Internal Feature Selection Method of CSP Based on L1-Norm and Dempster–Shafer Theory

Jing Jin, *Senior Member, IEEE*, Ruocheng Xiao, Ian Daly, Yangyang Miao, Xingyu Wang, and Andrzej Cichocki, *Fellow, IEEE*

***Abstract*— The common spatial pattern (CSP) algorithm is a well-recognized spatial filtering method for feature extraction in motor imagery (MI)-based brain–computer interfaces (BCIs). However, due to the influence of nonstationary in electroencephalography (EEG) and inherent defects of the CSP objective function, the spatial filters, and their corresponding features are not necessarily optimal in the feature space used within CSP. In this work, we design a new feature selection method to address this issue by selecting features based on an improved objective function. Especially, improvements are made in suppressing outliers and discovering features with larger interclass distances. Moreover, a fusion algorithm based on the Dempster–Shafer theory is proposed, which takes into consideration the distribution of features. With two competition data sets, we first evaluate the performance of the improved objective functions in terms of classification accuracy, feature distribution, and embeddability. Then, a comparison with other feature selection methods is carried out in both accuracy and computational time. Experimental results show that the proposed methods consume less additional computational cost and result in a significant increase in the performance of MI-based BCI systems.**

***Index Terms*— Brain–computer interface (BCI), common spatial pattern (CSP), feature selection, motor imagery (MI), spatial filtering.**

Manuscript received January 20, 2020; revised April 25, 2020; accepted August 5, 2020. This work was supported in part by the National Key Research and Development Program under Grant 2017YFB13003002, Grant2018YFC2002300, and Grant 2018YFC2002301; in part by the Grant National Natural Science Foundation of China under Grant 61573142 and Grant 61773164; in part the Program of Introducing Talents of Discipline to Universities through the 111 Project under Grant B17017; in part by the ShuGuang Project supported by the Shanghai Municipal Education Commission and the Shanghai Education Development Foundation under Grant 19SG25; in part by the Ministry of Education and Science of the Russian Federation under Grant 14.756.31.0001, and in part by the Polish National Science Center under Grant UMO-2016/20/W/NZG/00354. *(Corresponding author: Jing Jin.)*

Jing Jin, Ruocheng Xiao, Yangyang Miao, and Xingyu Wang are with the Key Laboratory of Advanced Control and Optimization for Chemical Processes, Ministry of Education, East China University of Science and Technology, Shanghai 200237, China (e-mail: jinjingat@gmail.com; xiaoruocheng96@163.com; miaoyy1991@163.com; xywang@ecust.edu.cn).

Ian Daly is with the Brain-Computer Interfacing and Neural Engineering Laboratory, School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, U.K. (e-mail: i.daly@essex.ac.uk).

Andrzej Cichocki is with the Skolkovo Institute of Science and Technology (Skoltech), 121205 Moscow, Russia, and also with the Nicolaus Copernicus University (UMK), 87-100 Torun, Poland (e-mail: a.cichocki@skoltech.ru).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2020.3015505

Nomenclature

$\mathbf{X}$	Electroencephalography (EEG) signals.
N	Number of channels.
T	Number of sampling points.
$c(\bar{c})$	Class index.
n	Number of trials in a class.
Σ	Spatial filter matrix of EEG signals.
J	Objective function.
w	Spatial filter (eigenvector).
$\mathbf{W}_{\text{csp}}$	Set of spatial filters in the feature space of CSP.
$\mathbf{Z}$	Projection signal.
λ	Eigenvalue.
m	Number of pairs of features.
f	Feature produced by CSP.
S_w	Interclass distance.
S_b	Within-class distance.
$y \in \{c, \bar{c}\}$	Class label.
$m(A)$	Mass function of the focal element "A."
ξ	Slack variable of SVM.
b	Bias of SVM.
C	Penalty parameter of SVM.
l	Average L1-norm of a class.
k	Position of a feature in descending order of the objective function value.

I. Introduction

BRAIN–COMPUTER interface (BCI) systems provide a novel communication path that allows humans to control external devices by using brainwaves only [1], [2]. In the past few decades, great progress has been achieved for BCI technology, and a range of BCI applications are beginning to find uses in daily life applications, such as wheelchair control, speller-based communication systems, and neurorehabilitation equipment [3]–[5]. With the advantages in terms of price and portability, electroencephalography (EEG) has become a common collection method in BCI systems. Currently, the most widely used paradigms in BCI systems include event-related potential (ERP), steady-state visual evoked potential (SSVEP), and motor imagery (MI). MI-based BCI systems depend on the phenomena of event-related synchronization (ERS) and event-related desynchronization (ERD), which occurs in the mu and beta rhythms during both actual movement and imaginary

movement [6]. Moreover, compared with other paradigms, MI-based BCI is more intuitive for users and does not rely on an external stimulus.

The common spatial pattern (CSP) algorithm is an effective spatial filtering method for feature extraction in MI-based BCI systems [7], [8]. CSP computes spatial filters in a data-driven manner, which maximizes the variance of one class while minimizing the variance of the other class [9]. However, CSP is easily affected by noise and is sensitive to parameters, such as specific EEG time window used, operational frequency band, and selected channels, which may produce suboptimal results [10], [11]. Several extensions have been proposed to address these problems. These solutions could be divided into the following three categories. The first category is summarized as an automatic selection for parameters. Filter bank CSP (FBCSP) [12] is proposed for selecting the optimal frequency bands automatically. The correlation-based time window selection (CTWS) algorithm [13] is used to choose a subject-specific time window for CSP. Correlation-based channel selection (CCS) algorithm [14] sorts the importance of channels based on the correlation among multichannel EEG signals and then selects channels according to the order. The methods of the second category change the goals of optimization. Regularized CSP (RCSP) [15]–[17] is a typical framework based on CSP, through which a variety of prior knowledge could be used in the optimization of spatial filters. CSP-L1 [18], [19] is another example that replaces L1-norm with L2-norm in the objective function. KLCSP [20] adds KL divergence to the optimization of CSP, in order to obtain spatial filters with minimum within-class dissimilarities. The third category uses information measure to select features. For example, Fisher's CSP (FCSP) [21] computes the Fisher ratio of each feature produced by CSP and then selects features based on the ratio size. In FBCSP, mutual information (MIN) is used for the selection of multiband fusion features. All the solutions improve the performance of the traditional CSP algorithm to varying degrees.

In this work, we will focus on feature selection for the feature space used within CSP. In the field of machine learning, feature selection refers to selecting some of the most effective features from the available feature set to reduce data set dimensions while improving algorithm performance. Feature selection methods can be divided into three categories: filter methods, wrapper methods, and embedded methods [22]. The feature selection methods mentioned earlier, including the Fisher ratio and MIN, are both filter methods. Meng *et al.* [6] proposed a wrapper method where a series of feature subsets is input to the classifier, and then, the best feature subset is selected corresponding to the maximum classification accuracy. Least absolute shrinkage and selection operator (Lasso) is a commonly used embedded feature selection method. Kumar *et al.* [23] used a lasso to select features from the tangent space, which is produced by the mapping of the manifold of spatial covariance matrices.

In the traditional CSP algorithm, the objective function is represented as a Rayleigh quotient between the class average covariance matrices, which is equal to the ratio of the class average power of the EEG signals [24]. The spatial filters, which produce features, are formed by the eigenvectors corresponding to the maximum and minimum eigenvalues of a joint covariance matrix appearing in the optimization of the Rayleigh quotient, and the selected eigenvalues are the extreme values of the CSP objective function. Hence, as shown in Fig. 1, traditional CSP can be viewed as consisting of three phases: extraction, internal selection, and generation. The extraction phase is represented as obtaining eigenvectors from a joint spatial covariance matrix, while the internal selection phase is described as selecting spatial filters corresponding to the maximum and minimum objective function values. Finally, the generation phase refers to projecting EEG signals through spatial filters and then using the logarithm variance of the projection signals as features.

However, there are some drawbacks to the CSP objective function. One is that the feature distribution will not be taken into consideration when only using average values, which has a significant impact on the classification accuracy. To address this problem, both Fisher's ratio spatial pattern (FSSP) [25] and FCSP use the Fisher ratio to analyze the distribution of features. The difference between the two is during the extraction phase, where the former uses the Fisher ratio as the objective function and the latter uses the CSP objective function.

