y.; Chang, R.; Guo, J. Emotion recognition of eeg signals based on the ensemble learning method: Adaboost. *Math. Probl. Eng.* **2021**, *2021*, 8896062. [\[CrossRef\]](#)
- Rahman, M.M.; Sarkar, A.K.; Hossain, M.A.; Hossain, M.S.; Islam, M.R.; Hossain, M.B.; Quinn, J.M.; Moni, M.A. Recognition of human emotions using EEG signals: A review. *Comput. Biol. Med.* **2021**, *136*, 104696. [\[CrossRef\]](#)
- Bahador, N.; Erikson, K.; Laurila, J.; Koskenkari, J.; Ala-Kokko, T.; Kortelainen, J. Automatic detection of artifacts in EEG by combining deep learning and histogram contour processing. In Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Montreal, QC, Canada, 20–24 July 2020; pp. 138–141. [\[CrossRef\]](#)
- Mehmood, R.M.; Bilal, M.; Vimal, S.; Lee, S.W. EEG-based affective state recognition from human brain signals by using Hjorth-activity. *Measurement* **2022**, *202*, 111738. [\[CrossRef\]](#)
- Singh, M.I.; Singh, M. Emotion recognition: An evaluation of ERP features acquired from frontal EEG electrodes. *Appl. Sci.* **2021**, *11*, 4131. [\[CrossRef\]](#)
- Alsolamy, M.; Fattouh, A. Emotion estimation from EEG signals during listening to Quran using PSD features. In Proceedings of the 2016 7th International Conference on Computer Science and Information Technology (CSIT), Amman, Jordan, 13–15 July 2016; pp. 1–5. [\[CrossRef\]](#)
- Fan, C.; Liu, X.; Gu, X.; Zhou, L.; Li, X. Research on emotion recognition of EEG signal based on convolutional neural networks and high-order cross-analysis. *J. Healthc. Eng.* **2022**, *2022*, 6238172. [\[CrossRef\]](#)
- Hwang, S.; Hong, K.; Son, G.; Byun, H. Learning CNN features from DE features for EEG-based emotion recognition. *Pattern Anal. Appl.* **2020**, *23*, 1323–1335. [\[CrossRef\]](#)
- Sharma, R.; Pachori, R.B.; Sircar, P. Automated emotion recognition based on higher order statistics and deep learning algorithm. *Biomed. Signal Process. Control* **2020**, *58*, 101867. [\[CrossRef\]](#)

20. Donmez, H.; Ozkurt, N. Emotion classification from EEG signals in convolutional neural networks. In Proceedings of the 2019 Innovations in Intelligent Systems and Applications Conference (ASYU), Izmir, Turkey, 31 October–2 November 2019; pp. 1–6. [[CrossRef](#)]
21. Gupta, V.; Chopda, M.D.; Pachori, R.B. Cross-subject emotion recognition using flexible analytic wavelet transform from EEG signals. *IEEE Sens. J.* **2018**, *19*, 2266–2274. [[CrossRef](#)]
22. Chen, J.; Jiang, D.; Zhang, Y. A common spatial pattern and wavelet packet decomposition combined method for EEG-based emotion recognition. *J. Adv. Comput. Intell. Intell. Inform.* **2019**, *23*, 274–281. [[CrossRef](#)]
23. Wang, J.; Wang, M. Review of the emotional feature extraction and classification using EEG signals. *Cognit. Robot.* **2021**, *1*, 29–40. [[CrossRef](#)]
24. Basar, M.D.; Duru, A.D.; Akan, A. Emotional state detection based on common spatial patterns of EEG. *Signal Image Video Process* **2020**, *14*, 473–481. [[CrossRef](#)]
25. Novi, Q.; Guan, C.; Dat, T.H.; Xue, P. Sub-band common spatial pattern (SBCSP) for brain-computer interface. In Proceedings of the 2007 3rd International IEEE/EMBS Conference on Neural Engineering, Kohala Coast, HI, USA, 2–5 May 2007; pp. 204–207. [[CrossRef](#)]
26. Li, Z.; Tian, X.; Shu, L.; Xu, X.; Hu, B. Emotion recognition from EEG using RASM and LSTM. In Proceedings of the International Conference on Internet Multimedia Computing and Service (ICIMS), Qingdao, China, 23–25 August 2017; pp. 310–318. [[CrossRef](#)]
