A Factorized Extreme Learning Machine and Its Applications in EEG-Based Emotion Recognition

Yong Peng[1,2,3(✉)], Rixin Tang[2], Wanzeng Kong[2], and Feiping Nie[1]

[1] Center for OPTIMAL, Northwestern Polytechnical University, Xi'an 710072, China
yongpeng@hdu.edu.cn
[2] School of Computer Science, Hangzhou Dianzi University, Hangzhou 310018, China
[3] Provincial Key Laboratory for Computer Information Processing Technology, Soochow University, Suzhou 215006, China

Abstract. Extreme learning machine (ELM) is an efficient learning algorithm for single hidden layer feed forward neural networks. Its main feature is the random generation of the hidden layer weights and biases and then we only need to determine the output weights in model learning. However, the random mapping in ELM impairs the discriminative information of data to certain extent, which brings side effects for the output weight matrix to well capture the essential data properties. In this paper, we propose a factorized extreme learning machine (FELM) by incorporating another hidden layer between the ELM hidden layer and the output layer. Mathematically, the original output matrix is factorized so as to effectively explore the structured discriminative information of data. That is, we constrain the group sparsity of data representation in the new hidden layer, which will be further projected to the output layer. An efficient learning algorithm is proposed to optimize the objective of the proposed FELM model. Extensive experiments on EEG-based emotion recognition show the effectiveness of FELM.

Keywords: Extreme learning machine · Factorized representation · Group sparsity · Emotion recognition · EEG

1 Introduction

ELM is an efficient training algorithm for single hidden layer feed forward neural networks (SLFNNs) in which the input weights are randomly generated and the output weights can be analytically obtained [6]. Compared with the back propagation-based network weights tuning methods, the tedious process of iterative parameter tuning is eliminated and the problems including slow convergence speed and local minima are avoided. From the perspective of model optimization, the consistency of ELM, SVM, least square SVM and proximal SVM has been fully investigated [5]. ELM provides us a unified solution to generalized SLFNNs, including but not limited to neural networks, support vector networks and regularized networks [5].

H. Yang et al. (Eds.): ICONIP 2020, CCIS 1333, pp. 11–20, 2020.
https://doi.org/10.1007/978-3-030-63823-8_2

In recent years, a lot of efforts have been made on ELM from perspectives of theory and application. Huang et al. proposed the incremental ELM to enhance the universal approximation performance of SLFNNs, which can randomly select hidden nodes and adjust the output weights accordingly [3,4]. Zong et al. applied ELM as a ranking algorithm from the pointwise and pairwise perspectives [12]. In order to reduce the influence of outliers, Horata et al. proposed a robust ELM [2]. A fuzzy ELM was proposed to make different contributions to the learning of output weights through inputs with different fuzzy matrices [11]. To simultaneously utilize the benefits of ℓ_1 and ℓ_2 norms, an elastic net regularized ELM was proposed to perform EEG-based vigilance estimation [10]. As a feature extraction model, the discriminative extreme learning machine with supervised sparsity preserving in which the constraints were imposed on the output weights to preserve the sparsity achieved promising performance in data classification [8]. Besides, ELM has been widely employed in diverse fields such as face recognition, human action recognition, speaker recognition and data privacy.

However, the random generation of input weights may cause some distortions to the ELM hidden layer data representation in comparison with the original structure information of data. Therefore, when given complicated data sets, it will be hard to obtain a well-formed output weight matrix to get good generalization performance. To this end, we propose a structured matrix factorized extreme learning machine (FELM) in this paper. Our FELM model acts as the matrix factorization on the output weight matrix by introducing another hidden layer in which we enforce the group sparsity representation of data to achieve local dependencies of hidden units. Particularly, the mixed-norm regularization (ℓ_1/ℓ_2 norm) is incorporated in the model to obtain the group sparsity. We verify the ability of FELM on EEG-based emotion recognition task. Experimental results demonstrate that it can obtain better performance than SVM and ELM.