This difference also leads to changes in the feature space. Since the objective function of FSSP cannot be expressed by the Rayleigh quotient, the optimization of the spatial filter must be based on an iterative algorithm (e.g., gradient descent), which increases the computational time required by the method considerably, with the final results depending, in part, on the initial value used at the start of the search process. Although FCSP is almost the same as the traditional CSP in computational time, in the feature space produced by CSP, only using the Fisher ratio to discriminate features, sometimes, yields worse results than those achieved by the traditional CSP algorithm. Another drawback is that the CSP objective function only maximizes the ratio of the two-class average power, without considering the difference between the two classes. Thus, even if there is a large power ratio between two classes, the interclass distance of the corresponding features is still short if the power difference is small. For example, consider 3/1 versus 8/3; though the former has a larger ratio, the interclass distance of the latter is clearly larger. In addition, the variance ratio (power ratio) is quite sensitive to outliers, which may result in the selected features that are not expected.

Since the selected features correspond to the extreme values of the CSP objective function, the abovementioned problems mean that the selected features may not be optimal in the feature space used within CSP. In other words, features not selected may achieve better classification results.

To address the abovementioned issue, while retaining the efficiency of CSP algorithm, we suggest still using CSP in the extraction phase then changing the feature selection rules of the internal selection phase based on an improved objective function. This work made the following contributions.

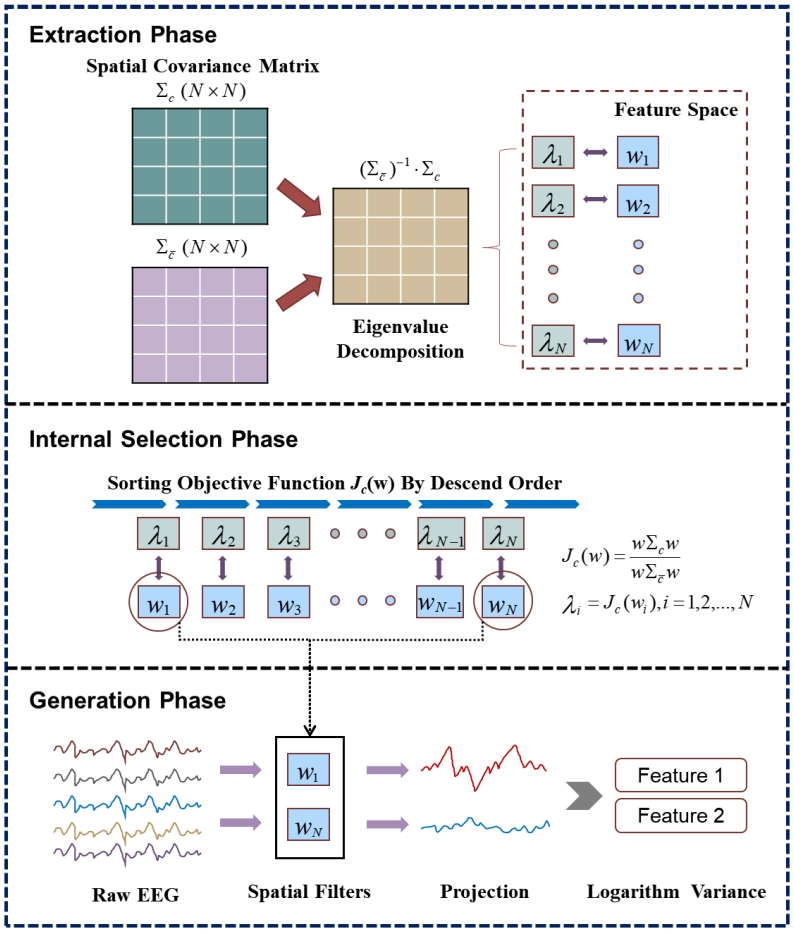

Fig. 1. Schematic of the traditional CSP algorithm.

1) A new design is proposed for the objective functions to solve defects inherent in traditional CSP. Especially, improvements are made in suppressing outliers and discovering features with larger interclass distances.
2) A new fusion algorithm for internal feature selection based on the Dempster–Shafer theory (DST) is proposed, which fuses different feature selection rules.

II. Preliminaries for Internal Feature Selection

A. Common Spatial Pattern

To explain the variables more clearly, the relevant nomenclature is shown in Nomenclature. The CSP algorithm is an effective spatial filtering method commonly used to extract features in MI-based BCI systems. The spatial filters are considered as projection vectors and are calculated to maximize the variance of one class while minimizing the variance of the other class. Consider two classes of EEG signals $X_{i,1}$, $X_{i,2} \in R^{N\times T}$ from the experimental ith trial, where N is the number of channels and T denotes the number of sampling points. The spatial covariance matrix of class c is computed as follows:

$$\mathbf{\Sigma}_c = \frac{1}{n_c}\sum_{i=1}^{n_c}\mathbf{X}_{i,c}\mathbf{X}_{i,c}^T \tag{1}$$

where n_c represents the number of trials in class c. Then, the spatial filter that maximizes the variance of one class and minimizes the variance of the other can be calculated by

$$J_C(w) = \frac{w^T\mathbf{\Sigma}_c w}{w^T\mathbf{\Sigma}_{\bar{c}} w} \quad \text{s.t. } \|w\|_2 = 1 \tag{2}$$

where w is the spatial filter. The optimization of the Rayleigh quotient can be converted to the generalized eigenvalue problem

$$\mathbf{\Sigma}_c w = \lambda\mathbf{\Sigma}_{\bar{c}} w \tag{3}$$

where λ and w are the generalized eigenvalue and eigenvector, respectively. The spatial filters $\mathbf{W}_{\text{csp}}$ are formed by eigenvectors corresponding to m maximum and minimum eigenvalues.

The projection signal $\mathbf{Z}$ of the single trial is given by

$$\mathbf{Z} = \mathbf{W}_{\text{csp}}^T\mathbf{X}. \tag{4}$$

Then, the pth feature of the single trial can be obtained as follows:

$$f^p = \log\left(\mathbf{Z}_p\mathbf{Z}_p^T\right) \tag{5}$$

where $\mathbf{Z}_p$ is the pth row of $\mathbf{Z}$ ($p = 1, 2, \ldots, 2m$).

B. Dempster–Shafer Theory

DST, which was established by Dempster [26] and improved by Shafer [27], is also known as evidence theory and was

first used in expert systems. DST may be considered to be a generalization of classical probability theory, which is a framework that combines evidence from different sources [28]. A brief introduction to this theory is given as follows.

Let X be a finite set (frame of discernment) that contains all possible answers to a question. The power set of X is represented by 2^X, which includes all possible subsets of X. DST assigns a probability for each subset. This probability is called the basic probability assignment (BPA) or mass function, which satisfies the following conditions:

$$m(\phi) = 0 \tag{6}$$

$$\sum_{A \in 2^X} m(A) = 1 \tag{7}$$

$$0 \le m(A) \le 1, \quad A \in 2^X \tag{8}$$

where ϕ denotes the empty set. The focal element represents the subset of 2^X whose mass function is not zero. The focal element "A" is considered as a hypothesis, which may include one or more answers. DST defines a belief function and a plausibility function to determine the probability range of element "A," which is expressed by mass functions as follows:

$$\mathrm{Bel}(A) = \sum_{B \subseteq A} m(B) \tag{9}$$

$$Pl(A) = \sum_{B \cap A \neq \phi} m(B) \tag{10}$$

where "B" is another focal element, and "B" satisfies the relationship in (9) and (10) with "A." Therefore, the probability range of "A" (interval belief) is obtained by

$$\mathrm{Bel}(A) \le P(A) \le Pl(A). \tag{11}$$

Next, a framework provided by DST is presented, which combines independent evidence from different sources. Suppose that m_1 and m_2 are two mass functions associated with focal element "A" in the same frame of discernment, which are from different sources. The fusion rules of the two sets of evidence are as follows:

$$m_{1,2}(A) = \begin{cases} \dfrac{\sum_{B \cap C = A} m_1(B) \cdot m_2(C)}{1 - \sum_{B \cap C = \phi} m_1(B) \cdot m_2(C)}, & A \neq \phi \\ 0, & A = \phi. \end{cases} \tag{12}$$

Based on DST, a fusion algorithm for internal feature selection is proposed in Section III.