27. Zhang, G.; Yu, M.J.; Chen, G.; Han, Y.H.; Zhang, D.; Zhao, G.Z.; Liu, Y.-J. A review of EEG features for emotion recognition. *Sci. China Inf. Sci.* **2019**, *49*, 1097–1118. (In Chinese) [[CrossRef](#)]
28. Thammasan, N.; Fukui, K.I.; Numao, M. Application of deep belief networks in eeg-based dynamic music-emotion recognition. In Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 24–29 July 2016; pp. 881–888. [[CrossRef](#)]
29. Li, J.; Zhang, Z.; He, H. Hierarchical convolutional neural networks for EEG-based emotion recognition. *Cognit. Comput.* **2018**, *10*, 368–380. [[CrossRef](#)]
30. Cui, H.; Liu, A.; Zhang, X.; Chen, X.; Wang, K.; Chen, X. EEG-based emotion recognition using an end-to-end regional-asymmetric convolutional neural network. *Knowl. Based Syst.* **2020**, *205*, 106243. [[CrossRef](#)]
31. Liu, Y.; Ding, Y.; Li, C.; Cheng, J.; Song, R.; Wan, F.; Chen, X. Multi-channel EEG-based emotion recognition via a multi-level features guided capsule network. *Comput. Biol. Med.* **2020**, *123*, 103927. [[CrossRef](#)] [[PubMed](#)]
32. Gong, S.; Xing, K.; Cichocki, A.; Li, J. Deep learning in EEG: Advance of the last ten-year critical period. *IEEE Trans. Cogn. Develop. Syst.* **2022**, *14*, 348–365. [[CrossRef](#)]
33. Peng, Y.; Qin, F.; Kong, W.; Ge, Y.; Nie, F.; Cichocki, A. GFIL: A unified framework for the importance analysis of features, frequency bands, and channels in EEG-based emotion recognition. *IEEE Trans. Cogn. Develop. Syst.* **2022**, *14*, 935–947. [[CrossRef](#)]
34. Peng, Y.; Jin, F.; Kong, W.; Nie, F.; Lu, B.L.; Cichocki, A. OGSSL: A semi-supervised classification model coupled with optimal graph learning for EEG emotion recognition. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2022**, *30*, 1288–1297. [[CrossRef](#)]
35. Ekman, P. Facial expression and emotion. *Am. Psychol.* **1993**, *48*, 384. [[CrossRef](#)]
36. Van den Broek, E.L. Ubiquitous emotion-aware computing. *Pers. Ubiquitous. Comput.* **2013**, *17*, 53–67. [[CrossRef](#)]
37. Posner, J.; Russell, J.A.; Peterson, B.S. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Dev. Psychopathol.* **2005**, *17*, 715–734. [[CrossRef](#)]
38. Lang, P.J. The emotion probe: Studies of motivation and attention. *Am. Psychol.* **1995**, *50*, 372. [[CrossRef](#)] [[PubMed](#)]
39. Chen, X.; Nie, F.; Yuan, G.; Huang, J.Z. Semi-supervised feature selection via rescaled linear regression. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI), Melbourne, Australia, 19–25 August 2017; Volume 2017, pp. 1525–1531. [[CrossRef](#)]
40. Xiang, S.; Nie, F.; Meng, G.; Pan, C.; Zhang, C. Discriminative least squares regression for multiclass classification and feature selection. *IEEE Trans. Neural Netw. Learn. Syst.* **2012**, *23*, 1738–1754. [[CrossRef](#)]
41. Nie, F.; Yuan, J.; Huang, H. Optimal mean robust principal component analysis. In Proceedings of the International Conference on Machine Learning, Beijing, China, 21–26 June 2014; pp. 1062–1070.
42. Zheng, W.L.; Liu, W.; Lu, Y.; Lu, B.L.; Cichocki, A. Emotionmeter: A multimodal framework for recognizing human emotions. *IEEE Trans. Cybern.* **2018**, *49*, 1110–1122. [[CrossRef](#)] [[PubMed](#)]
43. Wang, J.; Xie, F.; Nie, F.; Li, X. Robust Supervised and Semisupervised Least Squares Regression Using $\ell_{2,p}$ -Norm Minimization. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, 1–15. [[CrossRef](#)]
44. Friedman, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Am. Stat. Assoc.* **1937**, *32*, 675–701. [[CrossRef](#)]
45. Zhou, Z. *Machine Learning*; Tsinghua University Press: Beijing, China, 2016; pp. 41–45. (In Chinese)