2017 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 25–28, 2017, TOKYO, JAPAN

CORRENTROPY INDUCED METRIC BASED COMMON SPATIAL PATTERNS

Jiyao Dong[1]*,Badong Chen*[1]*,Na Lu*[1]*,Haixian Wang*[2]*,Nanning Zheng*[1]

[1]School of Electronic and Information Engineering, Xian Jiaotong University, Xian 710049, China
[2]Institute of Child Development and Education,Research Center for Learning Science, Southeast University, Nanjing, Jiangsu 210096, China

ABSTRACT

Common spatial patterns (CSP) is a widely used method in the field of electroencephalogram (EEG) signal processing. The goal of CSP is to find spatial filters that maximize the ratio between the variances of two classes. The conventional CSP is however sensitive to outliers because it is based on the L_2-norm. Inspired by the correntropy induced metric (CIM), we propose in this work a new algorithm, called CIM based CSP (CSP-CIM), to improve the robustness of CSP with respect to outliers. The CSP-CIM searches the optimal solution by a simple gradient based iterative algorithm. A toy example and a real EEG dataset are used to demonstrate the desirable performance of the new method.

Index Terms— Common spatial patterns, robustness, correntropy induced metric, kernel bandwidth

1. INTRODUCTION

Brain-computer interfaces (BCI) can translate brain signals to control commands and help severely paralyzed people communicate with external worlds without muscle operations [1]. Electroencephalography (EEG) is a widely used brain signal for BCI due to its non-invasiveness and convenience. One of the most important issues of EEG-based BCI is to classify the EEG signals robustly and accurately [2].

In order to extract useful discriminative features, so far many methods have been developed and the common spatial patterns (CSP) is one of the most effective algorithms to analyze multichannel EEG data of two classes [3]. The CSP searches the spatial filters which maximize the variance of one class and minimize the variance of the other class. Due to its popularity and efficiency, researchers have developed many variants of CSP, such as common spatio-spectral patterns (CSSP) [4], local temporal CSP (LTCSP) [5], regularized CSP (RCSP) [6], and canonical correlation approach to CSP (CCACSP) [7].

The conventional CSP and most of its variants are sensitive to outliers because they are developed based on the estimation of covariance matrices where L_2-norm is used. The L_2-norm may amplify the negative effects of outliers in multichannel EEG data and result in inappropriate spatial filters [8]. Therefore, not only the spatial filters of CSP but also the features of every single trial are affected by the outliers. It is necessary and important to develop robust variants of CSP to alleviate the negative effects of outliers [9–11]. The CSP-L_1 is one of the robust variants, which improves the conventional CSP method by using the L_1-norm instead of L_2-norm in the objective function. The literatures in the field of machine learning show that L_1-norm can alleviate the negative influence of outliers and achieve good learning performance [12, 13]. In [14], the common spatial patterns based on generalized norms was proposed. But the authors of [14] pointed out that the L_1-norm is still the best choice considering the implementation and complexity of CSP-L_p.

In this paper, we propose a new robust CSP algorithm based on the correntropy induced metric (CIM) which can approximate the L_2, L_1 and L_0 norms in different dynamic regions [15,16]. The CIM is defined based on the correntropy concept [15–19], a generalized correlation in Reproducing Kernel Hilbert Space (RKHS). In essence the CIM is an "L_2-norm" in kernel space, which corresponds to a nonlinear distance measure in input space. The CIM will saturate when two vectors are far apart from each other in the input space and this property makes it appropriate for the learning problems with outliers in the data. The CIM-based CSP will be robust to outliers and hence can yield better spatial filters comparing with conventional methods especially when outliers appear.

The rest of the paper is organized as follows. Section 2 describes the original CSP method. Section 3 briefly introduces the concepts of correntropy and CIM. Section 4 derives the CSP-CIM algorithm. In section 5, experiments are conducted to demonstrate the robustness of the proposed method. Finally, conclusion is given in section 6.