2 The Proposed FELM Model

Our proposed FELM model keeps the randomly generated input weights and hidden biases unchanged as those of ELM. The difference between FELM and ELM is the introduction of another hidden layer between the original ELM hidden layer and the output layer, which works as partitioning the original output weight matrix into two matrices. Then, we can enforce the data representation in the newly added hidden layer to have desirable properties which are beneficial for improving the learning performance.

As shown in Fig. 1, FELM includes the input layer, hidden layer H1, hidden layer H2 and output layer. The hidden layer H2 is the newly added one. Let $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N] \in \mathbb{R}^{D\times N}$ represent the input data, where D is the number of features and N is the number of samples. The number of input units is D. $\mathbf{w}_i$ and b_i represent the input weights and hidden bias respectively. They are both randomly determined. Let $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_P] \in \mathbb{R}^{D\times P}$ represent the input weight matrix. P is the number of units in the hidden layer H1. $\mathbf{a}_i$ indicate the input weights of the hidden layer H2, let $\mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_P]^T \in \mathbb{R}^{P\times Q}$ represent the corresponding input weight matrix. $\mathbf{b}_j$ indicate the output weights of the

hidden layer H2 and $\mathbf{B} = [\mathbf{b}_1, \cdots, \mathbf{b}_Q]^T \in \mathbb{R}^{Q \times C}$ represent the corresponding output weight matrix. Q is the number of units in the hidden layer H2. $f(\mathbf{X})$ indicate the output data. Let $\mathbf{T} = [\mathbf{t}_1, \cdots, \mathbf{t}_N] \in \mathbb{R}^{C \times N}$ indicate the expected output data. C represents the number of categories and the output units.

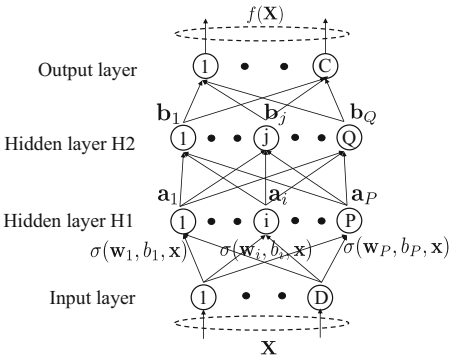

Fig. 1. Schematic diagram of the FELM model.

Specifically, the representation of the hidden layer H1 can be calculated as $\mathbf{h}(\mathbf{x}) = \sigma(\mathbf{w}_i, b_i, \mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x} + b)$, where $\sigma(a) = \frac{1}{1+e^{-a}}$ is the activation function in sigmoid form. The matrix form representation in H1 can be denoted as $\mathbf{H} = [\mathbf{h}(\mathbf{x}_1), \cdots, \mathbf{h}(\mathbf{x}_N)] \in \mathbb{R}^{P \times N}$. Then, the representation in hidden layer H2 can be obtained by $\hat{\mathbf{H}} = \phi(\mathbf{A}^T\mathbf{H})$. If $\phi(a) = a$ represents a linear function, equation (2) is equivalent to $\hat{\mathbf{H}} = \mathbf{A}^T\mathbf{H}$. The mapping relationship between the hidden layer H2 and the output layer is $f(\mathbf{X}) = \mathbf{B}^T\hat{\mathbf{H}}$.

For convenience, let $\mathcal{H} = \{1, 2, \cdots, Q\}$ denote the set of all units in the hidden layer H2. $\mathcal{H}$ can be partitioned into G groups and the gth group is represented by $\mathcal{G}_g$, where $\mathcal{H} = \cup_{g=1}^{G}\mathcal{G}_g$ and $\cap_{g=1}^{G}\mathcal{G}_g = \emptyset$. Therefore, $\hat{\mathbf{H}}$ can be expressed as $\hat{\mathbf{H}} = [\hat{\mathbf{H}}_{\mathcal{G}_1,:}; \cdots; \hat{\mathbf{H}}_{\mathcal{G}_g,:}; \cdots; \hat{\mathbf{H}}_{\mathcal{G}_G,:}]$. Therefore, the objective function of the FELM model can be expressed as follows:

$$\min_{\mathbf{A},\mathbf{B}} \ f = \|\mathbf{B}^T\hat{\mathbf{H}} - \mathbf{T}\|_F^2 + \alpha\Omega(\hat{\mathbf{H}}) + \beta\|\mathbf{B}\|_F^2, \tag{1}$$

where α is a regularization constant of the activation of the units in the hidden layer H2 and β is a regularization parameter of the hidden layer H2 output weight matrix $\mathbf{B}$. $\Omega(\hat{\mathbf{H}})$ represents the imposed penalty on sparse representations $\hat{\mathbf{H}}$.

Luo et al. [7] pointed out that group sparse representation can learn the statistical dependencies between hidden units, thereby improving model performance. Therefore, in order to implement the dependencies, we divide the units in the hidden layer H2 into non-overlapping groups on average to limit the dependencies within these groups and constrain the hidden units in a group to compete with each other. In addition, a mixed-norm regularization (ℓ_1/ℓ_2-norm) can achieve group sparse representation. So, we conduct the mixed-norm regularization $\Omega(\hat{\mathbf{H}}) = \sum_{g=1}^{G}\|\hat{\mathbf{H}}_{\mathcal{G}_g,:}\|_{1,2}$, where $\hat{\mathbf{H}}_{\mathcal{G}_{g,:}}$ is the representation matrix

associated to the data within modality belonging to the gth group. The ℓ_1/ℓ_2-norm can be expressed as $\|\hat{\mathbf{H}}_{\mathcal{G}_g,:}\|_{1,2} = \sum_{i=1}^{N} \sqrt{\sum_{j\in\mathcal{G}_g} \hat{h}_{j,i}^2}$.

Objective (1) has two variables, $\mathbf{A}$ and $\mathbf{B}$. We can alternately optimize one with the other fixed.

1) Update $\mathbf{B}$. The objective $\mathcal{O}(\mathbf{B})$ is $\min_{\mathbf{B}} \ f = \|\mathbf{B}^T\hat{\mathbf{H}} - \mathbf{T}\|_F^2 + \beta\|\mathbf{B}\|_F^2$, which is a convex optimization problem with the closed-form solution as

$$\mathbf{B} = (\hat{\mathbf{H}}\hat{\mathbf{H}}^T + \beta\mathbf{I})^{-1}\hat{\mathbf{H}}\mathbf{T}^T. \tag{2}$$

2) Update $\mathbf{A}$. The objective $\mathcal{O}(\mathbf{A})$ is $\min_{\mathbf{A}} \ f = \|\mathbf{B}^T\hat{\mathbf{H}} - \mathbf{T}\|_F^2 + \alpha\Omega(\hat{\mathbf{H}})$. We can use a gradient descent algorithm to solve the above squared error objective. By deriving the gradient, we obtain

$$\frac{\partial f}{\partial \mathbf{A}} = 2\mathbf{H}[d\phi(\hat{\mathbf{H}}^T) \circ (\mathbf{B}\mathbf{B}^T\hat{\mathbf{H}} - \mathbf{B}\mathbf{T})^T] + 2\alpha\mathbf{H}[d\phi(\hat{\mathbf{H}}^T) \circ \hat{\mathbf{H}}^T \circ /\tilde{\mathbf{H}}^T], \tag{3}$$

where $\circ$ means element-wise multiplication, $\circ/$ means element-wise division. The element of $\tilde{\mathbf{H}}$ is denoted as $\tilde{h}_{j,i} = \sqrt{\sum_{j\in\mathcal{G}_g} \hat{h}_{j,i}^2}$. $d\phi(a)$ represents the gradient of the function $\phi(a)$. When it is a sigmoid function, $d\phi(a) = \sigma(a) \times (1 - \sigma(a))$ and when it is a linear function, $d\phi(a) = 1$. So in Eq. (3), $d\phi(\hat{\mathbf{H}}^T) = 1$. Equation (3) can be further simplified as

$$\frac{\partial f}{\partial \mathbf{A}} = 2\mathbf{H}(\mathbf{B}\mathbf{B}^T\hat{\mathbf{H}} - \mathbf{B}\mathbf{T})^T + 2\alpha\mathbf{H}(\hat{\mathbf{H}}^T \circ /\tilde{\mathbf{H}}^T). \tag{4}$$

