A Novel Feature Extraction Method for Multi-Class Kernel CSSD

Yuyu GAO, Chunfang WANG*, Ying YANG

LiRen College of Yanshan University, Hebei Province, Qinhuangdao Hebei, 066004, China

Abstract — **In order to relax the presumption of strictly linear patterns between source signals and recorded EEG in Common Spatial Subspace Decomposition (CSSD), in this paper we study Kernel CSSD and propose to extend it to multi-class. To obtain the solution of the spatial filter, we adopt cross validation on the training data to get the best number m of the eigenvector, and the filter is optimized through KL divergence to determine the optimal spatial filter. We selects III_3a data set in 2005 BCI competition in the experiment, where a linear kernel function and a nearest neighbor classifier are used. The experiment shows that our method in this paper can extract the corresponding effective feature from multi-class EEG data of single trial, and get better classification results.**

Keywords - Common spatial subspace decomposition (CSSD); cross validation; KL divergence; multi-class; spatial filter

I. INTRODUCTION

Brain-Computer Interface (BCI) is a new control path of external communication established between brain and external environment using computer or other electronic equipment which does not depend on peripheral nerve and muscle tissue, and provides a new way to communicate with the outside world for those patients lost parts or all of muscle control function [1]. BCI is a non-muscular communication system and can make person's brain intention and external environment communicate directly, so its system function is that a new communication channel between the computer and the brain is established [2]. A part of the BCI languages is from the brain signals (external devices are controlled through the feature extraction of brain signals); the other part is from consultation (the user and the system adapt mutually through continuous adjustment) [3]. Like any communication system, BCI has input (signals from a user's brain), output (i.e. equipment directive), the component of classifying the former to the latter and operating protocol of begin, offset and timing. Therefore, any BCI system can be composed of four parts: signal acquisition, signal processing, output devices and operating protocol [4].

In recent years, the research of BCI system has become a rising trend. BCI technology has important application values in rehabilitation engineering, military and other fields, and has attract attention of more and more scientists and researchers. The core part of the BCI system is the analysis process of EEG signal-feature extraction. Quickly and effectively extract the features related to the task is the biggest issue placed in front of researchers currently.

II. RELATED WORKS

In order to extract the signal features fast and effectively, a feature extraction method which is called Common Spatial Patterns (CSP) [5] is proposed. CSP can create an optimal spatial filter, which is used to discriminate the importance between different signal classification and get the eigenvector has the greatest ability to distinguish between-class. It can handle multi-channel Electroencephalogram (EEG) simultaneously, which is conducive to learning concealed features which have overlapping noise with the frequency range in EEG or do not have the ability to distinguish. It can make these recorded signals become more clear, so as to obtain more accurate and credible brain signals or remove the effect of reference electrode from the signals. Traditional CSP is only suitable for multi-channel two class EEG signal, because it only considers the projection of two classes tasks in space has maximum separability, but the effects are also affected by factors such as non stationarity of EEG and frequency filtering. In order to overcome the disadvantages of CSP, many researchers have improved CSP method. For example, Lemm S et al. [6] propose a Common Spatio-Spectral Patterns (CSSP). In order to optimize spatial pattern and frequency pattern, the algorithm adds a delay function in each lead data to filter and improves classification performance, but the flexibility of frequency selection is low and the choice of delay parameters is very difficult. Regularized Common Spatial Pattern with Aggregation (R-CSP-A) is proposed by Lu et al. [7], which estimate two parameters using regulated covariance matrix to reduce the estimation variance and enhance the classification performance of the algorithm and solve the problem of selecting regularization parameter using the aggregation of a large number of regularized common spatial. Nguyen et al. [8] introduce real-time update feature mode to adapt to the non-stationarity of EEG, and designed incremental CSP. Then in order to solve the problem of pollution source of non-stationary signal and separate EEG source, stationary CSP (sCSP) is proposed [9]. Common Spatial Spectrum Sparsity Pattern (CSSSP) is proposed in literatures [10] and [11], but the choice of parameters is very difficult. If selected parameters are too small, overfitting is caused easily; if the selected parameters are too large, it is infinitely close to CSP algorithm and has poor generalization ability. Wu et al. [12] extend CSP algorithm to multi-class case and introduces One-Versus-the-Rest (OVR) CSP strategy, namely computing spatial filter with one class to all other class.

DOI 10.5013/ IJSSST.a.17.36.59

ISSN: 1473-804x online, 1473-8031 print

CSSD is a very similar to CSP, where a pseudo whitening matrix is used for solving spatial filter [13]. In order to relax the presumption of strictly linear patterns between source signals and recorded EEG in CSSD, we studied Kernel CSSD and extended to multi-class. Finally, using KL divergence we minimize within-class dissimilarities and obtain good classification performance.

III. METHOD DESCRIPTION

Common Average Reference (CAR) is a spatial filter of EEG data, whose purpose is to remove the correlation between channels [14], which is subtract the mean of original signal of all leads from each lead signal. Suppose there are N channels, namely the number of leads is N. The formula is:

$$x'_s(t) = x_s(t) - (1/n)\sum_{i=1}^{N} x_i(t) \tag{1}$$

A. *Common Spatial Subspace Decomposition*

Suppose $\mathbf{X}_{(1)}$ and $\mathbf{X}_{(2)}$ are recorded EEG signal of motor 1and motor 2 respectively (such as the left and right hand motor) and each row of the matrix corresponds to a EEG channel [15]. In CSSD, the above signal can be decomposed into:

$$\begin{cases} \mathbf{X}_{(1)} = [\mathbf{C}_1 \, \mathbf{C}_{cm}]\begin{bmatrix}\mathbf{S}_1 \\ \mathbf{S}_{cm}\end{bmatrix} \\ \mathbf{X}_{(2)} = [\mathbf{C}_2 \, \mathbf{C}_{cm}]\begin{bmatrix}\mathbf{S}_2 \\ \mathbf{S}_{cm}\end{bmatrix} \end{cases} \tag{2}$$

Where $\mathbf{S}_{cm}$, $\mathbf{S}_1$ and $\mathbf{S}_2$ represents the source component of common, motor 1 and motor 2 respectively; $\mathbf{C}_{cm}$, $\mathbf{C}_1$ and $\mathbf{C}_2$ represents corresponding spatial pattern of the source component of common, motor 1 and motor 2 respectively.

