Efficient Sample and Feature Importance Mining in Semi-Supervised EEG Emotion Recognition

Xing Li, Fangyao Shen, Yong Peng, *Member, IEEE*, Wanzeng Kong, *Member, IEEE*, and Bao-Liang Lu, *Fellow, IEEE*

***Abstract*—Recently, electroencephalogram (EEG)-based emotion recognition has attracted increasing interests in research community. The weak, non-stationary, multi-rhythm and multi-channel properties of EEG data easily cause the extracted EEG samples and features contribute differently in recognizing emotional states. However, existing studies either failed to consider both the issues of sample and feature importance or only considered one of them. In this brief, we propose a new model termed sJSFE (semi-supervised Joint Sample and Feature importance Evaluation) to quantitatively measure the sample and feature importance by self-paced learning and feature self-weighting respectively. Experimental results on the SEED-IV data set show that the emotion recognition performance is greatly improved by mining both the sample and feature importance. Specifically, the average accuracy obtained by sJSFE across the three cross-session recognition tasks is 82.45%, which is respectively 3.72% and 7.21% and 10.47% and 18.82% higher than the results of traditional models. Besides, the feature importance vector depicts that the *Gamma* frequency band contributes the most, and the brain regions of prefrontal, left/right temporal and (central) parietal lobes correlate more to emotion recognition. The sample importance descriptor shows that continual transitions of video types in consecutive trials might weaken the feature-label consistency of the collected EEG data.**

***Index Terms*—EEG, emotion recognition, feature importance, sample importance, semi-supervised learning.**

I. Introduction

IN PAST decades, rapid progresses have been made in physiological data-based human-computer interactions from sensation and perception to higher-level cognitive activities [1]. Especially, EEG-based emotion recognition has been a research hotspot in multiple disciplines such as information sciences and neural engineering [2]. Different from the popular data modalities in emotion recognition such as image and speech, EEG has some special characteristics. First, EEG is weak which is easily contaminated by noises during its collection process. Typically, the noises are from hardware devices, electrooculogram and electromyography. Though current EEG preprocessing methods such as filtering and artifact removal can moderately improve the data reliability [3], it is common that different samples contribute differently in characterizing the emotional states. Second, EEG is inherently multi-rhythm and multi-channel; therefore, EEG features extracted from different rhythms and channels should correlate differently to the occurrence of affective effect. However, existing studies either failed to consider both the issues of sample and feature importance or only considered one of them [4]–[7].

In this brief, we propose a model termed sJSFE to jointly explore the sample and feature importance in EEG-based emotion recognition. Specifically, the self-paced learning (SPL) and feature self-weighting techniques are seamlessly incorporated into the semi-supervised least square regression framework [8], [9]. SPL trains the model gradually from easy to complex samples iteratively in a self-paced fashion, in which weight parameters are adaptively learned to describe the complexities of samples. Simultaneously, a feature importance vector is incorporated to learn the contributions of different feature dimensions in characterizing the emotional states. To the best of our knowledge, sJSFE is the first model that both sample and feature importance are explored jointly.

Besides, to handle the small sample size problem, sJSFE works in the semi-supervised learning paradigm [10]. On one hand, the underlying EEG data properties can be better captured by involving unlabeled samples into the model learning; on the other hand, the label indicator matrix of unlabeled samples is jointly optimized with the other variables, which further facilitates the sample complexity estimation. Experiments on the benchmark EEG data set show that the emotion recognition performance is greatly improved by sJSFE. Moreover, the feature importance vector and the sample importance descriptor respectively provide us with more insights into the affective activation patterns and EEG data collection paradigm.

Manuscript received March 8, 2022; accepted March 24, 2022. Date of publication March 29, 2022; date of current version June 29, 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFE0116800; in part by the National Natural Science Foundation of China under Grant 61971173; in part by the Fundamental Research Funds for the Provincial Universities of Zhejiang under Grant GK209907299001-008; in part by the CAAC Key Laboratory of Flight Techniques and Flight Safety under Grant FZ2021KF16; and in part by the Guangxi Key Laboratory of Optoelectronic Information Processing (Guilin University of Electronic Technology) under Grant GD21202. This brief was recommended by Associate Editor G. Jovanovic Dolecek. *(Corresponding author: Yong Peng.)*

Xing Li, Fangyao Shen, and Wanzeng Kong are with the School of Computer Science, Hangzhou Dianzi University, Hangzhou 310018, China.

Yong Peng is with the School of Computer Science, Hangzhou Dianzi University, Hangzhou 310018, China, and also with the Key Laboratory of Flight Techniques and Flight Safety, Civil Aviation Administration of China, Guanggan 618307, China (e-mail: yongpeng@hdu.edu.cn).