C. Classification Tool

The support vector machine (SVM) is used as a classification tool in this work. SVM finds a hyperplane to segment two classes of samples, and this hyperplane can be represented as $w^T x + b = 0$, where $w \in R^d$ denotes the weight vector and b denotes the bias [29]. The principle of the segmentation is to maximize margins between two classes and, finally, transform the problem into a convex quadratic programming problem [30], in which

$$\min_{b,w,\xi} \frac{1}{2}\|w\|^2 + C \cdot \sum_{i=1}^{n} \xi_i$$
$$\text{s.t. } y_i\left(w^T x^{(i)} + b\right) \ge 1 - \xi_i, \quad \xi_i > 0, \ i = (1, \ldots, n) \tag{13}$$

where $x^{(i)}$ denotes the feature vector of the ith training sample, ξ denotes the slack variable, C denotes the penalty parameter of the error term, and y denotes the class label [31]. A radial basis function (RBF) kernel was used in this work.

III. Design of Internal Feature Selection Methods

A. Internal Feature Selection Based on the Difference and Ratio of Average L1-Norm for CSP (DRL1-CSP)

Equation (2) can be rewritten as

$$J_C(w) = \frac{\frac{1}{n_c}\sum_{\substack{i=1 \\ y_i = c}}^{n_c} w^T \mathbf{X}_{i,c} \mathbf{X}_{i,c}^T w}{\frac{1}{n_{\bar{c}}}\sum_{\substack{j=1 \\ y_j = \bar{c}}}^{n_{\bar{c}}} w^T \mathbf{X}_{i,\bar{c}} \mathbf{X}_{i,\bar{c}}^T w} = \frac{\frac{1}{n_c}\sum_{\substack{i=1 \\ y_i = c}}^{n_c} \left\| w^T \mathbf{X}_{i,c} \right\|_2^2}{\frac{1}{n_{\bar{c}}}\sum_{\substack{j=1 \\ y_j = \bar{c}}}^{n_{\bar{c}}} \left\| w^T \mathbf{X}_{j,\bar{c}} \right\|_2^2}. \tag{14}$$

The objective function of the traditional CSP algorithm could also be considered as the ratio between the average L2-norm squared of the two classes, whose extreme values are equal to the λ obtained in (3). However, the L2-norm is quite sensitive to outliers, which can be replaced with a more robust L1-norm [18]. Hence, we replace the L2-norm squared ratio with the ratio of L1-norm in the internal selection phase of CSP, in which

$$l_c(w) = 1/n_c \sum_{\substack{i=1 \\ y_i = c}}^{n_c} \left\| w^T \mathbf{X}_{i,c} \right\|_1 \tag{15}$$

$$J(w) = \frac{l_c(w)}{l_{\bar{c}}(w)} \tag{16}$$

where $l(w)$ denotes the average L1-norm of a class of projection signals. Equation (16) only maximizes the ratio of two L1-norm, without considering the difference between the two, which may make the interclass distance of the selected features not the largest. This could be solved by modifying the objective function as the product of the difference and ratio between the average L1-norm of the two classes, in which

$$J_D(w) = (l_c(w) - l_{\bar{c}}(w)) \cdot \left(\frac{l_c(w)}{l_{\bar{c}}(w)}\right)^{\mathrm{sgn}\left(\frac{l_c(w)}{l_{\bar{c}}(w)} - 1\right)} \tag{17}$$

where $\mathrm{sgn}(\cdot)$ is the symbolic function. After the sequence sorted by $J_D(w)$ is obtained, we select features with the largest and smallest $J_D(w)$ values. In the traditional CSP algorithm, the maximum and minimum eigenvalues are not comparable, so the spatial filters are always formed by the same number of two types eigenvectors that correspond to the largest and smallest eigenvalues, respectively. However, in the proposed method, the objective function values correspond to all features that can be compared with each other after taking absolute values, in which

$$\tilde{J}_D(w) = \left| (l_c(w) - l_{\bar{c}}(w)) \cdot \left(\frac{l_c(w)}{l_{\bar{c}}(w)}\right)^{\mathrm{sgn}\left(\frac{l_c(w)}{l_{\bar{c}}(w)} - 1\right)} \right|. \tag{18}$$

In (18), only the features corresponding to the maximum values are selected, instead of being equally selected between the maximum and minimum values. The methods represented

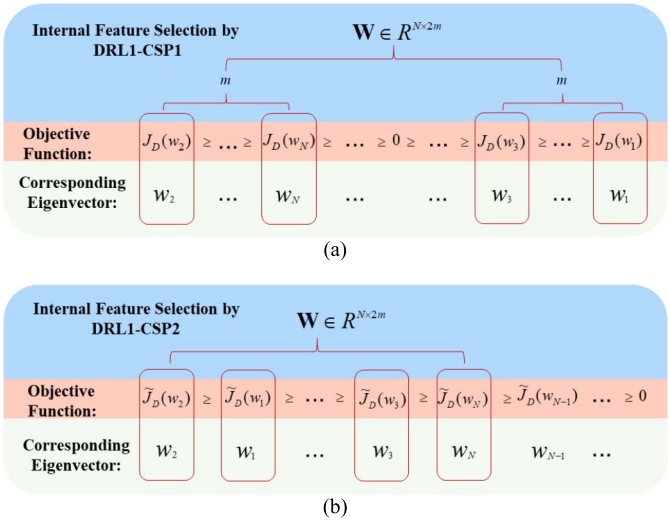

Fig. 2. Schematic of the internal selection phase for the proposed methods. Notice the position where the objective function is zero for each subgraph. (a) Internal feature selection by DRL1-CSP1. (b) Internal feature selection by DRL1-CSP2.

by (17) and (18) are named DRL1-CSP1 and DRL1-CSP2, respectively. Fig. 2(a) shows the schematic of the internal selection phase of DRL1-CSP1, while Fig. 2(b) shows the internal selection phase of DRL1-CSP2.

B. Fisher's Common Spatial Pattern

As mentioned in Section I, FCSP follows the CSP objective function in the extraction phase of CSP and then uses the Fisher ratio for each feature in the feature space, in which:

$$J_F(w) = \frac{S_w}{S_b} \tag{19}$$

where S_w and S_b are the interclass and within-class distances of a feature, respectively

$$S_w = (\mu_c - u_{\bar{c}})^2 \tag{20}$$

$$S_b = (\sigma_c)^2 + (\sigma_{\bar{c}})^2 \tag{21}$$

where μ_c is the mean of class c in a feature and $\mu_{\bar{c}}$ is that of the other class, $(\sigma_c)^2$ denotes the variance of class c in a feature, and $(\sigma_{\bar{c}})^2$ denotes that of the other class

$$\mu_c = \frac{1}{n_c} \sum_{\substack{i=1 \\ y_i=c}}^{n_c} f_i \tag{22}$$

$$(\sigma_c)^2 = \frac{1}{n_c} \sum_{\substack{i=1 \\ y_i=c}}^{n_c} (f_i - \mu_c)^2 \tag{23}$$

where y denotes the label. After calculating the Fisher ratio for all features, we select features with the largest Fisher ratio values.

C. Fusion Algorithm for Internal Feature Selection Based on Dempster–Shafer Theory

In Section I, we have discussed the advantages and disadvantages of using the Fisher ratio for feature selection.

Algorithm 1 Fusion Algorithm for Internal Feature Selection Based on DST

Input: Two classes training data set $X_{i,1}$, $X_{i,2} \in R^{N\times T}$, and the dimension of the required feature set N_r.
Output: Optimal spatial filters.
begin
Calculate all features of each trial by Eq. (1) to Eq. (5);
Calculate $\tilde{J}_D(w)$ of each feature by Eq. (18);
Calculate $J_F(w)$ of each feature by Eq. (19) to Eq. (23);
Calculate $m_D(\cdot)$ and $m_F(\cdot)$ of each feature by Eq. (25) and Eq. (26);
Calculate $m_{D.F}(\cdot)$ of each feature by Eq. (27);
Sort features by the descending order of fused mass function value, then select the first N_r eigenvectors as optimal spatial filters.
End

The instability of performance is where there is the greatest need for improvement. On the other hand, both traditional CSP and DRL1-CSP do not consider the feature distribution. To solve these problems, we use DST to fuse two different feature selection rules that are FCSP and DRL1-CSP.

There have been some examples of DST applied in the field of BCI. Without exception, these examples used DST to fuse the classification results of multiple classifiers [32], [33]. An important reason for this use is that the results of the classifier could be represented by a probabilistic structure, which could be directly used in DST. Therefore, if we want to use DST to fuse feature selection rules, we also need to convert the rules into probabilistic structures.