So, the update rule of $\mathbf{A}$ using the gradient defined in (4) is $\mathbf{A} = \mathbf{A} - \epsilon\frac{\partial f}{\partial \mathbf{A}}$, where ϵ is a learning rate. We summarize the optimization of FELM in Algorithm 1.

Algorithm 1 The optimization to FELM objective in equation (1)

Input: Data $\mathbf{X}$, label $\mathbf{T}$, parameters $\theta = \{\alpha, \beta, P, Q, \epsilon, G\}$.
Output: The output weight matrix $\mathbf{A}$ in the hidden layer H1 and the output weight matrix $\mathbf{B}$ in the hidden layer H2.

```
Randomly initialize the input weights w_i, bias b_i, i = 1, 2, ..., P, A and B and fix w_i and b_i.
while not converged do
    Update B according to (2);
    Update A according to A = A − ε ∂f/∂A;
end while
```

The optimization of each variable in FELM is iterative. The objective function in terms of variable $\mathbf{B}$ is a convex function and the solution obtained to $\mathbf{B}$ is in closed-form. So the convergence of FELM mainly depends on the update rule of variable $\mathbf{A}$, which is based on the gradient descent method. As the number of iterations increases, the value of the objective function decreases along the gradient until it converges. We terminate the iteration when the objective function value $\frac{\|obj^{(t+1)} - obj^{(t)}\|_2}{\|obj^{(t)}\|_2} < 10^{-4}$ in the experiment.

The main complexity of FELM is the loop containing two blocks. The main cost lies in calculating the inverse of $Q \times Q$ matrices $\hat{\mathbf{H}}\hat{\mathbf{H}}^T + \beta\mathbf{I}$ for the updating of variable $\mathbf{B}$, which needs $O(Q^3)$ complexity in each iteration. For the updating to $\mathbf{A}$, we need $O(PQN)$ complexity to calculate in each iteration by a gradient descent method. As a whole, the complexity for FELM is $O(t(Q^3+PQN))$ where t is the number of iterations.

3 Experiments

In the experiments, we evaluate the effectiveness of FELM on emotion recognition from EEG signals. The publicly available three-class emotional EEG data set, SEED (http://bcmi.sjtu.edu.cn/~seed/), was used in our experiments. The differential entropy feature smoothed by the linear dynamic system is used due to its effectiveness in expressing the emotional effect [1,9]. The EEG data of each subject has three different sessions and there were about 3400 samples in each session. We perform experiments in three different paradigms, which are within-session and cross-session of the same subject, and cross-subject experiments.

We compare FELM with SVM and ELM in terms of the classification performance on the given EEG data. Linear kernel was used in SVM and the regularization parameter C was selected from 2^{-7} to 2^{10}. The regularization parameters α and β in FELM were chosen from 10^{-4} to 10^4. If the dimension of input data satisfied $D < 100$, the numbers of hidden units P and Q were searched from 100 to 500 with step size 100. If the dimension of input data satisfied $100 < D < 500$, P was chosen from 500 to 1000 and Q from 100 to 500. For simplicity, we set the learning rate ϵ to 0.01 and the group numbers G to 4. The input weights and bias in ELM were the same as FELM which has P hidden units.

A. Experimental Paradigm 1. In order to test the ability of FELM model to classify DE features on different frequency bands, we choose about 2000 samples from one session of each subject as training set, the rest within the same session as test set. Table 1 and 2 show the classification results of linear-SVM, ELM and FELM models using the differential entropy features of *delta*, *theta*, *alpha*, *beta* and *gamma* frequency bands as input, where the best results are highlighted in boldface. We can find that the classification accuracies of FELM are higher than those of ELM and SVM in most cases in Table 1. As shown in Table 2, the average classification accuracy of FELM in each of the five frequency bands is higher than that of ELM and SVM. In addition, the classification results on *beta* and *gamma* frequency bands are higher than those of other frequency bands, meaning that the variation of emotional states may be more closely related to these two frequency bands.