In order to extract respective source component of motor 1 and motor 2 from recorded EEG signal, CSSD try to find the spatial filters $\mathbf{F}_1$ and $\mathbf{F}_2$.

$$\mathbf{S}_c = \mathbf{F}_c\mathbf{X}_{(c)}, c = 1,2 \tag{3}$$

In addition to different calculation method of spatial filter, CSSD is similar to CSP, and only solving spatial filter uses a pseudo whitening matrix in CSSD .

Assuming that $\mathbf{u}$ is a $N \times 1$ column vector (N is the channel number). A typical EEG signal $\mathbf{X}$ can be generated a single channel signal by a spatial filter $\mathbf{u}^T$, whose energy can be expressed as $\mathbf{u}^T(\mathbf{X}\mathbf{X}^T)\mathbf{u}$. The average energy about signal of motor 1 and motor 2 is $\mathbf{u}^T(\mathbf{R}_1)\mathbf{u}$ and $\mathbf{u}^T(\mathbf{R}_2)\mathbf{u}$ respectively, so the average covariance matrix belonging to each class is:

$$\mathbf{R}_c = \frac{1}{n_c}\sum_i \mathbf{X}_{i,(c)}\mathbf{X}_{i,(c)}^{\ T}, \ c = 1,2 \tag{4}$$

Where $\mathbf{X}_{i,(c)}$ represents i-th trial data of C class, and is a $N \times T$ matrix; N is the number of channels, and T is the number of sampled points for each channel. $\mathbf{X}_{i,(c)}^{\ T}$ represents transport matrix of $\mathbf{X}_{i,(c)}$ and n_c represents the trial number of C class.

The standardized average covariance matrix is expressed as:

$$\mathbf{R}_c = \frac{1}{n_c}\sum_i \frac{\mathbf{X}_{i,(c)}\mathbf{X}_{i,(c)}^{\ T}}{trace(\mathbf{X}_{i,(c)}\mathbf{X}_{i,(c)}^{\ T})}, c = 1,2 \tag{5}$$

Where $trace(\cdot)$ represents matrix trace. The standardized data can guarantee that the sum of energy in all signals channel is equal to one. So it make $\mathbf{R}_c$ is robust to possible changes in the aspect of signal amplification. However, when the number of EEG channels is relatively less, it may cause loss of data.

CSSD make the variance of a class (energy) maximum, through finding the optimal projection direction is the optimal linear spatial filter, while ensuring the variance of another class (energy) minimum, which is the same with CSP.

Common covariance matrix is:

$$\mathbf{R} = \mathbf{R}_1 + \mathbf{R}_2 \tag{6}$$

With the eigenvalue decomposition of common covariance matrix $\mathbf{R}$, denote $\mathbf{V}$ for the matrix composed by eigenvector of $\mathbf{R}$. $\mathbf{\Lambda}$ is diagonal matrix composed by corresponding eigenvalues, it can get:

$$\mathbf{R} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T \tag{7}$$

Pseudo whitening matrix $\mathbf{P}_1$ in CSSD algorithm is defined as:

$$\mathbf{P}_1 = \mathbf{V}\mathbf{\Lambda}^{1/2} \tag{8}$$

Then the whitening matrix is:

$$\mathbf{P} = \mathbf{\Lambda}^{-1/2}\mathbf{V}^T \tag{9}$$

That is $\mathbf{P}(\mathbf{R}_1 + \mathbf{R}_2)\mathbf{P}^T = \mathbf{I}$. Define $\mathbf{Y}_1$ and $\mathbf{Y}_2$ respectively as:

$$\mathbf{Y}_1 = \mathbf{P}\mathbf{R}_1\mathbf{P}^T \tag{10}$$

$$\mathbf{Y}_2 = \mathbf{P}\mathbf{R}_2\mathbf{P}^T = \mathbf{I} - \mathbf{Y}_1 \tag{11}$$

Suppose the eigenvector of $\mathbf{R}$ and the diagonal matrix composed by corresponding eigenvalue is respectively:

$$\mathbf{V}_1 = [\mathbf{v}_1 \, \mathbf{v}_2 \ldots \mathbf{v}_N] \tag{12}$$

$$\mathbf{\Lambda}_1 = diag([\lambda_1\lambda_2\ldots\lambda_N]), \ \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N \tag{13}$$

$$\mathbf{Y}_1\mathbf{V}_1 = \mathbf{V}_1\mathbf{\Lambda}_1 \Rightarrow (\mathbf{I} - \mathbf{Y}_2)\mathbf{V}_1 = \mathbf{V}_1\mathbf{\Lambda}_1 \Rightarrow \mathbf{Y}_2\mathbf{V}_1 = \mathbf{V}_1(\mathbf{I} - \mathbf{\Lambda}_1) \tag{14}$$

Where $\mathbf{V}_1$ also is the eigenvector of $\mathbf{V}_2$, and the diagonal matrix composed by corresponding eigenvalue is $\mathbf{I} - \mathbf{\Lambda}_1$ .The sum of the eigenvalues of $\mathbf{Y}_1$ and $\mathbf{Y}_2$ is one, which means that the same eigenvector in $\mathbf{V}_1$ make a class have a larger eigenvalue, while it can make another class have a smaller eigenvalue.

Then select m biggest eigenvalue and the corresponding eigenvector respectively for the two classes, which two new eigenvectors ($\mathbf{V}_{11}$ and $\mathbf{V}_{12}$) are composed of. $\mathbf{V}_{11}$ and $\mathbf{V}_{12}$ denote different optimal direction of two classes respectively, whose size are $N \times m$.

Define optimal spatial filters as:

$$\mathbf{F}_1 = \mathbf{P}_1\mathbf{V}_{11}\mathbf{V}_{11}{}^T\mathbf{P} \quad (15)$$

$$\mathbf{F}_2 = \mathbf{P}_1\mathbf{V}_{12}\mathbf{V}_{12}{}^T\mathbf{P} \quad (16)$$

A new EEG data is mapped to a new space s follows.

$$\mathbf{S}_c = \mathbf{F}_c\mathbf{X}_k,\ c = 1,2 \quad (17)$$

Where the size of $\mathbf{S}_c$ is $N \times T$.