Assume that each feature in the feature space has been assigned weight. Since the sum of the normalized weight coefficients of all features is one, the normalized weight coefficient could be regarded as the probability that the feature is selected, which is thought of as the mass function value in the DST framework. Note that we only assign values to all the individual features in the power set, and all the mass functions of combined features are set to 0. In subsequent formulas, these zero items will no longer appear. This is done to simplify the parameters required for DST and then to simplify the calculation. We do not have enough prior knowledge to estimate the mass function of the combined features.

Next, we need to think about how to weight the features reasonably. A natural idea is to use the objective function value [i.e., the function value obtained by (18) or (19)] corresponding to the feature as a weight coefficient. However, this solution has a major defect: Since the values of different features are not of the same order of magnitude, it is possible that very few features contribute most of the weight, which will cause the weight of the remaining features to be too small.

The purpose of using DST is to comprehensively evaluate features from different views (i.e., different objective functions). If the distribution of a view's mass function values is severely polarized, the order of magnitude of the mass function value may not be changed after fusion, which will weaken the effect of DST.

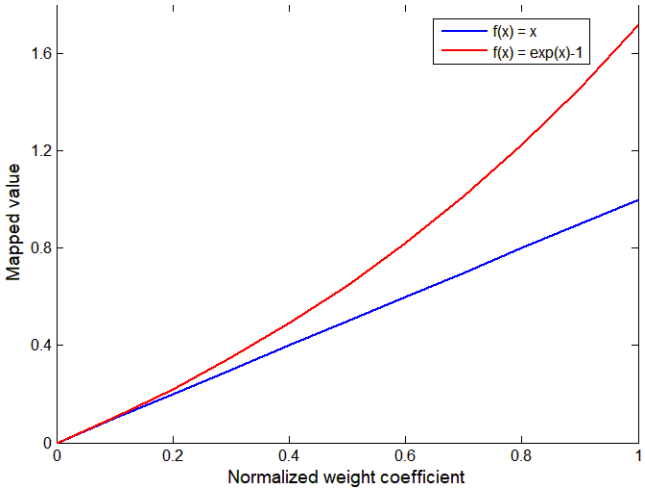

Fig. 3. Comparison diagram of different mapping relationships.

We notice that the objective function value is only used for comparison between different features, and its numerical value has no practical significance. What is important is the order of the features. Thus, the position labels of ordered feature sequences for the reverse assignment are used, which can also avoid the polarization of the distribution of the mass function. For example, after sorting the objective function in descending order of value, ten features are assigned values decreasing from 10 to 1, and then, the normalized weight coefficient is 10/55–1/55.

As mentioned earlier, the performance of FCSP is unstable, and the fundamental reason is that the Fisher ratio cannot reasonably evaluate the features in the cases of small interclass distances. In order to reduce the impact of this instability, we further modify the weighting strategy. Especially, a nonlinear mapping is added, in which

$$f(x) = \exp(x) - 1 \quad \text{s.t. } 0 \le x \le 1 \tag{24}$$

where x denotes the normalized weight coefficient obtained by reverse assignment. As shown in Fig. 3, through the nonlinear mapping, larger weights will increase, and smaller weights will decrease. For example, the normalized sequence of weight coefficients [0.4, 0.3, 0.2, 0.1], after being nonlinearly mapped and normalized, becomes [0.42, 0.30, 0.19, 0.09]. According to (19), the interclass distance is positively correlated with the Fisher ratio. Therefore, the nonlinear mapping reduces the weight of those features whose interclass distance may be small, thereby suppressing the probability that such features are selected.

The complete assignment strategy for the mass function (normalized weight coefficient) of the pth feature is as follows:

$$m_D(p) = \frac{1}{\sum_{i=1}^{n} i}(N + 1 - k^p) \tag{25}$$

$$m_F(p) = \frac{\exp\left(\frac{1}{\sum_{i=1}^{n} i}(N + 1 - k^p)\right) - 1}{\sum_{j=1}^{N} \exp\left(\frac{1}{\sum_{i=1}^{n} i}(N + 1 - k^j)\right) - N} \tag{26}$$

where $m_D(\cdot)$ and $m_F(\cdot)$ denote the mass function of DRL1-CSP and FCSP, respectively, and k denotes the position of a feature in descending order of the objective function value. N denotes the number of channels. Note that only DRL1-CSP2 is used in (25) since not all objective function values of DRL1-CSP1 can be compared with each other.

After completing the assignment of the mass function to all features, the framework of DST could be used, in which

$$\begin{aligned} m_{D\cdot F}(p) &= \frac{\sum_{B\cap C=p} m_D(B) \cdot m_F(C)}{1 - \sum_{B\cap C=\phi} m_D(B) \cdot m_F(C)} \\ &= \frac{m_D(p) \cdot m_F(p)}{\sum_{\substack{1\le i,j\le N \\ i=j}} m_D(i) \cdot m_F(j)}. \end{aligned} \tag{27}$$

Finally, sort features by descending order of $m_{D\cdot F}(\cdot)$, and select features from front to back in the sequence. The whole framework of the proposed fusion algorithm is shown in Fig. 4, and the flow of the fusion algorithm is shown in Algorithm 1.

IV. Experimental Study

A. Description of EEG Data

To evaluate the performance of the proposed methods for internal feature selection, two public data sets from the BCI Competition have been used in this study:

Data Set 1 (BCI Competition III Data Set IVa) [34]: This data set was recorded from five healthy participants who performed MI of their right hand and foot during cued trials; each participant conducted 280 trials (half for each class). The data were measured from 118 channels and subsampled at 100 Hz. In each trial, a visual cue was displayed on screen for 3.5 s to indicate to participants to perform MI. Participants were then given a rest time between 1.75 and 2.25 s. More details can be found at http://www.bbci.de/competition/iii/.

Data Set 2 (BCI Competition IV Data Set I) [35]: This data set includes seven healthy participants, each participant conducted 200 trials (half for each class) without feedback. In each trial, there was a fixation cross at the center of the screen for the first 2 s. Then, an arrow pointing left or right or down was displayed on the screen for 4 s, which cued participants to perform corresponding MI. Finally, a blank screen was displayed for 2 s. This data set was measured from 59 channels and subsampled at 100 Hz. See website http://www.bbci.de/competition/iv/ to find more details. Note that there were three artificially generated participants (named "c," "d," and "e"), and we only use the remaining four data sets recorded from real participants for testing.

B. Preprocessing and Experiment Setting

In this work, data set 1 has been extracted from 0.5 to 2.5 s for each trial after the visual cue, while data set 2 has been extracted from 0 to 3 s for each trial after the visual cue. The raw EEG signals were filtered with a fifth-order Butterworth filter between 8 and 30 Hz. Tenfold cross-validation was used in all experiments. Please note that the accuracy of the algorithm shown in the experiment is the highest accuracy that the algorithm can achieve under a different number of features.

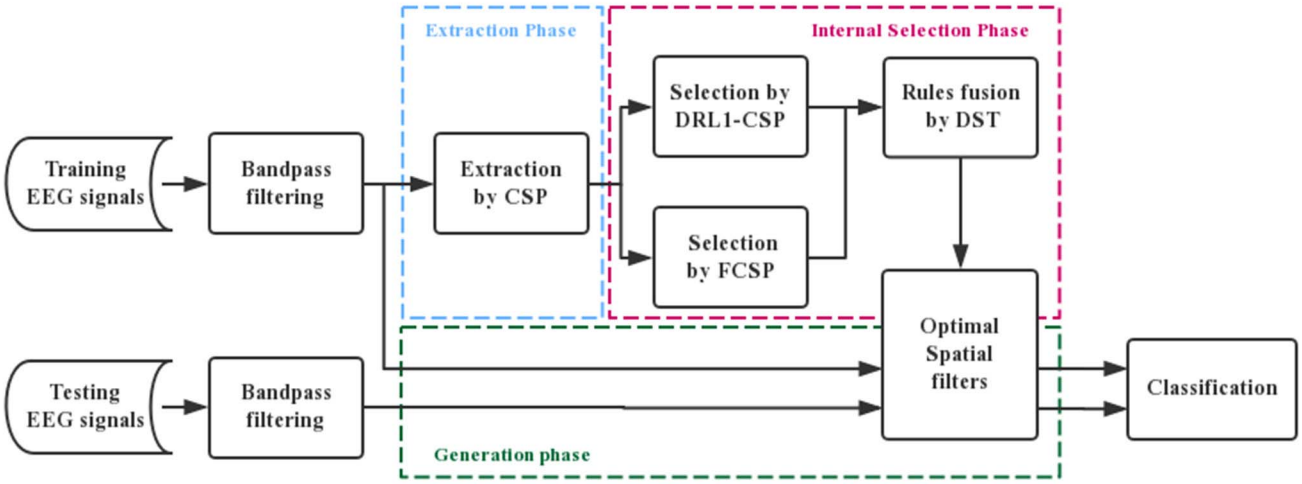

Fig. 4. Block diagram of the proposed fusion algorithm.