Bao-Liang Lu is with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSII.2022.3163141.

Digital Object Identifier 10.1109/TCSII.2022.3163141

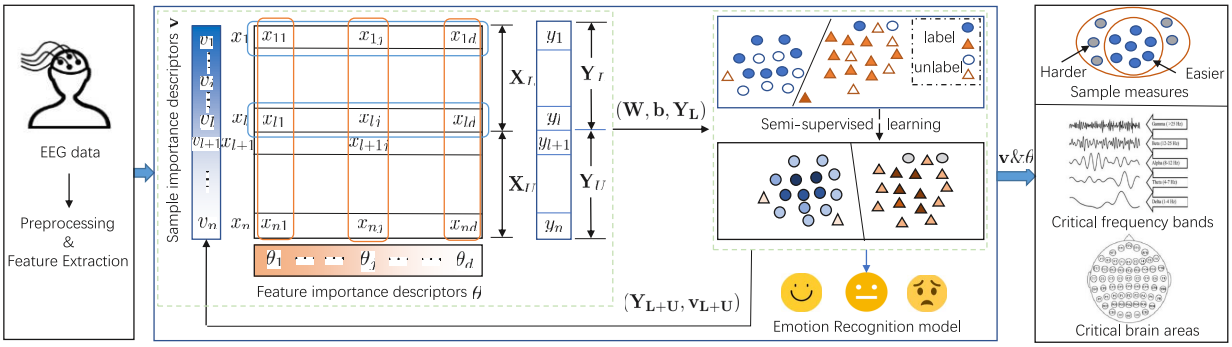

Fig. 1. The general framework of sJSFE.

II. The Proposed Method

A. sJSFE Model Formulation

Consider that we have an EEG data matrix $\mathbf{X} = [\mathbf{X}_l, \mathbf{X}_u] \in \mathbb{R}^{d\times n}$ consisting of l labeled samples and u unlabeled samples. $\mathbf{Y}_l \in \mathbb{R}^{l\times c}$ is the indicator matrix of labeled samples in one-hot encoding. Specifically, if sample $\mathbf{x}_i|_{i=1}^{l}$ belongs to the j-th emotional state and $\mathbf{y}^i$ is the i-th row of $\mathbf{Y}_l$, then the j-th element of $\mathbf{y}^i$ is one and all the others of $\mathbf{y}^i$ are zeros. $\mathbf{Y}_u \in \mathbb{R}^{u\times c}$ is an unknown label matrix corresponding to the unlabeled samples, and $\mathbf{Y} = [\mathbf{Y}_l; \mathbf{Y}_u] \in \mathbb{R}^{n\times c}$ is the combined label matrix corresponding to $\mathbf{X}$. Here, d is the sample dimensionality, c is the number of emotional states, $n = l + u$ is the total number of EEG samples. Our task is to estimate $\mathbf{Y}_u$ as accurate as possible given $\mathbf{X}$ and $\mathbf{Y}_l$.

By connecting the EEG data matrix with emotional label matrix in least square regression (LSR) formula, we have

$$\begin{aligned} \min_{\mathbf{W},\mathbf{b},\mathbf{Y}_u} \quad & \|\mathbf{X}^T\mathbf{W} + \mathbf{1}_n\mathbf{b}^T - \mathbf{Y}\|_2^2 + \gamma\|\mathbf{W}\|_2^2, \\ s.t. \quad & \mathbf{Y}_u \geq \mathbf{0}, \mathbf{Y}_u\mathbf{1}_c = \mathbf{1}_u, \end{aligned} \tag{1}$$

where $\mathbf{1}_c$ is an all-one column vector with length c. The second constraint enforces the summation of elements in each row of $\mathbf{Y}_u$ to be one. Together with the non-negative constraint, elements in each row of $\mathbf{Y}_u$ act as the probabilities of a certain sample to different emotional states. For example, if the fourth row of $\mathbf{Y}_u$ is [0.03, 0.83, 0.08, 0.06], then we mark the third unlabeled sample as the second emotional state. In (1), $\mathbf{Y}_u$ is an variable which is jointly optimized with $\mathbf{W}$ and $\mathbf{b}$.