Finally, in order to obtain the feature, the variance for the row of generated new signal $\mathbf{S}_c$ is taken, and get:

$$f_k = \log\left(\frac{\operatorname{var}(\mathbf{S}_1)}{\sum_{i=1}^{2}\operatorname{var}(\mathbf{S}_i)}\right) \quad (18)$$

Where $\operatorname{var}(\cdot)$ represents vector variance; f_k corresponds the feature of the EEG data $\mathbf{X}_k$, whose size is $1\times N$. So the feature of n trial is $f = [f_1; f_2; \dots f_n]$, whose size is $n \times N$ .Multi-class kernel CSSD.

B. *Multi-Class Kernel CSSD*

Because original CSSD method requires that it is a strict linear relationship between original EEG and recorded EEG signal, in order to alleviate the hypothetical relationship, a multi-class kernel CSSD (MKCSSD) method is studied in the paper. Using kernel method the OVR method is extended to nonlinear form, which can ease the strict hypothesis of linear pattern. a signal component corresponding with the class is extract from EEG data of multiple classification using the nuclear space pattern through a class to all other class.

In the original multi-class EEG data, average covariance matrix of each class is:

$$\mathbf{R}_c = \frac{1}{n_c}\sum_i \mathbf{X}_{i,(c)}\mathbf{X}_{i,(c)}{}^T,\ c = 1,2,\dots,K \quad (19)$$

The standardized average covariance matrix is expressed as:

$$\mathbf{R}_c = \frac{1}{n_c}\sum_i \frac{\mathbf{X}_{i,(c)}\mathbf{X}_{i,(c)}{}^T}{trace(\mathbf{X}_{i,(c)}\mathbf{X}_{i,(c)}{}^T)}, c = 1,2,\dots,K \quad (20)$$

Make $\mathbf{R}_1{}' = \mathbf{R}_2 + \mathbf{R}_3 + \dots + \mathbf{R}_K$, so common covariance matrix is:

$$\mathbf{R} = \mathbf{R}_1 + \mathbf{R}_1{}' \quad (21)$$

In MKCSSD method, the original EEG data are mapped to a high dimensional space. All trial signals $\mathbf{X}_{i,(c)}$ are horizontally connected to form a new data $\mathbf{X}_t$, whose size is $N\times(T\times n)$, where n represents the total number of trials. Suppose j-th column vector $\mathbf{X}_t$ is $\mathbf{x}_j$ .which is mapped into a high dimensional space for $\phi(\mathbf{x}_j)$ through a kernel function. So high dimensional form of $\mathbf{X}_t$ is $\mathbf{\Phi}_t$. High dimension space form corresponding to one class and other remaining all classes is $\mathbf{\Phi}_1$ and $\mathbf{\Phi}_1{}'$ respectively. Now in the new space, we can obtain:

$$\mathbf{R}_1 = \sum_{\substack{j \\ class(\mathbf{x}_j)=1}} w_j\phi(\mathbf{x}_j)\phi(\mathbf{x}_j)^T = \mathbf{\Phi}_1\mathbf{W}_1\mathbf{\Phi}_1{}^T \quad (22)$$

$$\mathbf{R} = \mathbf{R}_1 + \mathbf{R}_1{}' = \sum_j w_j\phi(\mathbf{x}_j)\phi(\mathbf{x}_j)^T = \mathbf{\Phi}_t\mathbf{W}\mathbf{\Phi}_t{}^T \quad (23)$$

Where $\mathbf{W}$ represents a diagonal matrix composed by all w_j and $w_j = 1/n_{y_j}$. n_{y_j} represents the total number of trials of the class that x_j belongs to.

In order to complete multi-class CSSD method in high dimensional space, first of all, we must solve the problem of the following eigenvalue decomposition:

$$\mathbf{R}\mathbf{V}' = \mathbf{\Phi}_t\mathbf{W}\mathbf{\Phi}_t{}^T\mathbf{V}' = \mathbf{V}'\mathbf{\Lambda}' \quad (24)$$

Where $\mathbf{V}'$ is a matrix composed by the eigenvectors of $\mathbf{R}$; $\mathbf{\Lambda}'$ is a diagonal matrix composed by the corresponding eigenvalue.

We can use the following formula for the eigenvalue decomposition:

$$(\mathbf{\Phi}_t{}^T\mathbf{\Phi}_t\mathbf{W})\mathbf{V} = (\mathbf{K}\mathbf{W})\mathbf{V} = \mathbf{V}\mathbf{\Lambda} \quad (25)$$

Where $\mathbf{K}$ is kernel matrix and $\mathbf{K} = \mathbf{\Phi}_t{}^T\mathbf{\Phi}_t$.

At the same time both sides of eqn (25) are multiplied by $\mathbf{\Phi}_t\mathbf{W}$:

$$\mathbf{\Phi}_t\mathbf{W}\mathbf{\Phi}_t{}^T(\mathbf{\Phi}_t\mathbf{W}\mathbf{V}) = (\mathbf{\Phi}_t\mathbf{W}\mathbf{V})\mathbf{\Lambda} \quad (26)$$

According to eqn (24) and eqn (26), we can get that $\mathbf{V}'$ and $\mathbf{\Lambda}'$ are respectively corresponding to $\mathbf{\Phi}_t\mathbf{W}\mathbf{V}$ and $\mathbf{\Lambda}$. Standardized $\mathbf{\Phi}_t\mathbf{W}\mathbf{V}$ is $\mathbf{\Phi}_t\mathbf{W}\mathbf{V}\mathbf{\Psi}^{-1/2}\mathbf{\Lambda}^{-1/2}$, where $\mathbf{\Psi}$ is a diagonal matrix, whose j-th element is $\mathbf{v}_j{}^T\mathbf{W}\mathbf{v}_j$. So we can get a pseudo whitening matrix:

$$\mathbf{P}_1 = \mathbf{\Phi}_t\mathbf{W}\mathbf{V}\mathbf{\Psi}^{-1/2} \quad (27)$$

Corresponding whitening matrix is:

$$\mathbf{P} = \mathbf{\Lambda}^{-1}\mathbf{\Psi}^{-1/2}\mathbf{V}^T\mathbf{W}\mathbf{\Phi}_t \quad (28)$$