TABLE I
COMPARISON OF CLASSIFICATION ACCURACY (%) AMONG PROPOSED METHODS AND TRADITIONAL CSP

Participant	Methods		
	CSP	DRL1-CSP1	DRL1-CSP2
aa	77.1	**83.2**	77.5
al	**96.8**	**96.8**	**96.8**
av	49.2	68.9	**69.3**
aw	85.0	91.1	**92.5**
ay	88.6	**93.6**	89.6
Mean±std	79.3±18.3	**86.7±11.2**	85.1±11.4
a	54.5	64.5	**66.5**
b	54.5	**62.0**	58.5
f	47.0	60.5	**63.5**
g	62.5	**85.0**	**85.0**
Mean±std	54.6±6.3	68.0±11.5	**68.4±11.6**
p-value	-	0.003	0.013

C. Performance of Proposed Objective Function in Internal Feature Selection of CSP

1) Comparison of Classification Accuracy: To evaluate the performance of the proposed methods, we compare the classification accuracies among the proposed methods and traditional CSP for the two data sets. As shown in Table I, the proposed methods achieve higher classification accuracies in all cases, which indicates that the proposed methods could select better features at the same feature space. Especially, in data set 1, the average classification accuracies are 79.3% (with CSP), 86.7% (with DRL1-CSP1), and 85.1% (with DRL1-CSP2); the average classification accuracies of data set 2 are 54.6% (with CSP), 68.0% (with DRL1-CSP1), and 68.4% (with DRL1-CSP2). Test results from two data sets indicate that the proposed methods are superior to traditional CSP (paired t-test, $p < 0.05$, with eight degrees of freedom).

2) Embeddability Analysis of the Proposed Methods: Since the proposed methods only change the rules of feature selection inside the CSP, it can be easily embedded into other types of extensions of CSP. In this work, we embed the proposed methods in CSP-rank [36]. CSP-rank is a filtering channel selection algorithm [37], which is used to remove redundant channels while improving the performance of BCI systems. It sorts the importance of channels according to the coefficients of spatial filters, and the optimal channel sets correspond to the highest cross-validation accuracy. The algorithms after embedding are named DRL1-CSP1-rank and DRL1-CSP2-rank, respectively. The classification accuracies and their corresponding numbers of channels are shown in Table II, and the classification accuracies of the proposed methods are significantly higher than that of CSP-rank (paired t-test, $p < 0.05$, with eight degrees of freedom), which proves the effectiveness and practicality of the proposed methods.

3) Comparison of Feature Distributions: To further observe the differences in features selected by the different methods, the feature distributions are displayed from all participants except "al." Due to dimensional constraints, only the two features preferred by each method are shown. Note that each subgraph uses all the trials belonging to a participant to plot the feature distributions. As shown in Figs. 5 and 6, it is obvious that the features selected by the proposed methods have feature distribution that can be more easily discriminated than those produced by traditional CSP.

Participant "aw" is used here to explain the differences in selected features between CSP and DRL1-CSP1. Especially, the highest accuracy of both algorithms is achieved when $m = 1$ (i.e., the number of features is 2). As shown in Fig. 5, after calculation, it is found that the features corresponding to the horizontal axis of the two are the same, and the difference lies in the vertical axis, and we use f_C and f_D to represent the features selected by the vertical axis of CSP and DRL1-CSP1, respectively. In traditional CSP, the feature selected by the vertical axis corresponds to the minimum value of the CSP objective function, so the CSP objective function value of f_C is smaller than f_D ($J_C(f_C) = 0.412$ versus $J_C(f_D) = 0.514$). In DRL1-CSP1, the feature selected by the vertical axis also corresponds to the minimum value of the DRL1-CSP1 objective function. Thus, the DRL1-CSP1 objective of f_D is smaller than f_C ($J_D(f_C) = -231.7$ versus $J_D(f_D) = -236.7$), which also indicates that f_D has a larger

TABLE II

COMPARISON OF CLASSIFICATION ACCURACY (%) AND SELECTED NUMBER OF CHANNELS FOR DIFFERENT METHODS

Participant	Methods					
	CSP-rank		DRL1-CSP1-rank		DRL1-CSP2-rank	
	Acc(%)	Num	Acc(%)	Num	Acc(%)	Num
aa	83.6	15	**86.4**	22	85.4	20
al	97.5	52	**97.9**	83	**97.9**	11
av	57.5	37	73.2	72	**73.6**	51
aw	91.1	28	**92.5**	41	**92.5**	90
ay	92.9	46	**94.6**	38	93.9	22
Mean±std	84.5±15.9	36	**88.9±9.7**	51	88.7±9.6	39
a	66.5	6	75.5	15	**82.0**	15
b	61.0	7	61.5	14	**67.5**	25
f	60.0	8	**70.0**	28	65.0	47
g	81.0	33	**87.5**	37	**87.5**	19
Mean±std	67.1±9.7	14	73.6±10.9	24	**75.5±11.0**	27
p-value	-		0.017		0.017	

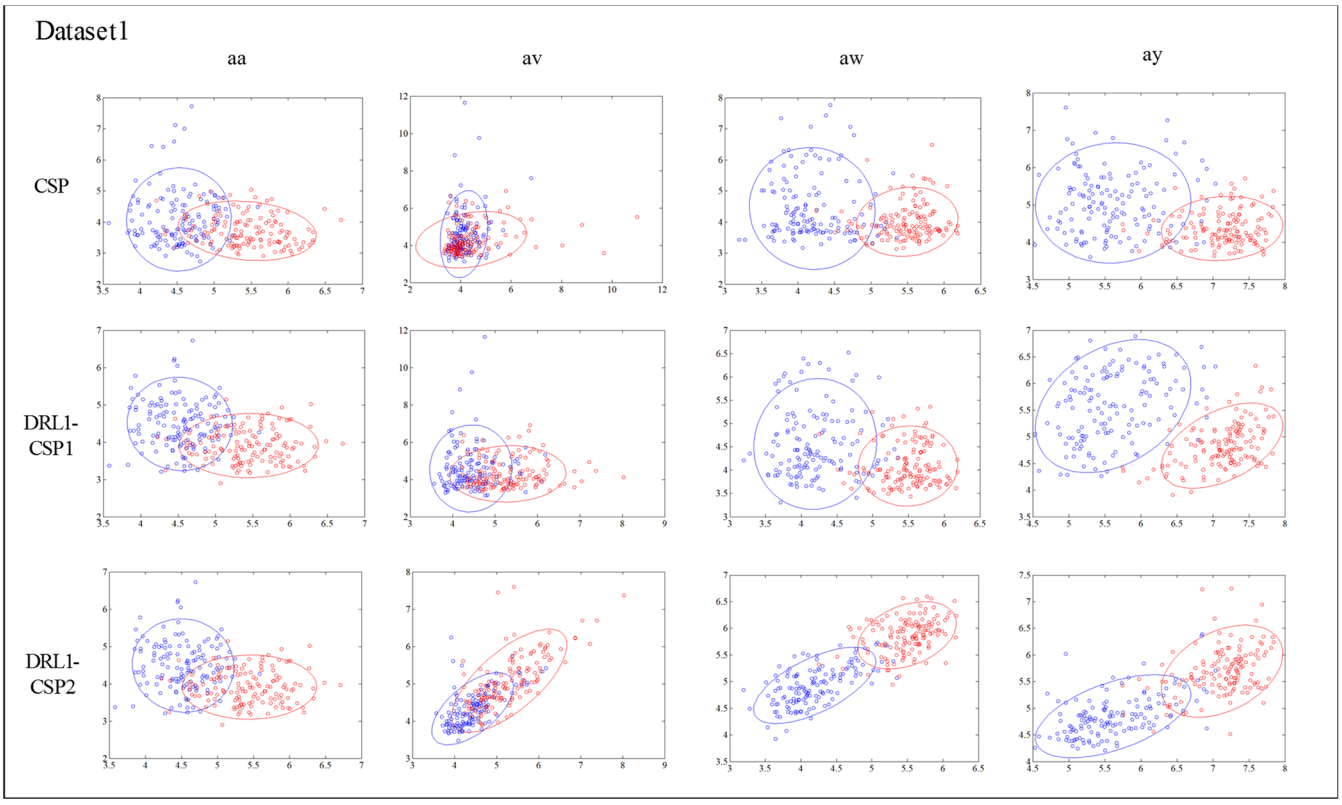

Fig. 5. Comparison of feature distributions from four participants of data set 1 ("aa," "av," "aw," and "ay"). The columns from left to right are feature distribution for participants "aa," "av," "aw," and "ay" with three different methods, respectively. Small circles of different colors represent the trials of different classes, while the large circles reflect the feature distribution range of most trials in the class of the corresponding color.

interclass distance than f_C. The difference in the features, corresponding to the vertical axis, means that the average accuracy of the two methods differs by 6.1% (85.0% with CSP versus 91.1% with DRL1-CSP1, paired t-test, $p = 0.074$, with nine degrees of freedom), which proves that the feature selected by DRL1-CSP1 is better.