Table 3 shows the average confusion matrices of three models based on the 310-dimensional feature vector of all frequency bands. We can find that positive and neutral emotional states are easier to be identified than the negative state. The FELM model estimates the negative state more accurately than both SVM and ELM. The average classification accuracy of FELM for negative state is 60.31% which is much higher than those of ELM (55.43%) and SVM (58.73%).

Table 1. Emotion recognition accuracies for six subjects A, B, C, D, E and F.

A	Session 1			Session 2			Session 3		
	SVM	ELM	FELM	SVM	ELM	FELM	SVM	ELM	FELM
delta	49.93	53.76	**54.55**	37.57	**40.53**	**40.53**	46.75	48.84	**49.64**
theta	**60.26**	58.02	60.12	49.35	49.28	**51.66**	**58.31**	55.71	57.95
alpha	65.17	**66.62**	66.55	54.41	54.48	**56.58**	48.63	52.67	**58.16**
beta	**84.10**	81.29	81.00	**65.46**	63.08	64.96	57.15	63.08	**67.12**
gamma	81.50	83.02	**84.90**	67.27	68.06	**72.04**	59.54	60.77	**69.44**
Total	**82.59**	83.74	81.14	**75.65**	64.74	67.63	59.90	61.27	**65.39**
B	Session 1			Session 2			Session 3		
	SVM	ELM	FELM	SVM	ELM	FELM	SVM	ELM	FELM
delta	53.47	56.72	**57.15**	38.73	**48.12**	47.83	52.02	51.81	**52.75**
theta	57.59	60.12	**60.55**	55.92	57.15	**59.47**	52.38	59.18	**59.90**
alpha	72.83	82.01	**83.74**	65.75	64.45	**67.85**	65.10	70.30	**72.04**
beta	**90.17**	86.71	88.87	**69.44**	67.85	68.86	78.97	81.36	**82.01**
gamma	89.52	87.57	**91.26**	**70.66**	66.33	66.91	**77.24**	75.07	76.23
Total	**88.15**	85.84	87.79	65.82	70.01	**72.40**	71.82	73.63	**74.64**
C	Session 1			Session 2			Session 3		
	SVM	ELM	FELM	SVM	ELM	FELM	SVM	ELM	FELM
delta	50.79	54.26	**55.20**	35.77	39.38	**41.19**	44.73	**45.52**	42.34
theta	**69.44**	62.28	63.08	**49.57**	49.78	46.39	43.93	39.67	**46.60**
alpha	61.13	60.04	**63.15**	50.43	51.16	**53.32**	**49.21**	41.91	45.66
beta	**77.24**	70.38	72.25	90.03	**90.82**	90.39	58.60	47.83	**59.68**
gamma	76.37	72.04	**76.73**	**89.45**	85.04	89.38	59.18	53.61	**61.05**
Total	**76.52**	74.64	75.14	**91.11**	88.80	87.50	**61.20**	50.43	60.12
D	Session 1			Session 2			Session 3		
	SVM	ELM	FELM	SVM	ELM	FELM	SVM	ELM	FELM
delta	75.87	**77.67**	75.65	**60.33**	58.89	58.02	58.09	61.63	**63.08**
theta	**73.92**	69.22	69.94	56.00	57.88	**62.07**	55.78	54.62	**65.32**
alpha	70.16	80.78	**83.38**	80.56	76.01	**81.07**	80.27	**89.45**	87.79
beta	92.99	**96.10**	93.71	88.08	91.04	**95.38**	**97.18**	95.23	96.03
gamma	90.68	93.93	**95.23**	91.98	92.49	**94.58**	**96.32**	95.74	95.88
Total	96.68	96.10	**96.89**	91.04	**96.89**	**96.89**	97.25	**97.40**	95.30
E	Session 1			Session 2			Session 3		
	SVM	ELM	FELM	SVM	ELM	FELM	SVM	ELM	FELM
delta	**58.89**	51.16	50.65	**55.85**	55.06	53.97	48.70	49.28	**50.00**
theta	**66.47**	63.29	64.09	40.25	50.07	**58.53**	40.10	43.35	**43.79**
alpha	46.89	54.48	**59.39**	34.39	40.17	**44.15**	60.69	63.15	**66.40**
beta	67.12	71.53	**75.14**	53.90	65.97	**72.18**	63.08	67.34	**78.11**
gamma	76.59	77.60	**80.27**	70.66	70.88	**73.63**	63.29	**65.53**	64.60
Total	70.01	69.87	**72.18**	60.19	68.50	**69.51**	73.99	67.20	**76.81**
F	Session 1			Session 2			Session 3		
	SVM	ELM	FELM	SVM	ELM	FELM	SVM	ELM	FELM
delta	**69.65**	64.88	64.31	**45.16**	34.75	42.85	**55.85**	53.25	53.61
theta	**58.24**	**58.24**	57.73	46.82	49.78	**51.52**	**63.44**	60.62	60.26
alpha	60.48	**62.64**	62.57	**53.11**	48.55	52.24	66.84	65.53	**67.63**
beta	73.19	77.53	**78.54**	59.25	53.25	**61.56**	88.29	90.68	**90.97**
gamma	69.80	82.01	**85.26**	58.82	57.30	**61.05**	93.86	91.26	**95.30**
Total	73.19	76.30	**78.76**	56.50	**58.96**	58.24	87.50	89.23	**90.10**