Then

$$\mathbf{Y}_1 = \mathbf{P}\mathbf{R}_1\mathbf{P}^T$$

$$= \mathbf{\Lambda}^{-1}\mathbf{\Psi}^{-1/2}\mathbf{V}^T\mathbf{W}\left(\mathbf{\Phi}_t{}^T\mathbf{\Phi}_1\right)\mathbf{W}_1\left(\mathbf{\Phi}_t{}^T\mathbf{\Phi}_1\right)^T\mathbf{W}\mathbf{V}\mathbf{\Psi}^{-1/2}\mathbf{\Lambda}^{-1}$$

(29)

At this time, the eigenvectors and eigenvalues of $\mathbf{Y}_1$ can be calculated explicitly:

$$\mathbf{Y}_1 = \mathbf{V}_1\mathbf{\Lambda}_1\mathbf{V}_1{}^T \quad (30)$$

Where $\mathbf{V}_1$ is a matrix composed by the eigenvectors of $\mathbf{Y}_1$; $\mathbf{\Lambda}_1$ is a diagonal matrix composed by the corresponding eigenvalue. So

$$\mathbf{I} = \mathbf{P}\mathbf{R}\mathbf{P}^T = \mathbf{P}\mathbf{R}_1\mathbf{P}^T + \mathbf{P}\mathbf{R}_1{}'\mathbf{P}^T = \mathbf{Y}_1 + \mathbf{Y}_1{}' \quad (31)$$

$$\mathbf{Y}_1{}' = \mathbf{I} - \mathbf{Y}_1 = \mathbf{I} - \mathbf{V}_1\mathbf{\Lambda}_1\mathbf{V}_1{}^T = \mathbf{V}_1(\mathbf{I} - \mathbf{\Lambda}_1)\mathbf{V}_1{}^T \quad (32)$$

Above two formulas show that $\mathbf{V}_1$ is also the eigenvector of $\mathbf{Y}_1$ and $\mathbf{I} - \mathbf{\Lambda}_1$ is the diagonal matrix composed by the corresponding eigenvalue. Because the sum of corresponding eigenvalues of $\mathbf{Y}_1$ and $\mathbf{Y}_1{}'$ is one, which means that the projection in the feature space can make the variance of one class higher and ensure the variance of all other class smaller.

So we can get spatial filter $\mathbf{F}_1$:

$$\begin{aligned}\mathbf{F}_1 &= \mathbf{P}_1\mathbf{V}_{1m}\mathbf{V}_{1m}{}^T\mathbf{P} \\ &= \mathbf{\Phi}_t\mathbf{W}\mathbf{V}\mathbf{\Psi}^{-1/2}\mathbf{V}_{1m}\mathbf{V}_{1m}{}^T\mathbf{\Lambda}^{-1}\mathbf{\Psi}^{-1/2}\mathbf{V}^T\mathbf{W}\mathbf{\Phi}_t{}^T\end{aligned} \tag{33}$$

Where $\mathbf{V}_{1m}$ represents the eigenvector corresponding to the m largest eigenvalue of $\mathbf{Y}_1$. m can be obtained by cross validation.

A new EEG data $\mathbf{X}_k$ is mapped to a new space s follows:

$$\begin{aligned}\mathbf{F}_1\mathbf{\Phi}_k &= \mathbf{\Phi}_t\mathbf{W}\mathbf{V}\mathbf{\Psi}^{-1/2}\mathbf{V}_{1m}\mathbf{V}_{1m}{}^T\mathbf{\Lambda}^{-1}\mathbf{\Psi}^{-1/2}\mathbf{V}^T\mathbf{W}\mathbf{K}_{tk}, \mathbf{K}_{tk} \\ &= \mathbf{\Phi}_t^T\mathbf{\Phi}_k\end{aligned} \tag{34}$$

The data mapped to the new space is an implicit function, which cannot be calculated, therefore we must construct kernel function form:

$$\begin{aligned}\mathbf{S}_1 &= \left(\mathbf{F}_1\mathbf{\Phi}_k\right)^T\mathbf{F}_1\mathbf{\Phi}_k \\ &= \mathbf{K}_{kt}\mathbf{W}\mathbf{V}\mathbf{\Psi}^{-1/2}\mathbf{\Lambda}^{-1}\mathbf{V}_{1m}\mathbf{V}_{1m}{}^T\mathbf{\Psi}^{-1/2}\mathbf{V}^T\mathbf{W}\mathbf{K}_{tt} \\ &\quad \mathbf{W}\mathbf{V}\mathbf{\Psi}^{-1/2}\mathbf{V}_{1m}\mathbf{V}_{1m}{}^T\mathbf{\Lambda}^{-1}\mathbf{\Psi}^{-1/2}\mathbf{V}^T\mathbf{W}\mathbf{K}_{tk}\end{aligned} \tag{35}$$

Where $\mathbf{K}_{kt} = \mathbf{\Phi}_k^T\mathbf{\Phi}_t$.

Finally, take variance for the row of generated new signal $\mathbf{S}_1$. Then, take logarithm and standardize as features:

$$f_k = \log\left(\frac{\operatorname{var}\left(\mathbf{S}_1\right)}{\sum_{i=1}^{K}\operatorname{var}\left(\mathbf{S}_i\right)}\right) \tag{36}$$

According to the above derivation, we can get $f_2, f_3, ..., f_K$, which correspond to the features of two class, three class,…, K class. So finally, corresponding to EEG data $\mathbf{X}_k$ of K classes, whose feature is described as $F^K = [f_1, f_2, ..., f_K]$.

C. *K-Means Clustering*

Considering the tremendous computational pressure eigenvalue decomposition for in eqn (26), we put the data clustering of the initial data in the low-dimensional space firstly. In the experiment, we used the K-means clustering method, whose results show that K-means clustering can effectively reduce the size of the data, so that it would be beneficial to further manipulation of the data.

K-means clustering is one of the widely used partition method in clustering at present, and usually uses the squared error function as the clustering criterion function. The operational principles of K-means clustering are as follows: firstly, choose K points randomly from the original data as the initial cluster center; then measure the distance from every sample to the selected cluster center and classify the sample as the class of its nearest cluster center; finally, calculate the mean of each new clustering data for the new cluster center. If two adjacent clustering centers do not change, sample adjustment is complete, and clustering criterion function has been convergence.

D. *Minimizing Within-Class Dissimilarities*

The large discrimination between the class means does not guarantee to have compact features with small scatters around the means. Since the EEG signals are non-stationary, there may be high trial-to-trial variations within a class resulting in deteriorated BCI performances [17]. Minimizing within-class dissimilarities through optimization function using KL divergence is extended to multi-class kernel space in the paper.