D. Comparison of Fusion Algorithm With Different Feature Selection Methods

With the two data sets described in Section IV, the comparison is carried out in the proposed fusion method and the following algorithms.

MIN-CSP: MIN is used to select features in the feature space used within CSP.

Lasso-CSP: Lasso is used to select features in the feature space used within CSP.

FCSP: The objective function of the Fisher ratio is used to select features in the feature space used within CSP.

DRL1-CSP: Adopt DRL1-CSP2 that is presented in Section III.

1) Comparison of Classification Accuracy: The classification accuracy of each method is shown in Table III. The average accuracy of the fusion method is 6.8% and 15.4% higher than that of the traditional CSP for the two data sets and

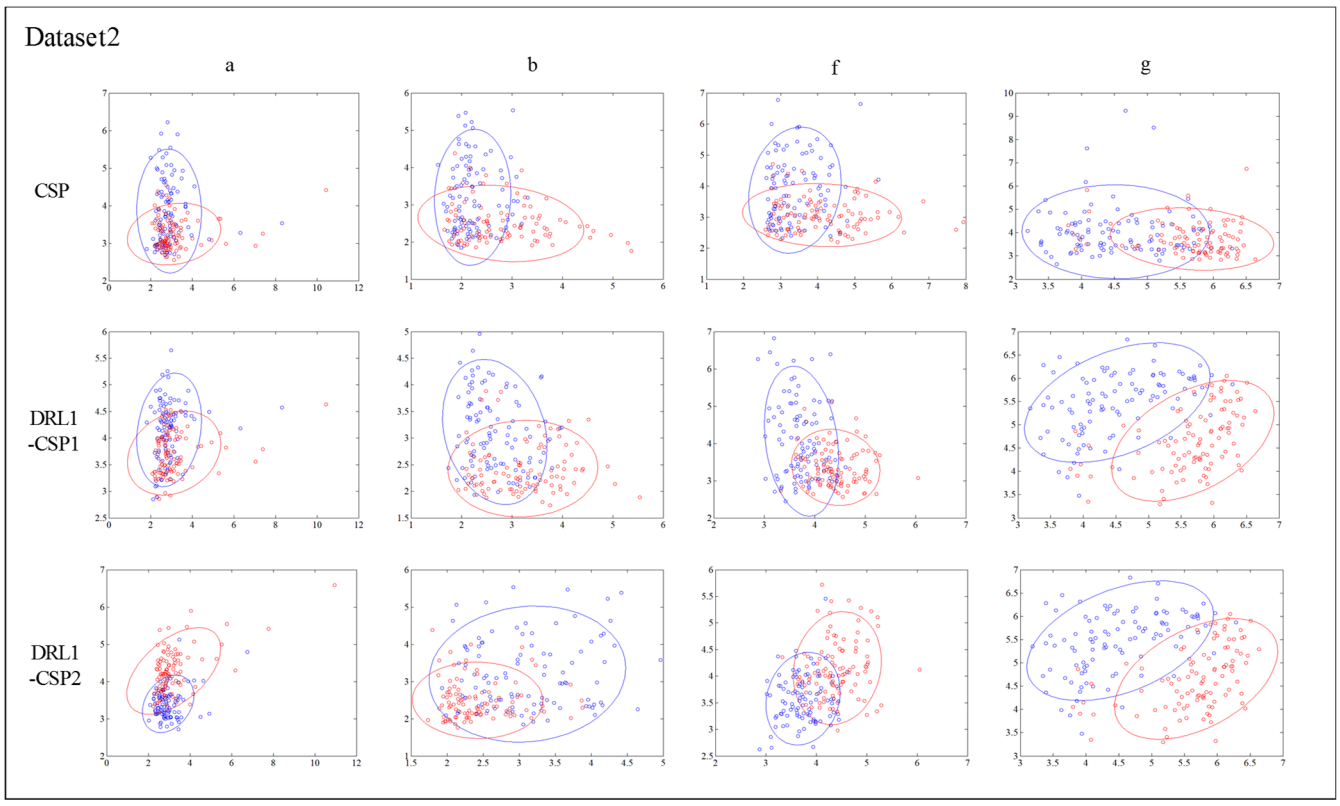

Fig. 6. Comparison of feature distributions from four participants of data set 2 ("a," "b," "f," and "g").

is the highest among the algorithms used in the comparison (paired t-test, $p < 0.05$, with eight degrees of freedom).

The advantages of the proposed fusion algorithm can also be seen in the table: in some cases, the performance of the fusion algorithm is equal to the higher of DRL1-CSP and FCSP, which guarantees the performance of the fusion algorithm. In other cases, the performance of the fusion algorithm is higher than both algorithms before the fusion, indicating that a better feature combination is found. Through the fusion algorithm, the interclass distance and feature distribution can be considered at the same time so that the pros and cons of the features can be judged more comprehensively.

2) Comparison of Computational Time: When evaluating our proposed extension of CSP for internal feature selection, the computational time is extremely important. A number of other algorithms treat CSP as a base algorithm and run it over multiple iterations (e.g., channel selection algorithms [38]). We first calculate the time complexity of the feature extraction process of different methods, in which the time complexity of CSP is $O(M)$, the time complexity of DRL1-CSP, FCSP, fusion algorithm, and MIN-CSP is $O(M + N)$, and the time complexity of lasso is $O(M * N)$. Here, M represents the number of trials, and N represents the number of channels. Table IV shows the specific computational time (in seconds) of algorithms used in the comparison. Each algorithm is run within a 10×10-fold cross-validation scheme on all participants' data, and then, the average running time is calculated over the data set. The calculations are made on a Windows computer with i5-6300HQ 2.3-GHz CPU/16-GB RAM. It may be clearly seen that the computational time of MIN-CSP is about two times longer than traditional CSP, and Lasso-CSP takes dozens of times longer, while the computational times of FCSP, DRL1-CSP, and their fusion algorithm are close to that of traditional CSP. Hence, the latter three methods are more suitable for internal feature selection in terms of efficiency.

V. Discussion

CSP is a commonly used algorithm for feature extraction in MI-BCI systems that achieved a good balance between the effectiveness and computational cost [39]. In recent years, lots of extensions have been proposed to improve the shortcomings of CSP, such as common spatiospectral pattern (CSSP) [40], SCSP [9], FBCSP [12], and RCSP [16], but few studies focus on the selection rules for the feature space used within CSP, which is closely related to the optimization of the CSP objective function. However, due to some drawbacks of the CSP objective function (mentioned in Section I), the selected features are not necessarily optimal in the feature space. In this work, efficient feature selection methods are designed to discover features that may have better classification results.