"Total" means concatenating features from all the five frequency bands.

Table 2. Average performances of different algorithms in paradigm 1 (mean ± std%).

Frequency band	mean ± std (%)		
	SVM	ELM	FELM
delta	52.12 ± 10.46	52.53 ± 9.90	**52.96 ± 8.94**
theta	55.43 ± 9.46	55.46 ± 7.32	**57.72 ± 7.05**
alpha	60.34 ± 12.07	62.47 ± 13.70	**65.09 ± 12.84**
beta	75.24 ± 14.00	75.62 ± 14.44	**78.71 ± 11.84**
gamma	77.35 ± 12.25	76.57 ± 13.07	**79.65 ± 12.28**
Total	76.62 ± 13.12	76.31 ± 13.92	**78.14 ± 12.10**

Table 3. Confusion matrices of different algorithms in paradigm 1 (mean ± std%).

SVM	Positive	Negative	Neural
Positive	**90.62 ± 10.15**	6.58 ± 7.75	2.80 ± 4.08
Negative	17.11 ± 15.91	**58.73 ± 31.53**	24.16 ± 22.28
Neural	9.86 ± 9.52	10.97 ± 11.68	**79.17 ± 16.24**
ELM	Positive	Negative	Neural
Positive	**91.22 ± 9.80**	5.10 ± 6.45	3.68 ± 5.20
Negative	17.65 ± 18.88	**55.43 ± 28.36**	26.92 ± 19.22
Neural	9.61 ± 14.76	9.65 ± 11.17	**80.75 ± 18.12**
FELM	Positive	Negative	Neural
Positive	**88.95 ± 9.95**	7.53 ± 7.41	3.52 ± 5.03
Negative	13.76 ± 13.55	**60.31 ± 27.19**	25.93 ± 20.50
Neural	6.62 ± 7.79	9.53 ± 9.95	**83.85 ± 14.29**