2. CSP ALGORITHM

CSP is proposed to deal with two-class problems. Let $\mathbf{X}_i \in \mathbb{R}^{c\times l}$ denotes the EEG data of one class and $\mathbf{Y}_j \in \mathbb{R}^{c\times l}$ denotes data of the other class, where c is the number of chan-

This work was supported by 973 Program (No. 2015CB351703,No. 2015CB351704) and National NSF of China (No. 91648208).

978-1-5090-6341-3/17/$31.00

nels and l the number of samples. Suppose that the number of trials of the two classes are N_x and N_y respectively. All the EEG trials can be represented as $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_{N_x}] \in \mathbb{R}^{c\times \mathrm{m}}$ and $\mathbf{Y} = \left[\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_{N_y}\right] \in \mathbb{R}^{c\times \mathrm{n}}$ where m equals to $l \times N_x$ and n equals to $l \times N_y$. Before these trial segments are applied, they should be band-pass filtered, centered, and scaled. The covariance matrices are $\mathbf{R}_x = \frac{1}{N_x}\mathbf{X}\mathbf{X}^T$ and $\mathbf{R}_y = \frac{1}{N_y}\mathbf{Y}\mathbf{Y}^T$. CSP is designed to calculate spatial filters that maximize the variance of one class and minimize the variance of the other class. Mathematically, this purpose can be achieved by maximizing the objective function [6]

$$J(\omega) = \frac{\omega^T \mathbf{R}_x \omega}{\omega^T \mathbf{R}_y \omega} \tag{1}$$

where ω is the spatial filter that needs to be solved. This problem can be solved by the generalized eigenvalue equation

$$\mathbf{R}_x \omega = \lambda \mathbf{R}_y \omega. \tag{2}$$

In the classification, only a few of eigenvectors corresponding to the largest and smallest eigenvalues are used.

3. CORRENTROPY AND CIM

Given two vectors

$$\mathbf{x} = [x_1, x_2, \cdots, x_N]$$
$$\mathbf{y} = [y_1, y_2, \cdots, y_N]$$

the correntropy is defined by

$$V(\mathbf{x}, \mathbf{y}) = \frac{1}{N}\sum_{i=1}^{N} \kappa(x_i, y_i) \tag{3}$$

where $\kappa(\cdot,\cdot)$ is a shift-invariant kernel and the most commonly used is the Gaussian kernel:

$$\kappa(x, y) = \kappa(x - y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right) \tag{4}$$

where σ is the kernel bandwidth. Since the Gaussian kernel satisfies Mercer's theorem [20], one can find a nonlinear mapping $\Phi : \mathbb{R} \to \mathbb{R}^{\mathbb{R}}$ to transfer data from the input space into a RKHS. The inner product between two vectors in RKHS can be evaluated by

$$\langle \Phi(x), \Phi(y) \rangle = \kappa(x, y). \tag{5}$$

With this nonlinear transformation, $\mathbf{x}$ and $\mathbf{y}$ are mapped as

$$\hat{\mathbf{x}} = [\Phi(x_1), \Phi(x_2), \ldots, \Phi(x_N)]$$
$$\hat{\mathbf{y}} = [\Phi(y_1), \Phi(y_2), \ldots, \Phi(y_N)].$$

In RKHS, the "Euclidean distance" between two vectors can be calculated by

$$\left[(\hat{\mathbf{x}} - \hat{\mathbf{y}})^T (\hat{\mathbf{x}} - \hat{\mathbf{y}})\right]^{\frac{1}{2}} = \sqrt{2N}[\kappa(0) - V(\mathbf{x}, \mathbf{y})]^{\frac{1}{2}}. \tag{6}$$

According to (6), CIM is defined by

$$CIM(\mathbf{x}, \mathbf{y}) = [\kappa(0) - V(\mathbf{x}, \mathbf{y})]^{\frac{1}{2}}. \tag{7}$$

As shown in [15], the CIM (7) behaves like an L_2-norm of $\mathbf{x} - \mathbf{y}$ when two vectors are close, and this area is Euclidean zone; with the distance becoming larger, CIM behaves like an L_1-norm which is Transition zone; finally, when two vectors are further apart, CIM saturates and approaches Rectification zone (L_0-norm). This interesting property suggests that the CIM can be applied as a robust or sparsity-aware cost for machine learning and signal processing. The kernel bandwidth σ is an important free parameter in CIM, which controls the scale of the CIM metric. As σ becomes larger, the linear region (Euclidean area) is enlarged.