The traditional CSP algorithm could be divided into three phases: extraction, internal selection, and generation. As shown in Fig. 2, new feature selection methods are achieved by calculating improved objective functions during the internal selection phase. These feature selection methods

TABLE III

COMPARISON OF CLASSIFICATION ACCURACY (%) FOR DIFFERENT FEATURE SELECTION ALGORITHMS

Participant	Methods					
	CSP	Lasso-CSP	MIN-CSP	FCSP	DRL1-CSP	Fusion
aa	77.1	76.4	**80.0**	78.9	77.5	**80.0**
al	**96.8**	92.1	96.4	96.4	**96.8**	**96.8**
av	49.2	55.7	**70.0**	63.6	69.3	**70.0**
aw	85.0	90.0	91.4	90.7	**92.5**	**92.5**
ay	88.6	88.2	89.6	86.1	89.6	**91.1**
Mean±std	79.3±18.3	80.5±15.1	85.5±10.5	83.1±12.7	85.1±11.4	**86.1±10.9**
a	54.5	64.0	65.5	64.5	66.5	**72.0**
b	54.5	55.5	58.5	**61.0**	58.5	**61.0**
f	47.0	62.0	61.5	62.5	63.5	**65.0**
g	62.5	79.5	84.0	80.5	**85.0**	84.5
Mean±std	54.6±6.3	65.3±10.3	67.4±11.4	67.1±9.0	68.4±11.6	**70.0±11.1**
p-value	0.006	0.002	0.037	0.007	0.038	-

TABLE IV

COMPARISON OF COMPUTATIONAL TIME FOR DIFFERENT METHODS (IN SECONDS, MEAN ± STD)

Methods	CSP	Lasso-CSP	MIN-CSP	FCSP	DRL1-CSP	Fusion
Dataset1	40.46±1.31	953.81±212.31	80.97±0.57	45.82±0.79	54.15±2.12	54.18±1.38
Dataset2	17.61±0.51	541.22±111.41	31.30±0.31	21.13±0.15	25.56±0.48	25.84±0.40

can search for features that are more in line with the new objective function from the feature space used within CSP, which is especially suitable for the objective functions that are difficult to optimize through gradient descent. In the proposed objective functions, improvements are made in suppressing outliers and discovering features with larger interclass distances (DRL1-CSP). As shown in Table I, both the proposed methods significantly improve the classification accuracies ($p < 0.05$). Moreover, the objective function values in DRL1-CSP2 could be compared with each other, where only features with the largest objective function values need to be selected. We also compare the feature distributions from all trials in the two data sets (except participant "al"). As shown in Figs. 5 and 6, it is obvious that the features selected by the proposed methods have better feature distributions than those selected by the traditional CSP algorithm. In addition, as extensions of CSP that only change the rules of internal feature selection, the proposed methods could be easily embedded in other type extensions of CSP. To test this, we evaluate the performance of the proposed methods after embedding them in CSP-rank. As shown in Table II, the proposed methods have a significant improvement compared to unembedded CSP-rank ($p < 0.05$), which proves the effectiveness and great potential of proposed methods.

Both DRL1-CSP and traditional CSP do not consider the distribution of features in their objective functions, while FCSP only uses the Fisher ratio to discriminate features that, sometimes, yield worse results. We use DST to fuse two different feature selection rules (DRL1-CSP2 and FCSP) so that the distribution of features could be taken into consideration in DRL1-CSP and the performance of FCSP could be more stable. As shown in Table III, the average classification accuracy of the proposed fusion algorithm is superior to other algorithms used in the comparison ($p < 0.05$). The computational time is another important indicator of feature selection rules. As shown in Table IV, the computational time of the proposed methods is close to traditional CSP, while MIN-CSP and Lasso-CSP are several times greater than traditional CSP. This is because the proposed methods perform simple computation only for the projected signals. Although it has achieved good results, the assignment of the mass function in this work only adopts one combination scheme (see Fig. 3). In future work, we will try other assignment strategies to get better performance.

The proposed methods consume less computational cost in exchange for a more significant increase in the performance of MI-based BCI systems, but there are several limitations to these methods. One of the main limitations is that the proposed methods rely on the feature space used within CSP. In other words, if the candidate spatial filters from the feature space lack diversity, the performance of the MI-BCI system may not be significantly improved by proposed methods. We can expand the dimensions of the feature space to solve this problem, which is worthy of more in-depth study in the future. Another limitation is that the performance can only be evaluated offline when using competition data set. Although the cross-validation has been used, overfitting may still happen, which raises concerns about the generalizability of the results. In future work, we will evaluate the method online and test its generalization ability.

VI. CONCLUSION

In this work, several methods are proposed to solve the problems existing in the feature space used within CSP, wherein the selected features may not be optimal. By changing the objective function during the internal selection phase, new feature selection methods are realized. Especially, we first replace the average L2-norm squared ratio with the ratio of the average L1-norm so that the influence of outliers on the

feature distribution could be suppressed. Then, we upgrade it to the product between the difference and ratio of the average L1-norm (DRL1-CSP) to discover the features with larger interclass distances. DRL1-CSP could be divided into two methods (one-side selection or two-side selection). Moreover, we use DST to fuse DRL1-CSP with FCSP so that the distribution of features could be taken into consideration in DRL1-CSP. The experimental results show that the proposed methods effectively improve the performance of the BCI system with a small increase in computational time. In summary, the proposed methods can make full use of the feature space of CSP, which contributes to the development of feature extraction for MI-based BCI systems.

REFERENCES

[1] J. Wolpaw, N. Birbaumer, D. McFarland, G. Pfurtscheller, and T. Vaughan, "Brain-computer interfaces for communication and control," *Clin. Neurophys.*, vol. 113, no. 6, pp. 767–791, 2002.

[2] F. Lotte and C. Guan, "Regularizing common spatial patterns to improve BCI designs: Unified theory and new algorithms," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 2, pp. 355–362, Feb. 2011.

[3] A. Herweg, J. Gutzeit, S. Kleih, and A. Kübler, "Wheelchair control by elderly participants in a virtual environment with a brain-computer interface (BCI) and tactile stimulation," *Biol. Psychol.*, vol. 121, pp. 117–124, Dec. 2016.

[4] R. Scherer, G. R. Müller, C. Neuper, B. Graimann, and G. Pfurtscheller, "An asynchronously controlled EEG-based virtual keyboard: Improvement of the spelling rate," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 979–984, Jun. 2004.

[5] R. Leeb, D. Friedman, G. R. Müller-Putz, R. Scherer, M. Slater, and G. Pfurtscheller, "Self-paced (asynchronous) BCI control of a wheelchair in virtual environments: A case study with a tetraplegic," *Comput. Intell. Neurosci.*, vol. 2007, p. 79642, Jan. 2007.

[6] J. Meng, G. Huang, D. Zhang, and X. Zhu, "Optimizing spatial spectral patterns jointly with channel configuration for brain–computer interface," *Neurocomputing*, vol. 104, pp. 115–126, Mar. 2013.

[7] H. Ramoser, J. Muller-Gerking, and G. Pfurtscheller, "Optimal spatial filtering of single trial EEG during imagined hand movement," *IEEE Trans. Rehabil. Eng.*, vol. 8, no. 4, pp. 441–446, Dec. 2000.

[8] Y. Jiao *et al.*, "Sparse group representation model for motor imagery EEG classification," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 2, pp. 631–641, Mar. 2019.

[9] M. Arvaneh, C. Guan, K. K. Ang, and C. Quek, "Optimizing the channel selection and classification accuracy in EEG-based BCI," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 6, pp. 1865–1873, Jun. 2011.

[10] W. Samek, M. Kawanabe, and K.-R. Müller, "Divergence-based framework for common spatial patterns algorithms," *IEEE Rev. Biomed. Eng.*, vol. 7, pp. 50–72, Nov. 2014.

[11] Y. Zhang, Y. Wang, J. Jin, and X. Wang, "Sparse Bayesian learning for obtaining sparsity of EEG frequency bands based feature vectors in motor imagery classification," *Int. J. Neural Syst.*, vol. 27, no. 2, Mar. 2017, Art. no. 1650032.

[12] K. K. Ang, Z. Y. Chin, C. Wang, C. Guan, and H. Zhang, "Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b," *Frontiers Neurosci.*, vol. 6, p. 39, Mar. 2012.

[13] J. Feng *et al.*, "Towards correlation-based time window selection method for motor imagery BCIs," *Neural Netw.*, vol. 102, pp. 87–95, Jun. 2018.

[14] J. Jin, Y. Miao, I. Daly, C. Zuo, D. Hu, and A. Cichocki, "Correlation-based channel selection and regularized feature optimization for MI-based BCI," *Neural Netw.*, vol. 118, pp. 262–270, Oct. 2019.

[15] W. Samek, C. Vidaurre, K. R. Müller, and M. Kawanabe, "Stationary common spatial patterns for brain-computer interfacing," *J. Neural Eng.*, vol. 9, no. 2, p. 026013, Apr. 2012.

[16] H. Lu, H.-L. Eng, C. Guan, K. N. Plataniotis, and A. N. Venetsanopoulos, "Regularized common spatial pattern with aggregation for EEG classification in small-sample setting," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 12, pp. 2936–2946, Dec. 2010.

[17] H. Kang, Y. Nam, and S. Choi, "Composite common spatial pattern for subject-to-subject transfer," *IEEE Signal Process. Lett.*, vol. 16, no. 8, pp. 683–686, Aug. 2009.