For the spatial filter $\mathbf{F}_1$ in MKCSSD, there is:

$$\mathbf{Y}_1\mathbf{F}_1^T = \mathbf{Y}_1^{'}\mathbf{F}_1^T\mathbf{\Lambda}_1 \tag{37}$$

Therefore, the MKCSSD algorithm, in computing the spatial filter $\mathbf{F}_1$ can be formulated as an optimization problem given by:

$$\begin{aligned}&\min_{\mathbf{w}_i}\left(\sum_{i=1}^{i=m}\mathbf{w}_i\mathbf{Y}_1^{'}\mathbf{w}_i^T + \sum_{i=1+m}^{i=2m}\mathbf{w}_i\mathbf{Y}_1\mathbf{w}_i^T\right) \\ &\textit{subjuct to}: \mathbf{w}_i\left(\mathbf{Y}_1 + \mathbf{Y}_1^{'}\right)\mathbf{w}_i^T = 1, i = \{1, 2, ..., 2m\} \\ &\mathbf{w}_i\left(\mathbf{Y}_1 + \mathbf{Y}_1^{'}\right)\mathbf{w}_j^T = 0,\ i, j = \{1, 2, ..., 2m\}, i \neq j\end{aligned} \tag{38}$$

Where $\mathbf{w}_i$ represents the first and the last m rows of $\mathbf{F}_1$.

In order to minimize the within-class dissimilarities, the dissimilarities between the trials of each class need to be measured. The Kullback-Leibler (KL) divergence is selected to compare the probability distribution function as a dissimilarite metric in the paper.

The hypodispersion of EEG trail passband filtered is defined as the covariance matrix of a zero mean, which is acquired through mean covariance matrix of multi-class EEG trails. According to the maximum entropy principle, the distribution pattern conforming to EEG data of zero mean covariance matrix is Gaussian distribution. The KL divergence between multivariate Gaussian distributions $N_0\left(\mu_0, \Sigma_0\right)$ and $N_1\left(\mu_1, \Sigma_1\right)$ is:

$$\begin{aligned}D\left(N_0 \middle| N_1\right) = 0.5\Big[&\left(\mu_1 - \mu_0\right)^T\mathbf{\Sigma}_1^{-1}\left(\mu_1 - \mu_0\right) \\ &+ trace\left(\mathbf{\Sigma}_1^{-1}\mathbf{\Sigma}_0\right) - \ln\left(\frac{\det\left(\mathbf{\Sigma}_0\right)}{\det\left(\mathbf{\Sigma}_1\right)}\right) - d\Big]\end{aligned} \tag{39}$$

Where det and d represents the determinant function and the dimensionality of the data respectively.

Based on the OVR method, using discrimination pattern of minimizing within-class dissimilarities for one class to all other classes, we can minimize signal dissimilarities within a class from EEG data of multiple classification. Therefore, minimizing the within-class dissimilarities of the EEG data is equivalent to minimizing the loss function:

$$\begin{aligned}&L\left(\begin{bmatrix}\mathbf{w}_1 \\ \mathbf{w}_2 \\ \vdots \\ \mathbf{w}_{2m}\end{bmatrix}\right) = L\left(\mathbf{w}\right) \\ &= \frac{1}{2}\Bigg[\frac{1}{N_1}\sum_{t=1}^{N_1}D\left(N\left(0, \mathbf{w}\mathbf{Y}_1^t\mathbf{w}^T\right)\middle|N\left(0, \mathbf{w}\mathbf{Y}_1\mathbf{w}^T\right)\right) \\ &\quad + \frac{1}{N^{'}}\sum_{t=1}^{N^{'}}D\left(N\left(0, \mathbf{w}\mathbf{Y}^{'t}\mathbf{w}^T\right)\middle|N\left(0, \mathbf{w}\mathbf{Y}^{'}\mathbf{w}^T\right)\right)\Bigg]\end{aligned} \tag{40}$$

Where $\mathbf{w}$ represents matrix which includes the first and last m spatial filters; N_1 or $N^{'}$ represents the number of epochs belonging to one class or all other classes respectively, and $N_1 = N^{'}$. $D(\square)$ represents the distribution of the t-th epoch in one class or all other classes from the average distribution in one class or all other classes.

So, In order to obtain the optimized spatial filters, the optimization function is given as:

$$\begin{aligned} &\min_{\mathbf{w}_i}(1-r)\left(\sum_{i=1}^{i=m}\mathbf{w}_i\mathbf{Y}_1^{'}\mathbf{w}_i^T+\sum_{i=1+m}^{i=2m}\mathbf{w}_i\mathbf{Y}_1\mathbf{w}_i^T\right)+rL(\mathbf{w}) \\ &\textit{subject to}:\mathbf{w}_i\left(\mathbf{Y}_1+\mathbf{Y}_1^{'}\right)\mathbf{w}_i^T=1, i=\{1,2,...,2m\} \\ &\mathbf{w}_i\left(\mathbf{Y}_1+\mathbf{Y}_1^{'}\right)\mathbf{w}_j^T=0, i,j=\{1,2,...,2m\}, i\neq j \end{aligned} \quad (41)$$

Where r $(0 \le r \le 1)$ is a regularization parameter to control the discrimination between and the similarity within the training class. The best r and N_1 values are selected by cross validation.

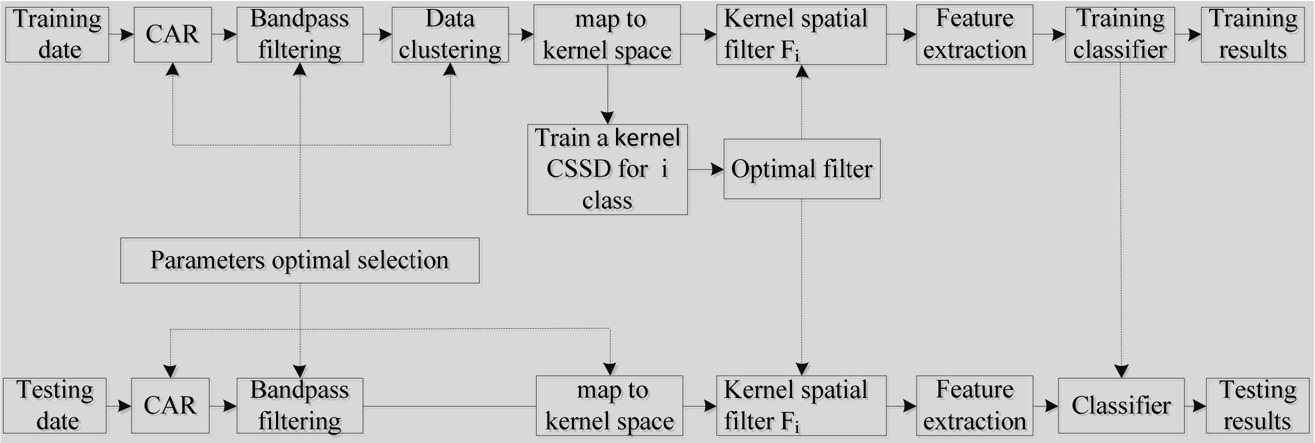


Figure1. Flowchart of classification based multi-class Kernel CSSD.