4. CIM-BASED CSP

The conventional CSP method is based on L_2-norm and one can rewrite (1) as

$$J(\omega) = \frac{\frac{1}{N_x}\left\|\omega^T \mathbf{X}\right\|_2^2}{\frac{1}{N_y}\left\|\omega^T \mathbf{Y}\right\|_2^2} \tag{8}$$

where $\|\cdot\|_2$ denotes the L_2-norm. According to (8), CSP is sensitive to outliers, because the L_2-norm will amplify the negative effects of large deviations. EEG signals are usually polluted by artifacts and noises. Therefore, it is necessary to use a robust cost function instead of the L_2-norm to improve the robustness of the conventional CSP. A robust CSP algorithm was proposed in [11] by using the L_1-norm. In the present work, to further improve the performance, we use the CIM to replace the L2-norm to develop a new robust CSP algorithm. Specifically, we propose the following objective function:

$$\tilde{J}(\omega) = \frac{CIM\left(\omega^T \mathbf{X}, \mathbf{0}\right)}{CIM\left(\omega^T \mathbf{Y}, \mathbf{0}\right)} = \frac{\left[\kappa(0) - V\left(\omega^T \mathbf{X}, \mathbf{0}\right)\right]^{\frac{1}{2}}}{\left[\kappa(0) - V\left(\omega^T \mathbf{Y}, \mathbf{0}\right)\right]^{\frac{1}{2}}}. \tag{9}$$

The above cost function can be rewritten as

$$\tilde{J}(\omega) = \left[\frac{\frac{1}{m}\sum_{i=1}^{m}(1 - E_i)}{\frac{1}{n}\sum_{j=1}^{n}(1 - E_j)}\right]^{\frac{1}{2}} \tag{10}$$

where

$$E_i = \exp\left(-\frac{\left(\omega^T \mathbf{x}_i\right)^2}{2\sigma^2}\right), E_j = \exp\left(-\frac{\left(\omega^T \mathbf{y}_j\right)^2}{2\sigma^2}\right).$$

Algorithm 1 Procedure of CSP-CIM

1: Set $t = 0$. Initialize the kernel bandwidth σ, $\omega(t)$ and a set of different values of learning rate η.
2: Calculate the gradient of $\log J(\omega)$ with respect to $\omega(t)$ according to equation (12).
3: For every η, compute the corresponding value of the objective function using the updated spatial filter (by (13)) and equation (11), and choose the one resulting in the largest value to update $\omega(t+1)$. Set $t \leftarrow t+1$.
4: Check if the stopping criteria is satisfied, otherwise, go to step 2.
5: Output $\omega(t)$.

Taking logarithm of both sides, we can obtain the equation

$$\begin{aligned}\log \tilde{J}(\omega) = &\frac{1}{2}\log\left[1-\frac{1}{m}\sum_{i=1}^{m}E_i\right] \\ &-\frac{1}{2}\log\left[1-\frac{1}{n}\sum_{j=1}^{n}E_j\right].\end{aligned} \tag{11}$$

The gradient descent based methods can be used to search the optimal solution of (11). We calculate the gradient of $\log \tilde{J}(\omega)$ with respect to ω as

$$\begin{aligned}&\frac{\partial \log \tilde{J}(\omega)}{\partial(\omega)} \\ &= \frac{1}{2}\frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i E_i\left(\frac{\omega^T\mathbf{x}_i}{\sigma^2}\right)\Big/\left(1-\frac{1}{m}\sum_{i=1}^{m}E_i\right) \\ &-\frac{1}{2}\frac{1}{n}\sum_{j=1}^{n}\mathbf{y}_j E_j\left(\frac{\omega^T\mathbf{y}_j}{\sigma^2}\right)\Big/\left(1-\frac{1}{n}\sum_{j=1}^{n}E_j\right).\end{aligned} \tag{12}$$