[18] H. Wang, Q. Tang, and W. Zheng, "L1-norm-based common spatial patterns," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 3, pp. 653–662, Mar. 2012.

[19] X. Li, X. Lu, and H. Wang, "Robust common spatial patterns with sparsity," *Biomed. Signal Process. Control*, vol. 26, pp. 52–57, Apr. 2016.

[20] M. Arvaneh, C. Guan, K. Keng Ang, and C. Quek, "Optimizing spatial filters by minimizing within-class dissimilarities in electroencephalogram-based brain–computer interface," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 4, pp. 610–619, Apr. 2013.

[21] D. Fattahi, B. Nasihatkon, and R. Boostani, "A general framework to estimate spatial and spatio-spectral filters for EEG signal classification," *Neurocomputing*, vol. 119, pp. 165–174, Nov. 2013.

[22] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, Oct. 2007.

[23] S. Kumar, K. Mamun, and A. Sharma, "CSP-TSM: Optimizing the performance of Riemannian tangent space mapping using common spatial pattern for MI-BCI," *Comput. Biol. Med.*, vol. 91, pp. 231–242, Dec. 2017.

[24] Y. Zhang, G. Zhou, J. Jin, X. Wang, and A. Cichocki, "Optimizing spatial patterns with sparse filter bands for motor-imagery based brain–computer interface," *J. Neurosci. Methods*, vol. 255, pp. 85–91, Nov. 2015.

[25] X. Li, C. Guan, H. Zhang, and K. K. Ang, "A unified Fisher's ratio learning method for spatial filter optimization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 11, pp. 2727–2737, Nov. 2017.

[26] A. P. Dempster, "Upper and lower probabilities induced by a multivalued mapping," *Ann. Math. Statist.*, vol. 38, no. 2, pp. 325–339, Apr. 1967.

[27] G. Shafer, *A Mathematical Theory of Evidence*. Princeton, NJ, USA: Princeton Univ., 1976.

[28] B. R. Mathon, M. M. Ozbek, and G. F. Pinder, "Dempster–Shafer theory applied to uncertainty surrounding permeability," *Math. Geosci.*, vol. 42, no. 3, pp. 293–307, Apr. 2010.

[29] T. N. Lal *et al.*, "Support vector channel selection in BCI," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 1003–1010, Jun. 2004.

[30] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[31] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Process. Lett.*, vol. 9, no. 3, pp. 293–300, Jun. 1999.

[32] S. Razi, M. R. Karami Mollaei, and J. Ghasemi, "A novel method for classification of BCI multi-class motor imagery task based on Dempster–Shafer theory," *Inf. Sci.*, vol. 484, pp. 14–26, May 2019.

[33] A. Yazdani, T. Ebrahimi, and U. Hoffmann, "Classification of EEG signals using Dempster Shafer theory and a k-nearest neighbor classifier," in *Proc. 4th Int. IEEE/EMBS Conf. Neural Eng.*, Antalya, Turkey, Apr. 2009, pp. 327–330.

[34] G. Dornhege, B. Blankertz, G. Curio, and K.-R. Müller, "Boosting bit rates in noninvasive EEG single-trial classifications by feature combination and multiclass paradigms," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 993–1002, Jun. 2004.

[35] B. Blankertz, G. Dornhege, M. Krauledat, K.-R. Müller, and G. Curio, "The non-invasive Berlin brain–computer interface: Fast acquisition of effective performance in untrained subjects," *NeuroImage*, vol. 37, no. 2, pp. 539–550, Aug. 2007.

[36] W.-K. Tam, Z. Ke, and K.-Y. Tong, "Performance of common spatial pattern under a smaller set of EEG electrodes in brain-computer interface on chronic stroke patients: A multi-session dataset study," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Boston, MA, USA, Aug. 2011, pp. 6344–6347.

[37] T. Alotaiby, F. E. A. El-Samie, S. A. Alshebeili, and I. Ahmad, "A review of channel selection algorithms for EEG signal processing," *EURASIP J. Adv. Signal Process.*, vol. 2015, no. 1, p. 66, Dec. 2015.

[38] Z. Qiu, J. Jin, H.-K. Lam, Y. Zhang, X. Wang, and A. Cichocki, "Improved SFFS method for channel selection in motor imagery based BCI," *Neurocomputing*, vol. 207, pp. 519–527, Sep. 2016.

[39] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Müller, "Optimizing spatial filters for robust EEG single-trial analysis," *IEEE Signal Process. Mag.*, vol. 25, no. 1, pp. 41–56, Jan. 2008.

[40] S. Lemm, B. Blankertz, G. Curio, and K.-R. Müller, "Spatio-spectral filters for improving the classification of single trial EEG," *IEEE Trans. Biomed. Eng.*, vol. 52, no. 9, pp. 1541–1548, Sep. 2005.

Jing Jin (Senior Member, IEEE) received the Ph.D. degree in control theory and control engineering from the East China University of Science and Technology (ECUST), Shanghai, China, in 2010.

His Ph.D. advisors were Prof. Gert Pfurtscheller at the Graz University of Technology, Graz, Austria, from 2008 to 2010, and Prof. Xingyu Wang at ECUST from 2006 to 2008. He is currently a Professor and the Chair of the Automation Department, ECUST. His research interests include brain–computer interface, signal processing, and pattern recognition.

Dr. Jin serves as an Associate Editor of *Frontiers in Neurorobotics*, an Action Editor of *Neural Networks*, and an Editor of the *Journal of Neural Engineering* and *Journal of Neuroscience Methods*.

Ruocheng Xiao received the B.S. degree in electrical engineering and automation from the East China University of Science and Technology, Shanghai, China, in 2018, where he is currently pursuing the master's degree.

His research interests include brain–computer interface, machine learning, and signal processing.

Ian Daly received the M.Eng. degree in computer science and the Ph.D. degree in cybernetics from the University of Reading, Reading, U.K., in 2006 and 2011, respectively.

From May 2011 to May 2013, he was a Post-Doctoral Researcher with the Laboratory of Brain-Computer Interfaces, Graz University of Technology, Graz, Austria. He is currently a Lecturer with the University of Essex, U.K. His research interests focus on BCIs, nonlinear dynamics, machine learning, signal processing, and connectivity analysis in EEG and Functional Magnetic Resonance Imaging (fMRI). He is also interested in the neurophysiological correlates of motor control and stimuli perception and how they differ between healthy participants and individuals with neurological and physiological impairments.

Yangyang Miao received the B.S. degree in electrical engineering and automation and the M.S. degree in control science and engineering from Nantong University, Nantong, China, in 2013 and 2016, respectively. He is currently pursuing the Ph.D. degree in control science and engineering with the East China University of Science and Technology, Shanghai, China.

His research interests include brain–computer interface, signal processing, and machine learning.

Xingyu Wang was born in Sichuan, China, in 1944. He received the B.S. degree in mathematics from Fudan University, Shanghai, China, in 1967, the M.S. degree in control theory from East China Normal University, Shanghai, in 1982, and the Ph.D. degrees in industrial automation from the East China University of Science and Technology, in 1984.

He is currently a Professor with the School of Information Science and Engineering, East China University of Science and Technology. His research interests include control theory, control techniques, the application to biomedical systems, and brain control.

Andrzej Cichocki (Fellow, IEEE) received the M.Sc. (Hons.), Ph.D., and Dr.Sc. (Habilitation) degrees from the Warsaw University of Technology, Warsaw, Poland, in 1972, 1975, and 1982, respectively, all in electrical engineering.

He spent several years at University Erlangen, Erlangen, Germany, as an Alexander-von-Humboldt Research Fellow and a Guest Professor. From 1995 to 2017, he was a Senior Team Leader and the Head of the Laboratory for Advanced Brain Signal Processing, RIKEN Brain Science Institute, Wako, Japan. He is currently a Professor with the Skolkovo Institute of Science and Technology (SKOLTECH), Moscow, Russia. He is the author of more than 500 technical journal papers and five monographs in English (two of them translated to Chinese). His joint publications currently report over 44,000 citations according to Google Scholar, with an h-index of 97. His current research focuses on multiway blind source separation, tensor decomposition, tensor networks for big data mining, and brain–computer interface.

Dr. Cichocki has served as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON CYBERNETICS, and the *Journal of Neuroscience Methods*. He is the founding Editor-in-Chief of the *Journal Computational Intelligence and Neuroscience*.