Finally, using fmincon function in MATLAB, we can find local minimum of these function and get the optimal filter. Similarly, the optimal filter of other classes can be obtained. We abbreviate the proposed algorithm as MK-KCSSD. Fig. 1 is the diagram of the classification based on multi-class kernel KLCSSD.

IV. THE RESULTS AND ANALYSIS OF SIMULATION EXPERIMENT

We select III_3a data set in 2005 BCI competition in the experiment, which is a full record of practical BCI's implement. There are three objects in this experiment, respectively for K3, K6 and L1. In the process of conducting experiments, subjects keep arms relaxed and natural in sitting in a chair. The experiment task is that subjects imagine the movement of left hand, right hand, foot, or the tongue based on random information. Experimental process: participants were to be in a quiet state in the first 2 seconds. Then the computer gave voice to say the trial begins and the screen appears "+". In the 3rd second when the computer screen appears to the left, right, up or down arrow, the subjects, respectively, acted in the direction shown by the arrow to imagine the movement of left hand, right hand, the tongue and foot. This continued until the 7th seconds. At this time, the sign "+" disappeared. The data was collected in by Neuroscan EEG amplifier, and the sampling frequency is 250 Hz. The data was 1~50 Hz band pass filter, which recorded the 60 EEG channel.

A. *Data Preprocessing*

As is known to all, EEG signals that were collected in scalp is a time-varying, nonlinear data. It is interfered by lots of noise signals. Therefore, choosing the appropriate preprocessing method for EEG feature extraction and classification based on the different data for the next step is very important. The preprocessing methods used in this experiment are mainly frequency domain and time window.

Frequency domain window: The processing data of the experiment is based on the experimental data of the imagination movement of four categories. According to the content of the second chapter, in the process of the imagination movement ERD and ERS appeared with mu and beta rhythm in. General imagination movement makes mu rhythm range from 8 to 12 Hz frequency band to produce rhythm changes, and it makes the beta rhythm change within 14-30 Hz frequency band. Because of individual differences, different object varied in its processing scope. The frequency domain windows of this experiment that we choose are: K3 and K6 choose 8~30 Hz, L1 choose 8~21 Hz.

Time window: Throughout the actual record in the EEG signal, a portion of the data is useful for classification, and part is redundant, whose contribution to the classification is not big. Those redundant data will even reduce the accuracy of classification. But if we select fewer data, we may lose some important information. Comprehensively considering, we use the 4~7 s imagination data, sampling of 250 Hz, and then the sampling points of each channel is 750. After the preprocessing, three objects of all experimental data were obtained. The K3 training data has 180 trials, the left hand, right hand, the tongue and feet 45 trials. Each trial data is matrix; In K3 test data, there are 180 trials. K6 and L1 training data each have 120 trials, the left hand, right hand, and the tongue and feet 30 trials. Each trial data is matrix; K6 and L1 test data each have 120 trials. Four categories of data in the left hand, right hand, tongue and feet respectively note for 1, 2, 3, and 4.

B. Comparative Analysis

In this experiment, CSSD, KCSSD and K-KCSSD were applied to multi-class classification on the test data of different objects respectively. The data preprocessing of three methods is the same, like we choice the same frequency window and time window. And the experimental simulation is finished under the same experimental platform, which is windows 7 operating syste, inter i3-2120 CPU, 3.3GHz main frequency and 4GB memory. Considering the data set used in the experiment is large, when the eigenvalue decomposition after mapping to kernel space through kernel methods, the computer can not calculate. In order to reduce the computational complexity, this problem is solved using K means clustering, namely in the low dimensional space data is gathered. The literature [18] handles also the computing problem of large matrices using this method, and conducts tests to prove. The large number of clusters means better clustering results, but it would take a long time. Generally if clustering number is more than 12, our computer will not calculate, considering we choose the clustering number is 12, so we consider selecting clustering number as 12.

Firstly, clustering is performed for sampling points in every test and each channel from the original data preprocessed. Then the date is mapped to the kernel space through kernel function. Because the relationship between the source signal in the brain and signals EEG recorded is mainly linear, but there exist nonlinear, so we choose the linear kernel function to solve the filter. In the training process using training data, we use cross validation to find the number the optimal spatial filters, namely selecting the best m value, and classify the training and testing data of each object. In order to exact solution, we use the method of 20 times and 10 fold cross validation, whose basic principle is: the training data is divided into 10 parts, nine of which in turn will be used as training and a parts of which will be used as testing. The mean of 20 results is used to estimate the accuracy of the algorithm. The best and values in the optimization function can be obtained by cross validation, and can also take a constant value. In order to save time, we selected and in this paper.