At time t, the spatial filter can be represented as $\omega(t)$, and $\omega(t+1)$ is updated by

$$\omega(t+1) = \omega(t) + \eta\frac{\partial \log \tilde{J}(\omega)}{\partial(\omega)}. \tag{13}$$

Based on (13), the procedure of optimizing the objective function (10) is summarized in Algorithm 1. A set of different values of η are tried in the step 3 to make the spatial filter converge to the optimal solution fast and stably. In every step, the filter should be rescaled as unit length as suggested in [11]. Following [11], we also require any two spatial filters to be orthogonal. Specifically, when the first h spatial filters $\omega_1, \omega_2, \ldots, \omega_h$ are obtained, the $(h+1)th$ spatial filter that maximizes the objective function (9) should satisfy the constraint $\omega_1^T\omega_{h+1} = \omega_2^T\omega_{h+1} = \ldots = \omega_h^T\omega_{h+1} = 0$.

After acquiring a set of orthonormal spatial filters $\omega_1, \omega_2, \ldots, \omega_p$ to maximize (9), we calculate another set of orthonormal spatial filters $\omega'_1, \omega'_2, \ldots, \omega'_q$ to maximize $1/\tilde{J}(\omega)$ using a similar procedure. With these filters, features of every

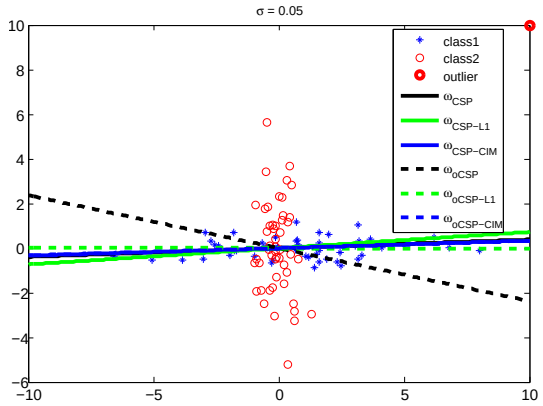

Fig. 1. Toy example. The data points of two classes and the spatial filters of conventional CSP, CSP-L_1 and the proposed CSP-CIM with (dashed) and without (solid) the outlier.

EEG trial segment $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_l] \in \mathbb{R}^{c\times l}$ are represented as $\mathbf{f} = \left[f_1, \ldots, f_p, {f'}_1, \ldots, {f'}_q\right]^T$, where

$$\begin{aligned}f_{k'} &= \left(\frac{1}{l}\sum_{i=1}^{l}\left(1-\exp\left(-\frac{\left(\omega_{k'}{}^T\mathbf{z}_i\right)^2}{2\sigma^2}\right)\right)\right)^{\frac{1}{2}} \\ {f'}_k &= \left(\frac{1}{l}\sum_{i=1}^{l}\left(1-\exp\left(-\frac{\left({\omega'}_k{}^T\mathbf{z}_i\right)^2}{2\sigma^2}\right)\right)\right)^{\frac{1}{2}}\end{aligned} \tag{14}$$

with $1 \le k' \le p, 1 \le k \le q$. The feature vector $\mathbf{f}$ is a $(p+q)$-dimensional vector.

In our study, Linear Discriminant Analysis (LDA) is used to predict the labels [21]. For two-class problems, LDA projects the $(p+q)$-dimensional feature vector into a one-dimensional space such that the ratio of the between-class distance to the within-class distance is maximized. During the training phase, the centers of the projected EEG features of each class and the projection vector are obtained. During the testing phase, we calculate the distances between the projected testing sample and the centers, and classify the sample into the class with smaller distance.

5. EXPERIMENTS

Two datasets are used to compare the performance of the proposed CSP-CIM to the conventional CSP and CSP-L_1 [11]. The first one is a toy dataset, in which an outlier is included. The second one is the dataset IIa of BCI competition IV, a publicly available real EEG dataset. We set the initial vectors of CSP-L_1 and CSP-CIM as the solutions of the standard CSP method.