As shown in Fig. 1, two vectors, $\mathbf{v} = [v_1, \ldots, v_n]^T \in \mathbb{R}^n$ and $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_d]^T \in \mathbb{R}^d$, are respectively used to characterize the sample importance and feature importance. That is, v_i characterizes the importance of the i-th sample, and θ_j is the importance measure of the j-th feature. By incorporating $\mathbf{v}$ and $\boldsymbol{\theta}$ into (1), we achieve the objective function of sJSFE as

$$\begin{aligned} \min \quad & \frac{C}{2}\sum_{i=1}^{n} v_i\|\mathbf{x}_i^T\boldsymbol{\Theta}\mathbf{W} + \mathbf{b}^T - \mathbf{y}^i\|_2^2 + \frac{1}{2}\|\mathbf{W}\|_2^2 + f(\lambda, \mathbf{v}) \\ s.t. \quad & \mathbf{W}, \mathbf{b}, \mathbf{v}, \boldsymbol{\theta} \geq \mathbf{0}, \boldsymbol{\theta}^T\mathbf{1}_d = 1, \mathbf{Y}_u \geq \mathbf{0}, \mathbf{Y}_u\mathbf{1}_c = \mathbf{1}_u, \end{aligned} \tag{2}$$

where $\boldsymbol{\Theta} \in \mathbb{R}^{d\times d}$ is a rescaled diagonal matrix with its j-th diagonal element $\Theta_{jj}|_{j=1}^{d} = \sqrt{\theta_j}$. Obviously, $\mathbf{x}_i$ is replaced by $\mathbf{x}_i^T\boldsymbol{\Theta}$ to enforce the different contributions of feature dimensions in emotion recognition. $\mathbf{v}_i$ is mediated by the approximation error of sample $\mathbf{x}_i$, i.e., $\|\mathbf{x}_i^T\boldsymbol{\Theta}\mathbf{W} + \mathbf{b}^T - \mathbf{y}^i\|_2^2$. The larger the approximation error, the harder the sample to fit. Then, v_i should be assigned a smaller value. It is worth mentioning that for the unlabeled EEG samples, their ground truth labels $\mathbf{y}^i|_{i=l+1}^{n}$ are unavailable. Therefore, $\mathbf{Y}_u$ is jointly optimized in (2) to offer the soft labels for calculating the approximation error of the unlabeled EEG samples.

The regularizer $f(\lambda, \mathbf{v})$ defines a self-paced function, where parameter λ specifies how EEG samples are selected and how their weights are calculated. Then, the learning pace is controlled to learn new samples. In this brief, we use the below linear self-paced regularization term [8]

$$f(\lambda, \mathbf{v}) = \frac{1}{2}\lambda\sum_{i=1}^{n}(v_i^2 - 2v_i). \tag{3}$$

By setting $\widetilde{\mathbf{W}} = \boldsymbol{\Theta}\mathbf{W}$, we have $\mathbf{W} = \boldsymbol{\Theta}^{-1}\widetilde{\mathbf{W}}$ and then (2) can be reformulated as

$$\begin{aligned} \min_{\widetilde{\mathbf{W}},\mathbf{v},\mathbf{b},\boldsymbol{\theta},\mathbf{Y}_u} & \frac{C}{2}\sum_{i=1}^{n} v_i\|\mathbf{x}_i^T\widetilde{\mathbf{W}} + \mathbf{b}^T - \mathbf{y}^i\|_2^2 + \frac{1}{2}\|\boldsymbol{\Theta}^{-1}\widetilde{\mathbf{W}}\|_2^2 \\ & + f(\lambda, \mathbf{v}), s.t. \boldsymbol{\theta} \geq \mathbf{0}, \boldsymbol{\theta}^T\mathbf{1}_d = 1, \mathbf{Y}_u \geq \mathbf{0}, \mathbf{Y}_u\mathbf{1}_c = \mathbf{1}_u. \end{aligned} \tag{4}$$

Since $\Theta_{jj}|_{j=1}^{d} = \sqrt{\theta_j}$ and $\boldsymbol{\theta}^T\mathbf{1}_d = 1$, we have

$$\min_{\boldsymbol{\theta}\geq\mathbf{0},\boldsymbol{\theta}^T\mathbf{1}_d=1} \|\boldsymbol{\Theta}^{-1}\widetilde{\mathbf{W}}\|_2^2 = \min_{\boldsymbol{\theta}\geq\mathbf{0},\boldsymbol{\theta}^T\mathbf{1}_d=1} \sum_{j=1}^{d}\frac{1}{\theta_j}\|\widetilde{\mathbf{w}}^j\|_2^2, \tag{5}$$

The corresponding Lagrangian function w.r.t. θ_j is

$$\mathcal{L}(\theta_j) = \frac{1}{\theta_j}\|\widetilde{\mathbf{w}}^j\|_2^2 - \eta(\boldsymbol{\theta}^T\mathbf{1}_d - 1). \tag{6}$$

By setting the derivative of $\mathcal{L}(\theta_j)$ w.r.t. θ_j to be zero and considering the normalization constraint, we have

$$\theta_j = \frac{\|\widetilde{\mathbf{w}}^j\|_2}{\sum_{j'=1}^{d}\|\widetilde{\mathbf{w}}^{j'}\|_2}. \tag{7}$$

The above derivation depicts that how the feature self-weighting variable $\boldsymbol{\Theta}$ can be absorbed in the intermediate variable $\widetilde{\mathbf{W}}$. Then we can use $\min_{\theta>0,\theta^T\mathbf{1}_d=1}\|\widetilde{\mathbf{W}}\|_{2,1}^2$ to replace the second term in (4).