TABLE 1. THE CLASSIFICATION ACCURACY OF K3 TRAINING DATE UNDER CROSS VALIDATION THROUGH DIFFERENT EXTRACTION METHODS

m	Feature extraction methods: mean (±standard variance)					
	MCSSD		MKCSSD		MK-KCSSD	
	A	B	A	B	A	B
1	0.8278 (0.0021)	0.8019 (0.0089)	**0.9186 (0.0020)**	**0.8832 (0.0089)**	0.9280 (0.0013)	0.8979 (0.0091)
2	**0.9213 (0.0116)**	**0.8706 (0.0181)**	0.9395 (0.0013)	0.8717 (0.0137)	**0.9615 (0.0030)**	**0.8974 (0.0023)**
3	0.9050 (0.0350)	0.7821 (0.0313)	0.9955 (0.0084)	0.8854 (0.0148)	0.9955 (0.0055)	0.8834 (0.0041)
4	0.9969 (0.0346)	0.8642 ((0.0323)	1.0000 (0.0000)	0.8084 (0.0141)	0.9340 (0.0034)	0.8668 (0.0032)
5	1.0000 (0.0000)	0.8456 (0.0112)	0.9748 (0.0088)	0.8691 (0.0187)	0.9948 (0.0112)	0.8952 (0.0112)
6	0.9987 (0.0059)	0.8302 (0.0188)	1.0000 (0.0000)	0.8440 (0.0186)	0.9979 (0.0066)	0.8946 (0.0086)

Table 1 shows the 20 times and 10 fold cross validation classification results of the K3 training data using different feature extraction methods according to different number of spatial filters. From the results of the table we can find the optimal spatial filter for K3. Because the increase of m can not significantly improve the accuracy, considering the amount and time of the computation, we may be appropriate to select m value. We do the 20 times and 10 fold cross validation classification on training set for the remaining two objects K6 and L1 respectively, whose experimental results are as shown in table 2 and Table 3.

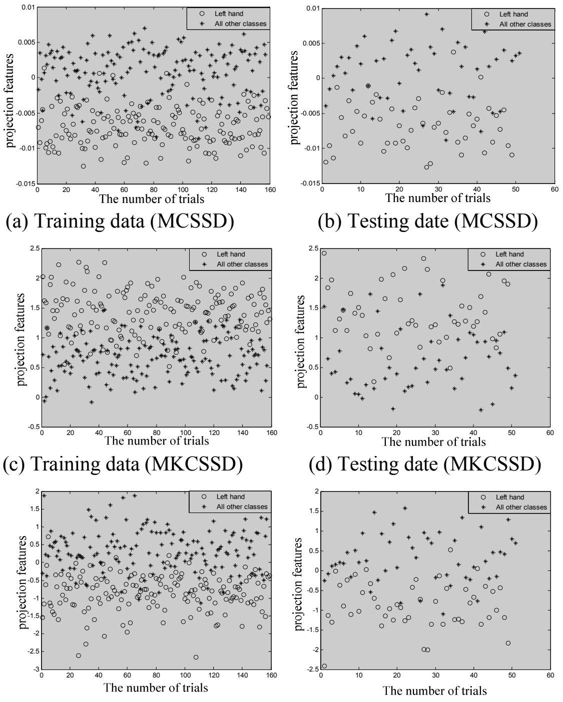
TABLE 2. THE CLASSIFICATION ACCURACY OF K6 TRAINING DATE UNDER CROSS VALIDATION THROUGH DIFFERENT EXTRACTION METHODS

m	Feature extraction methods: mean (±standard variance)					
	MCSSD		MKCSSD		MK-KCSSD	
	A	B	A	B	A	B
1	0.5036 (0.0058)	0.4238 (0.0265)	0.5642 (0.0052)	0.5121 (0.0230)	0.5889 (0.0053)	0.5289 (0.0150)
2	0.5390 (0.0046)	0.3971 (0.0215)	**0.6497 (0.0045)**	**0.5139 (0.0183)**	0.6032 (0.0052)	0.4676 (0.0239)
3	**0.6283 (0.0048)**	**0.4242 (0.0187)**	0.7218 (0.0045)	0.5447 (0.0219)	0.6828 (0.0052)	0.4892 (0.0202)
4	0.6515 (0.0065)	0.3594 (0.0258)	0.7522 (0.0104)	0.4820 (0.0236)	**0.7563 (0.0038)**	**0.5349 (0.0081)**
5	0.7658 (0.0059)	0.4737 (0.0234)	0.8175 (0.0063)	0.4910 (0.0302)	0.8185 (0.0073)	0.5005 (0.0223)
6	0.6996 (0.0087)	0.3912 (0.0269)	0.7506 (0.0179)	0.4251 (0.0186)	0.8518 (0.0055)	0.5223 (0.0259)

TABLE 3. THE CLASSIFICATION ACCURACY OF L1 TRAINING DATE UNDER CROSS VALIDATION THROUGH DIFFERENT EXTRACTION METHODS

m	Feature extraction methods: mean (±standard variance)					
	MCSSD		MKCSSD		MK-KCSSD	
	A	B	A	B	A	B
1	0.6613 (0.0037)	0.6225 (0.0165)	0.7311 (0.0043)	0.6746 (0.0147)	0.7020 (0.0044)	0.6567 (0.0137)
2	0.7482 (0.0039)	0.6838 (0.0139)	0.8084 (0.0042)	0.6942 (0.0234)	**0.8493** **(0.0015)**	**0.7837** **(0.0131)**
3	**0.8148** **(0.0030)**	**0.6958** **(0.0223)**	**0.8269** **(0.0029)**	**0.6871** **(0.0125)**	0.9183 (0.0036)	0.7450 (0.0157)
4	0.8659 (0.0057)	0.6471 (0.0309)	0.9079 (0.0137)	0.6704 (0.0271)	0.9277 (0.0422)	0.6946 (0.0324)
5	0.9492 (0.0053)	0.6567 (0.0216)	0.9610 (0.0263)	0.6317 (0.0380)	0.9635 (0.0369)	0.6267 (0.0324)
6	0.9604 (0.0187)	0.5651 (0.0300)	0.9501 (0.0365)	0.5842 (0.0334)	0.9834 (0.0280)	0.6109 (0.0423)

In the three tables, A and B represents respectively training and testing accuracy rate of training date. The literature shows that selecting multiple eigenvalue cannot improve the classification accuracy rate greatly, so in the experiment we will only get m to 6. For the testing set of object K3, the classification accuracy of the three algorithms have little difference. The classification accuracy of MK-KCSSD is slightly higher. For object L1, the classification accuracy of MKCSSD outperformed the MCSSD and MKCSSD algorithms by an average of 3.57% and 1.36%, respectively It can be shown that MK-KCSSD can extract effective features of multi-class EEG data, and can obtain better classification results. For object K6, the classification accuracy of the former two algorithms are both relatively low. The low classification accuracy could be due to the fact that different objects' own reasons, such as psychological state, state fatigue, which can affect the collection of experimental data and affect the next step of feature extraction and classification. Therefore, it is important that experimental objects maintain good mood, mental calm, and devote theirselves heart and soul to complete the experimental task. MK-KCSSD has obvious advantage. While the algorithm can greatly reduce within-class dissimilarities of EEG date by using the optimization functions, the between-classes distance is increased slightly, which can ease the unsteady and human factors damage effect for the test.