B. Experimental Paradigm 2. In order to identify the stable emotional patterns across different times, we choose the EEG data from one session of one subject as training set and the data from another session of such subject as test set. This paradigm can be termed as 'cross-session' emotion recognition. Table 4 shows the recognition results of each subject respectively obtained by linear-SVM, ELM and FELM models whose average performances are presented in Table 5. Here A1-A1 means that all the training and test samples are from the same session; specifically, we used the former 2000 of total 3400 samples as training and the rest as test, which follows the pipeline in [9]. We can find from Table 4 that the recognition accuracies of FELM is the highest in most cases. Generally, the classification accuracies by respectively choosing training and test samples from different sessions are significantly lower than choosing both training and test samples from the same session. This is caused by the non-stationary property of EEG data even if it was collected from the same subject but at different times. The average FELM classification accuracy for all subjects in the

experimental paradigm 1 is 78.14% while it is 67.44% in experimental paradigm 2. Nevertheless, 67.44% is still a relatively good result for the three-class emotion recognition task. This demonstrates that the transition of EEG patterns are stable among different sessions of the same subject.

Table 4. Emotion recognition accuracies (%) of different algorithms in paradigm 2.

		A1*	A2	A3		B1	B2	B3
SVM	A1	82.59	53.48	44.49	B1	**88.15**	32.62	54.10
ELM		**83.74**	**57.25**	**64.79**		85.84	59.75	57.75
FELM		81.14	55.24	57.04		87.79	**61.43**	**61.87**
SVM	A2	62.64	**75.65**	52.18	B2	65.09	65.82	67.47
ELM		59.66	64.74	47.53		59.05	70.01	**71.10**
FELM		**65.17**	67.63	**59.52**		**68.83**	**72.40**	69.51
SVM	A3	36.21	55.83	59.90	B3	**73.10**	44.49	71.82
ELM		55.42	50.65	61.27		65.91	62.32	73.63
FELM		**55.98**	**57.31**	**65.39**		71.66	**67.21**	**74.64**
ALG.s		C1	C2	C3		D1	D2	D3
SVM	C1	**76.52**	**80.41**	66.23	D1	96.68	80.55	83.50
ELM		74.64	80.32	66.82		96.10	82.20	**90.28**
FELM		75.14	78.99	**73.10**		**96.89**	**84.00**	82.12
SVM	C2	70.09	**91.11**	58.60	D2	**89.84**	91.04	**95.43**
ELM		70.86	88.80	**68.27**		85.50	**96.89**	91.54
FELM		**71.80**	87.50	63.20		84.03	**96.89**	93.64
SVM	C3	**79.29**	**81.35**	**61.20**	D3	**81.08**	**93.25**	97.25
ELM		56.48	72.04	50.43		79.17	87.71	**97.40**
FELM		78.61	78.34	60.12		78.99	88.80	95.30
ALG.s		E1	E2	E3		F1	F2	F3
SVM	E1	70.01	63.26	53.54	F1	73.19	**55.57**	**56.84**
ELM		69.87	59.78	48.26		76.30	54.71	53.89
FELM		**72.18**	**63.67**	**62.32**		**78.76**	51.94	53.95
SVM	E2	64.32	60.19	53.06	F2	61.82	56.50	69.27
ELM		46.32	68.50	**54.57**		**66.97**	**58.96**	65.00
FELM		**65.91**	**69.51**	52.83		61.14	58.24	**71.69**
SVM	E3	61.40	49.00	73.99	F3	43.69	52.53	87.50
ELM		43.25	44.61	67.20		59.58	49.18	89.23
FELM		**65.20***	**52.86**	**76.81**		**63.82**	**56.04**	**90.10**

"A1" is the first session of subject A. For example, the value 65.20 in bottom left corner is obtained by FELM in using E3 as training set and E1 as test set.

Table 5. Average performances of different algorithms in paradigm 2 (mean ± std%).