B. sJSFE Model Optimization

Below we derive the updating rule to each variable in (4).

■ **Update b.** By denoting $\mathbf{U} = \text{diag}(\sqrt{\mathbf{v}})$, $\mathbf{G} = \mathbf{U}\mathbf{X}^T$, $\mathbf{H} = \mathbf{U}\mathbf{1}$ and $\mathbf{T} = \mathbf{U}\mathbf{Y}$, we set the derivative of (4) w.r.t. $\mathbf{b}$ to zero and then get the optimal solution of $\mathbf{b}$ as

$$\mathbf{b} = (\mathbf{H}^T\mathbf{H})^{-1}(\mathbf{T}^T - \widetilde{\mathbf{W}}^T\mathbf{G}^T)\mathbf{H}. \tag{8}$$

■ **Update $\widetilde{\mathbf{W}}$.** The objective function in terms of $\widetilde{\mathbf{W}}$ is

$$\min_{\widetilde{\mathbf{W}}} \frac{C}{2}\sum_{i=1}^{n} v_i\|\mathbf{x}_i^T\widetilde{\mathbf{W}} + \mathbf{b}^T - \mathbf{y}^i\|_2^2 + \frac{1}{2}\|\widetilde{\mathbf{W}}\|_{2,1}^2. \tag{9}$$

Since $\mathbf{G} = \mathbf{U}\mathbf{X}^T$, $\mathbf{H} = \mathbf{U}\mathbf{1}$, and $\mathbf{T} = \mathbf{U}\mathbf{Y}$, we rewrite (9) as

$$\mathcal{L}(\widetilde{\mathbf{W}}) = \frac{C}{2}\|\mathbf{G}\widetilde{\mathbf{W}} + \mathbf{H}\mathbf{b}^T - \mathbf{T}\|_2^2 + \frac{1}{2}\mathbf{Q}\widetilde{\mathbf{W}}, \tag{10}$$

where $\mathbf{Q} \in \mathbb{R}^{d\times d}$ is a diagonal matrix with its j-th diagonal element as

$$q_{jj} = \frac{\sum_{p=1}^{d}\sqrt{\|\widetilde{\mathbf{w}}^p\|_2^2 + \epsilon}}{\sqrt{\|\widetilde{\mathbf{w}}^j\|_2^2 + \epsilon}}. \tag{11}$$

By setting the derivative of $\mathcal{L}(\widetilde{\mathbf{W}})$ w.r.t. $\widetilde{\mathbf{W}}$ to zero, we have the updating rule to $\widetilde{\mathbf{W}}$ as

$$\widetilde{\mathbf{W}} = (\mathbf{G}^T\mathbf{G} + \frac{1}{C}\mathbf{Q})^{-1}\mathbf{G}^T(\mathbf{T} - \mathbf{H}\mathbf{b}^T). \tag{12}$$

■ **Update $\mathbf{Y}_u$.** Note that the first term in (4) can be decoupled for each $i|_{i=l+1}^n$. If the associated v_i is not zero, $\mathbf{y}^i$ can be obtained by solving

$$\min_{\mathbf{y}^i\geq\mathbf{0},\mathbf{y}^i\mathbf{1}_c=1} \|\mathbf{x}_i^T\widetilde{\mathbf{W}} + \mathbf{b}^T - \mathbf{y}^i\|_2^2, \tag{13}$$

which defines an Euclidean projection on a simplex constraint. The detailed derivation can be found in [11].

■ **Update v.** By defining $l_i = \|\mathbf{x}_i^T\widetilde{\mathbf{W}} + \mathbf{b}^T - \mathbf{y}^i\|_2^2$ as the loss on sample $\mathbf{x}_i$ and absorbing parameter C into λ, we have the objective function in terms of $\mathbf{v}$ as

$$\min_{\mathbf{v}} \sum_{i=1}^{n} v_i l_i + \frac{1}{2}\lambda\sum_{i=1}^{n}(v_i^2 - 2v_i), s.t.\ 0 \leq v_i|_{i=1}^n \leq 1. \tag{14}$$

By taking the derivative of (14) w.r.t. v_i and setting it to be zero, we obtain the analytical solution to v_i as

$$v_i = \begin{cases} 1 - \frac{l_i}{\lambda}, & l_i < \lambda; \\ 0, & l_i \geq \lambda. \end{cases} \tag{15}$$

Obviously, the larger l_i, the smaller v_i.