(a) Training data (MCSSD) (b) Testing date (MCSSD)

(c) Training data (MKCSSD) (d) Testing date (MKCSSD)

(e) Training data (MK-KCSSD) (f) Testing date (MK-KCSSD)

Figure2. Features projection based on different extraction methods

The features projection for the object K6 which are extracted through the three methods and done linear discriminant analysis in the case of optimal parameters is shown in Fig. 2. The figure is intended only to look at the projection effect. We use all the dimensional features, in the classification of the features. According to the comparison of the training features distributions, it is clear that KCSSD and CSSD can reach the same projection effect. The features extracted by KL-MKCSSD are more compact and more separable. The phenomenon suggests that KL-MKCSSD can successfully extract most the non-smooth change from the testing data of the object K6, and the dissimilarities between the distribution of training and testing features are limited to small change.

V. CONCLUSIONS

The proposed MK-KCSSD algorithm is that CSSD algorithm is extended to the kernel space, and applied to multi-class classification. To obtain the optimal filter, we perform 20 times and 10 fold cross validation on the date of training set, and solve the problems of choosing the optimal parameters m. Then in order to improve further classification accuracy of the test set, we optimize the filters using the optimization function based on KL divergence. For the problem of computational complexity, the sampling points for each channel in each trial were clustered through K-means clustering, and it meets the requirements of computer operation. Based on the simulation of III_3a data set in 2005 BCI competition, MK-KCSSD classification accuracy on the testing set is higher than that of the traditional MCSSD, especially the change of the non-stationary feature, which shows that MK-KCSSD can extract better the corresponding class features of multi-class single trial EEG data.

ACKNOWLEDGMENTS

This work is supported by The National Natural Science Foundation, China (No. 61473339). This work is supported by HeBei National Natural Science Foundation, China (No.F2012203084). This work is supported by HeBei National Natural Science Foundation, China (No.A2012203124).

REFERENCES

[1] Wolpaw J R, Birbaumer N, Mcfarland DJ, et al. Brain-Computer Interfaces for Communication and Control. Clinical Neurophysiology, 113(6), pp. 767-791, 2002.

[2] Schalk G, Mellinger J. A Practical Guide to Brain-Computer Interfacing with BCI2000. Springer-Verlag: London, pp. 2-4, 2010.

[3] Donoghue J P, Nurmikko A, Black M, et al. Assistive technology and robotic control using motor cortex ensemble-based neural interface systems in humans with tetraplegia. J.Physiol, 579(3), pp. 603-611, 2007.

[4] Pfurtscheller G, Guger C, Muller G, et al. Brain oscillations control hand orthosis in a tetraplegic. Neuroscience Letters, 3(292), pp. 211-214, 2000.

[5] Guger C, Ramoser H, Pfurtscheller G. Real-time EEG analysis with subject-specific spatial patterns for a brain-computer interface (BCI). IEEE Trans Rehabil Eng, 8(4), pp. 447-456, 2000.

[6] Lemm S, Blankertz G B, Curio G, Muller K R. Spatio-spectral filters for improving the classification of single trial EEG. IEEE Transactions on Biomedical Engineering, 52 (9), pp. 1541-1548, 2005.

[7] Haiping Lu, How-Lung Eng, Cuntai Guan, et al. Regularized common spatial pattern with aggregation for EEG classification in small-sample setting. IEEE Transactions on Biomedical Engineering, 57(12), pp. 2936-2946, 2010.

[8] Thanh Ha Nguyen, Seung-Min Park, Kwang-Eun Ko, et al. Invariant common spatial pattern advanced feature extraction in mu rhythms of EEG signals. 2012-38th Annual Conference on IEEE Industrial Electronics Society, IECON, IEEE: Montreal, pp. 1535-1539, 2012.

[9] Thanh Ha Nguyen, Seung-Min Park, Kwang-Eun Ko, et al. Multi-class stationary CSP for optimal feature separation of brain source in BCI system. 2012 12th International Conference on Control, Automation and Systems (ICCAS), IEEE: JeJu Island, pp.1035 - 1039, 2012.

[10] Dornhege G, Blankertz B, Krauledat M,et al. Combined optimization of spatial and temporal filters for improving brain-computer interfacing. IEEE Transactions on Biomedical Engineering, 53(11), pp. 2274-2281, 2005.

[11] Dornhege G, Blankertz B, Krauledat M, et al. Optimizing spatio-temporal filters for improving brain-computer interfacing. Advances in Neural Information Processing Systems, NIPS 2005, DBLP: Vancouver, pp. 315-322, 2005.

[12] Dornhege G, Blankertz B, Curio G, et al. Boosting bit rates in noninvasive EEG single-trial classifications by feature combination and multiclass paradigms. Biomedical Engineering, 51(6), pp. 993–1002, 2004.

[13] Wang Y H, Berg P, Scherg M. Common spatial subspace decomposition applied to analysis of brain responses under multiple task conditions: a simulation study. Clinical Neurophysiology, 110 (4), pp. 604-614, 1999.

[14] Wang Lu, Wu Xiaopei, Gao Xiangping. Analysis and classification of four-class motor imagery EEG data. Computer Technology and Development, 18(10), pp. 23-26, 2008.

[15] Nasihatkon B. Boostani R. Jahromi M. Z. An efficient hybrid linear and kernel CSP approach for EEG feature extraction. Neurocomputing, 73 (1), pp. 432-437, 2009.

[16] Arvaneh M, Cuntai Guan, Kai Keng Ang, Chai Quek. Optimizing spatial filters by minimizing within-class dissimilarities in electroencephalogram-based Brain-Computer Interface. IEEE Transactions on Neural Networks and Learning Systems, 24(4), pp. 610-619, 2013.

[17] Wang Jinjia, Zhang Lingzhi, Hu Bei. Multi-class kernel CSP-based feature extraction methods. Journal of Biomediacal Engineering, 29(2), pp. 217-222, 2012.

[18] LIAO X, YAO DZ, WU.D. Combining spatial filters for the classification of single-trial EEG in a finger movement task. IEEE Trans BME, 54(5), pp. 821-831, 2007..

.