		Session 1	Session 2	Session 3
SVM	Session 1	81.19 ± 10.01	60.98 ± 18.20	59.78 ± 13.55
ELM		81.08 ± 9.44	65.67 ± 12.24	63.63 ± 14.75
FELM		$\mathbf{81.98 \pm 9.06}$	$\mathbf{65.88 \pm 12.90}$	$\mathbf{65.07 \pm 10.59}$
SVM	Session 2	68.97 ± 10.63	73.39 ± 15.15	66.00 ± 16.08
ELM		64.73 ± 13.20	74.65 ± 14.82	66.34 ± 15.20
FELM		$\mathbf{69.48 \pm 7.98}$	$\mathbf{75.36 \pm 14.19}$	$\mathbf{68.40 \pm 14.12}$
SVM	Session 3	62.46 ± 18.90	62.74 ± 19.75	75.28 ± 14.70
ELM		59.97 ± 11.97	61.09 ± 16.45	73.19 ± 17.55
FELM		$\mathbf{69.04 \pm 9.06}$	$\mathbf{66.76 \pm 14.27}$	$\mathbf{77.06 \pm 13.65}$

4 Conclusion

In this paper, we proposed an improved extreme learning machine model based on matrix factorization technique, termed as factorized ELM (FELM). This model performed matrix factorization on the ELM output weight matrix by adding an additional hidden layer between the hidden and output layers to mine the structured information of high-dimensional data. The group sparse representations was adopted to learn the local dependencies of hidden units. We applied FELM into emotion recognition from EEG signals. Based on the experimental results, we had three observations: 1) the EEG features from *beta* and *gamma* frequency bands might be more related to the transition of emotional states; 2) the positive emotional state is easier to recognize than the neutral and negative states; 3) there exist stable patterns in EEG features for performing cross-session recognition. In comparison with the baseline ELM and SVM, FELM obtained the best average classification performance.

Acknowledgments. This work was supported by NSFC (61971173,U1909202), Fundamental Research Funds for the Provincial Universities of Zhejiang (GK209907299001-008), Postdoctoral Science Foundation of China (2017M620470), Key Laboratory of Advanced Perception and Intelligent Control of High-end Equipment of Ministry of Education, Anhui Polytechnic University (GDSC202015) and Provincial Key Laboratory for Computer Information Processing Technology, Soochow University (KJS1841).

References

1. Duan, R.N., Zhu, J.Y., Lu, B.L.: Differential entropy feature for EEG-based emotion classification. In: NER, pp. 81–84 (2013)
2. Horata, P., Chiewchanwattana, S., Sunat, K.: Robust extreme learning machine. Neurocomputing **102**, 31–44 (2013)
3. Huang, G.B., Chen, L., Siew, C.K., et al.: Universal approximation using incremental constructive feedforward networks with random hidden nodes. IEEE Trans. Neural Netw. **17**(4), 879–892 (2006)

4. Huang, G.B., Li, M.B., Chen, L., Siew, C.K.: Incremental extreme learning machine with fully complex hidden nodes. Neurocomputing **71**(4–6), 576–583 (2008)
5. Huang, G.B., Zhou, H., Ding, X., Zhang, R.: Extreme learning machine for regression and multiclass classification. IEEE TSMC-B **42**(2), 513–529 (2012)
6. Huang, G.B., Zhu, Q.Y., Siew, C.K.: Extreme learning machine: theory and applications. Neurocomputing **70**(1–3), 489–501 (2006)
7. Luo, H., Shen, R., Niu, C., Ullrich, C.: Sparse group restricted Boltzmann machines. In: AAAI Conference on Artificial Intelligence, pp. 429–434 (2011)
8. Peng, Y., Lu, B.L.: Discriminative extreme learning machine with supervised sparsity preserving for image classification. Neurocomputing **261**, 242–252 (2017)
9. Peng, Y., Zhu, J.Y., Zheng, W.L., Lu, B.L.: EEG-based emotion recognition with manifold regularized extreme learning machine. In: EMBC, pp. 974–977 (2014)
10. Shi, L.C., Lu, B.L.: EEG-based vigilance estimation using extreme learning machines. Neurocomputing **102**, 135–143 (2013)
11. Zhang, W., Ji, H.: Fuzzy extreme learning machine for classification. Electron. Lett. **49**(7), 448–450 (2013)
12. Zong, W., Huang, G.B.: Learning to rank with extreme learning machine. Neural Process. Lett. **39**(2), 155–166 (2014)