The complete procedure to optimize the sJSFE objective function is summarized in Algorithm 1.

Below we provide the computational complexity analysis of sJSFE. We need $\mathcal{O}(n^2d + d^2n + d^3 + dnc + d^2c)$ complexity to update $\widetilde{\mathbf{W}}$. When updating $\mathbf{Q}$, its complexity is $\mathcal{O}(dc)$. The updating of $\mathbf{b}$ and $\mathbf{v}$ requires $\mathcal{O}(dnc+nc+n)$ and $\mathcal{O}(dnc+nc)$, respectively. Since the calculation of $\mathbf{x}_i^T\widetilde{\mathbf{W}}$ have been completed in updating $\mathbf{v}$, the complexity of updating $\mathbf{Y}_u$ is $\mathcal{O}(uc)$. Considering that it is usually $n \approx u > d >> c$, we have the overall computational complexity of sJSFE is $\mathcal{O}(tn^2d)$ where t is the number of iterations.

Algorithm 1 Optimization to sJSFE Objective Function (4)

Input: EEG data matrix $\mathbf{X} = [\mathbf{X}_l, \mathbf{X}_u]$ and the corresponding label matrix $\mathbf{Y}_l$, parameters C, λ and k;
Output: Weighted projection matrix $\widetilde{\mathbf{W}}$, bias vector $\mathbf{b}$ and sample importance descriptor $\mathbf{v}$ and estimated label $\mathbf{Y}_u$.

```
Initialize v = [1/n, ..., 1/n], b = 1/n (Y^T 1 − W̃^T X1) and Q as an identity matrix, λ = 0.001;
while not converged do
    Update W̃ by optimizing (12);
    Update Q by equation (11);
    Update b by equation (8);
    Update v by equation (15);
    Update Y_u by solving (13) for each i|_{i=l+1}^n;
    Update λ = kλ;
end while
Calculate the feature importance vector θ by equation (7).
```

TABLE I
SUMMARY OF THE SEED-IV EMOTIONAL EEG DATA SET

Item	Properties
# subject	15
feature	differential entropy (DE)
# electrode	62
frequency bands	*Delta*, *Theta*, *Alpha*, *Beta*, *Gamma*
emotional states	*sad*, *fear*, *happy*, *neutral*
channel order	FP1, FPZ, ⋯, CB2

The detailed channel order can be found in the SEED-IV description file.

III. EXPERIMENTS

A. Data Set and Experimental Settings

In the following experiments, we use the benchmark SEED-IV emotional EEG data set to evaluate the effectiveness of sJSFE, whose detailed descriptions can be found in [12]. In Table I, we only list its main properties.

We conduct cross-session emotion recognition in chronological order. That is, for each subject, we have three recognition tasks of 'session1→session2', 'session1→session3' and 'session2→session3'. In the 'session1→session2' task, EEG samples from session 1 are fully labeled while those from session 2 are unlabeled. We compare sJSFE with semi-supervised support vector machine with linear kernel (sSVM), rescaled LSR (RLSR) [9], semi-supervised least squares regression (sLSR) and semi-supervised SPL (sSPL). Here, RLSR explicitly takes the feature importance into account while sSPL considers the sample importance by reweighting the least square loss with a linear self-paced regularizer. Parameters involved in respective models are searched from $\{2^{-20}, 2^{-19}, \ldots, 2^{20}\}$. The step parameter k in sJSFE and sSPL controls the pace of self-paced learning, which is searched from $\{1.1, 1.2, \ldots, 3.0\}$.

B. Results and Analysis

The emotion recognition accuracies of these compared models are provided in Table II, where the best result in each case is highlighted in bold. s1, ..., s15 are indices of the 15 subjects in SEED-IV data set. From these results, we have the following findings. 1) sJSFE obtained very competitive emotion recognition performance despite the discrepancies for EEG data collected at different times. Specifically, the

TABLE II
CROSS-SESSION EMOTION RECOGNITION RESULTS (%) OF DIFFERENT MODELS ON SEED-IV

ID	session1→session2					session1→session3					session2→session3				
	sJSFE	sSPL	RLSR	sLSR	sSVM	sJSFE	sSPL	RLSR	sLSR	sSVM	sJSFE	sSPL	RLSR	sLSR	sSVM
s1	**77.16**	67.91	67.19	61.54	43.87	**87.35**	86.25	81.39	73.84	56.93	**69.59**	68.98	65.21	64.11	63.63
s2	**94.83**	94.11	88.94	85.10	90.63	**94.83**	85.64	87.59	84.55	83.21	86.13	**86.86**	85.28	85.40	83.58
s3	**82.81**	77.16	69.35	62.26	48.80	**78.83**	58.52	57.79	53.28	34.31	**78.85**	77.86	71.78	71.29	45.99
s4	**83.05**	76.32	74.88	72.60	48.56	**88.08**	82.60	83.21	74.09	48.30	**89.05**	86.37	86.86	75.67	57.06
s5	**89.54**	75.36	67.67	62.62	70.79	**78.49**	78.22	65.57	65.21	63.38	**88.56**	84.06	80.41	79.20	70.56
s6	**78.25**	76.20	70.67	69.83	57.21	**87.35**	87.10	83.09	82.48	82.12	**88.69**	86.74	85.77	78.83	73.24
s7	**94.23**	93.51	85.22	84.13	59.01	**91.61**	90.02	83.94	83.58	63.50	**92.58**	92.09	91.97	85.64	85.77
s8	83.53	**86.42**	81.61	83.77	66.23	**83.53**	80.17	82.73	80.05	79.93	**84.67**	81.27	83.21	76.52	80.05
s9	**83.65**	77.88	78.61	67.67	74.28	**83.65**	64.84	61.80	59.73	58.76	**76.32**	74.82	60.10	54.26	55.60
s10	**63.58**	63.34	56.37	52.40	41.35	**69.10**	66.18	56.33	56.93	58.52	**75.18**	73.84	70.07	72.75	63.50
s11	**67.91**	64.18	55.17	55.05	56.49	**84.31**	81.51	70.56	68.25	58.76	**75.91**	66.18	73.24	64.36	56.33
s12	**72.36**	69.47	69.59	66.23	31.13	**73.92**	66.67	63.02	62.65	49.51	**73.84**	72.26	71.78	71.17	63.02
s13	**77.28**	74.76	67.19	65.26	61.06	**67.19**	61.80	61.92	58.52	50.12	**71.88**	70.92	67.52	57.54	60.95
s14	**84.38**	83.65	80.77	79.33	81.01	**86.13**	84.31	84.43	77.98	65.94	89.17	**92.46**	83.45	91.36	66.91
s15	**97.36**	93.15	97.36	88.58	94.23	**92.58**	90.27	84.79	83.70	80.54	**93.03**	90.51	90.39	89.90	78.71
avg.	**82.00**	78.23	74.04	70.42	61.64	**83.13**	77.61	73.88	70.99	62.25	**82.23**	80.35	77.80	74.53	66.99

average accuracies of sJSFE in the three tasks are 82.00%, 83.13%, 82.23%, which respectively make improvements by 3.77%, 5.52% and 1.88% in comparison with the runner-up. 2) Based on the average performance, sSPL outperforms sLSR respectively by 7.81%, 6.62% and 5.82% in the three tasks, with the help of the sample importance descriptor in dynamically adjusting the effect of samples. That is, a smaller weight should be assigned to the sample if it is hard to fit by the learned model. 3) By adaptively exploring the different importance of EEG features, RLSR obtains improved performance of 3.62%, 2.89%, and 3.27% in comparison with its counterpart sLSR. Similarly, sJSFE further obtained average performance superiorities to sSPL, indicating the necessity of adaptive feature weights learning. 4) The results significantly show that the quality of both samples and features plays important roles in determining the recognition performance. In sJSFE, the sample importance descriptor and the feature importance vector are jointly optimized for better capturing the data characteristics, leading to improved performance.

In comparison with sJSFE, RLSR needs not to estimate the sample importance descriptor $\mathbf{v}$ and sLSR has no both the variables $\mathbf{v}$ and $\mathbf{Q}$. Therefore, their overall complexities are both $\mathcal{O}(tn^2d)$. Similarly, because there is no feature importance learning in sSPL, it does not involve the optimization of variable $\mathbf{Q}$, leading to the complexity $\mathcal{O}(tn^2d)$. On the sSVM, our implementation is based on the work [13], whose complexity is $\mathcal{O}(n^2 \ln n)$. We conclude that sJSFE does not incur more complexity based on the big $\mathcal{O}$ notation. Usually, EEG data sets are in moderate sizes, which are not time-consuming to perform emotion recognition.

C. Affective Activation Patterns Analysis

Based on the correspondence between spectra features and EEG frequency bands and channels [14], critical frequency bands and channels in emotion recognition can be identified by the learned feature importance vector $\boldsymbol{\theta}$. The average feature importance values across all the 45 experimental cases are shown by Fig. 2, from which we easily find that EEG features contribute differently in emotion recognition.

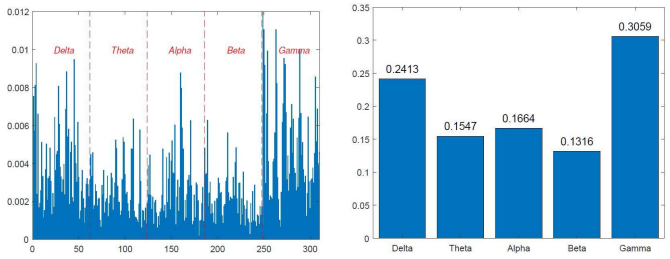

Fig. 2. The average importance of EEG features (left) and frequency bands (right) obtained by sJSFE.

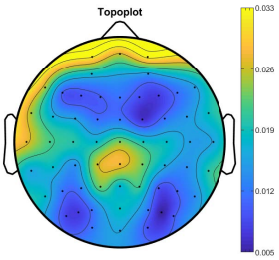

Fig. 3. The average importance of EEG channels obtained by sJSFE.

Moreover, the *Gamma* band offers the greatest contribution in differentiating different emotional states, which is consistent with the finding obtained by the trial-and-error manner [15].

Based on consensus that different brain regions might correlate differently to the occurrence of affective effects, we also can identify the importance of different EEG channels by $\boldsymbol{\theta}$. In Fig. 3, we project the channel importance values onto the brain topographical map for facilitating the identification of critical brain regions in EEG-based emotion recognition. It is obvious that the channels at the prefrontal, left/right central and (central) parietal lobes are generally more important. From the data-driven perspective, such explored affective activation patterns lay potential foundations for further understanding the neural mechanism of emotion processing, and the future design of emotion-customized EEG data acquisition devices.

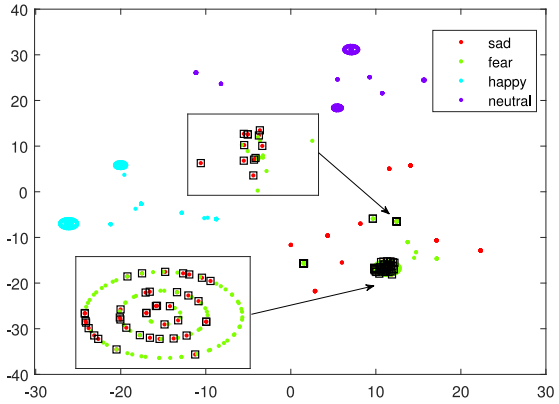

Fig. 4. An example to visualize the sample importance.

D. Sample Importance Analysis

Here we illustrate the effect of sample importance descriptor in improving the model robustness. Taking the case of subject 6: session1→session3 for example, we visualize the unlabeled samples from session 3 in a 2-D plane by the t-stochastic neighbor embedding (t-SNE) method in Fig. 4. As the legends show, four different colors respectively represent the four different emotional states. We can see there are two overlapping areas highlighted by rectangles.

According to the experimental paradigm in SEED-IV, after watching the film clip in one trial, there is a 45-second self-assessment. If a subject is overly immersed himself (herself) in this film and such self-assessment cannot completely make him (her) recover from the corresponding emotional state, this emotional state would serve as a *background* emotion in the next trial. This unexpected phenomenon, which is informally termed as 'feature-label inconsistency' in pattern recognition, easily causes the obtained EEG samples of the next trial cannot well depict the essential characteristics of their assigned emotional state. Specifically, we can consider that these two trials generated similar samples but were assigned different labels. This coincides with the overlaps shown in the rectangles and samples within the rectangles are undoubtedly difficult to differentiate. To improve the performance of recognition model, the most immediate approach is to decrease the weights of these samples during the model learning process. In this case, the threshold is set to 0.5, and the samples with weights less than 0.5 are circled. From Fig. 4, we find that many *sad* samples are mixed with *fear* samples; therefore, these samples are hard to differentiate and should be given smaller weights. As a result, the negative effect of noisy samples is reduced, leading to improved robustness.

In SEED-IV, the ground truth labels of trials in session 3 are [1, 2, 2, 1, 3, 3, 3, 1, 1, 2, 1, 0, 2, 3, 3, 0, 2, 3, 0, 0, 2, 0, 1, 0], where 0, 1, 2, and 3 respectively denote the *neutral*, *sad*, *fear*, and *happy* states. Obviously, there are four transitions between *sad* and *fear* states, which appear in the modes of '1, 2, 2, 1' and '1, 2, 1'. As stated above, these transitions easily cause the overlap of data distributions of *sad* and *fear* trials. This phenomenon suggests that it might be better to avoid the multiplicity of transitions of different emotion states in the design of video-evoked emotion EEG data collection experiments.

IV. Conclusion

In this brief, we proposed an efficient joint EEG sample and feature importance mining model sJSFE for semi-supervised emotion recognition. Specifically, two variables respectively characterizing the importance of samples and features were incorporated and adaptively learned in the semi-supervised least square regression model. Experimental results on the benchmark emotional EEG data set depicted that sJSFE made significant performance improvements in comparison with the baseline models which failed to consider the issues of sample and feature importance or only consider one of them. Besides, the critical EEG frequency bands and channels in emotion recognition were automatically explored by the learned feature importance vector in sJSFE. The sample importance descriptor suggested that reducing the number of emotional state transitions in EEG data collection experiment helps to improve the feature-label consistency of collected EEG data.

References

[1] X. Gao, Y. Wang, X. Chen, and S. Gao, "Interface, interaction, and intelligence in generalized brain-computer interfaces," *Trends Cognit. Sci.*, vol. 25, no. 8, pp. 671–684, 2021.

[2] N. S. Suhaimi, J. Mountstephens, and J. Teo, "EEG-based emotion recognition: A state-of-the-art review of current trends and opportunities," *Comput. Intell. Neurosci.*, vol. 2020, pp. 1–19, Sep. 2020, Art. no. 8875426. [Online]. Available: https://www.hindawi.com/journals/cin/2020/8875426/

[3] X. Chen *et al.*, "ReMAE: User-friendly toolbox for removing muscle artifacts from EEG," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 5, pp. 2105–2119, May 2020.

[4] W.-L. Zheng, J.-Y. Zhu, and B.-L. Lu, "Identifying stable patterns over time for emotion recognition from EEG," *IEEE Trans. Affect. Comput.*, vol. 10, no. 3, pp. 417–429, Jul./Sep. 2019.

[5] Z. Gao, T. Yuan, X. Zhou, C. Ma, K. Ma, and P. Hui, "A deep learning method for improving the classification accuracy of SSMVEP-based BCI," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 67, no. 12, pp. 3447–3451, Dec. 2020.

[6] D. Hu, J. Cao, X. Lai, Y. Wang, S. Wang, and Y. Ding, "Epileptic state classification by fusing hand-crafted and deep learning EEG features," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 68, no. 4, pp. 1542–1546, Apr. 2021.

[7] Y. Peng *et al.*, "Self-weighted semi-supervised classification for joint EEG-based emotion recognition and affective activation patterns mining," *IEEE Trans. Instrum. Meas.*, vol. 70, Nov. 2021, Art. no. 2517111. [Online]. Available: https://ieeexplore.ieee.org/document/9597546

[8] L. Jiang, D. Meng, T. Mitamura, and A. G. Hauptmann, "Easy samples first: Self-paced reranking for zero-example multimedia search," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 547–556.

[9] X. Chen, G. Yuan, F. Nie, and J. Z. Huang, "Semi-supervised feature selection via rescaled linear regression," in *Proc. Int. J. Conf. Artif. Intell.*, 2017, pp. 1525–1531.

[10] D. Wu, C.-T. Lin, and J. Huang, "Active learning for regression using greedy sampling," *Inf. Sci.*, vol. 474, pp. 90–105, Feb. 2019.

[11] Y. Peng, X. Zhu, F. Nie, W. Kong, and Y. Ge, "Fuzzy graph clustering," *Inf. Sci.*, vol. 571, pp. 38–49, Sep. 2021.

[12] W.-L. Zheng, W. Liu, Y. Lu, B.-L. Lu, and A. Cichocki, "EmotionMeter: A multimodal framework for recognizing human emotions," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 1110–1122, Mar. 2019. [Online]. Available: https://bcmi.sjtu.edu.cn/ seed/seed-iv.html

[13] Y.-F. Li and Z.-H. Zhou, "S4VM: Safe semi-supervised support vector machine," May 2011, *arXiv:1005.1545*.

[14] Y. Peng, F. Qin, W. Kong, Y. Ge, F. Nie, and A. Cichocki, "GFIL: A unified framework for the importance analysis of features, frequency bands and channels in EEG-based emotion recognition," *IEEE Trans. Cognit. Devel. Syst.*, early access, May 21, 2021, doi: 10.1109/TCDS.2021.3082803.

[15] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Trans. Auton. Mental Develop.*, vol. 7, no. 3, pp. 162–175, Sep. 2015.