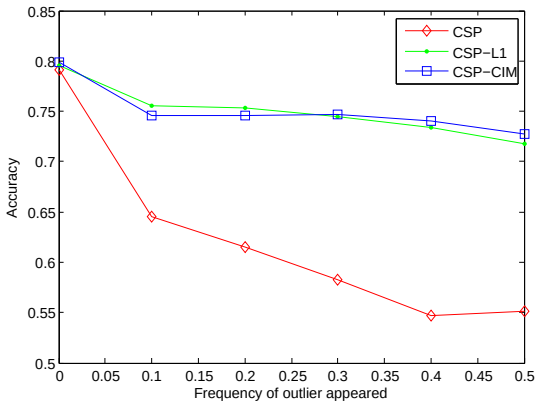

Fig. 2. Average classification accuracies across all subjects of conventional CSP, CSP-L_1 and the proposed CSP-CIM with increasing occurrence frequencies of outliers.

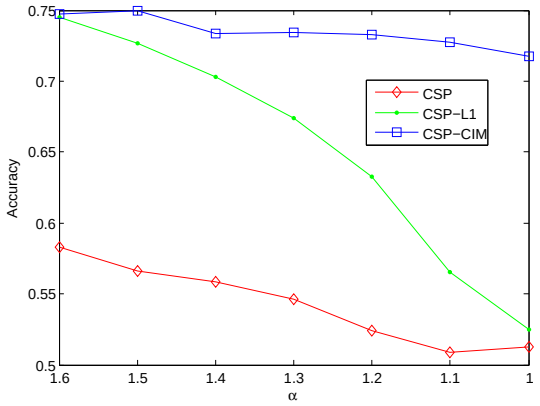

Fig. 3. Average classification accuracies across all nine subjects of conventional CSP, CSP-L_1 and the proposed CSP-CIM when changing the parameter α.

5.1. Toy dataset

A two-dimensional artificial dataset is utilized. We produce the simulated data according to two Gaussian distributions with zero means and covariance matrices diag(0.2,5) and diag(5,0.2). As shown in Fig.1, 50 points of each class are generated according to the two distributions. The samples of one class are specified by "o" and the other by "*". Furthermore, a point [10, 10] is added into class "o" in order to investigate the influence of outliers on the conventional CSP and the proposed CSP-CIM. We plot the filter which minimizes the dispersion of class "o".

First, we apply CSP, CSP-L_1 and CSP-CIM on the original data without the outlier. As shown in Fig.1, the spatial filters (in solid) of the three algorithms are similar, with $\omega_{CSP-CIM} = [-0.9994, -0.0335]$, $\omega_{CSP} = [-0.9867, -0.0392]$, $\omega_{CSP-L_1} = [-0.9973, -0.0728]$.

When the outlier occurs, the spatial filters(i.e. projection vectors) of CSP, CSP-L_1 and CSP-CIM can be computed as $\omega_{oCSP-CIM} = [-0.9995, -0.0327]$, $\omega_{oCSP} = [-0.9254, 0.2735]$ and $\omega_{oCSP-L_1} = [-1.0000, 0.0035]$ respectively. The learning rate parameter η of CSP-L_1 is 0.01. The kernel bandwidth σ and learning rate η of CSP-CIM are set at 0.05 and 0.1 respectively. From Fig.1, the filter obtained by the conventional CSP method is badly affected by the outlier, while those obtained by CSP-L_1 and CSP-CIM successfully suppress the negative influence of the outlier. In particular, the CSP-CIM performs better than the CSP-L1 and the projection vectors it obtains with and without the outlier are almost the same.

5.2. Real EEG Dataset

Next, we use the dataset IIa of BCI competition IV to illustrate the performance of the proposed approach. The EEG data are recorded from nine subjects with twenty-two electrodes. There are four different motor imagery tasks, i.e. left hand, right hand, both feet and tongue [22]. For each individual, two sessions on different days are recorded. Every session consists of 6 runs and every run is comprised of 48 trials. Therefore, there are 288 trials per session with 72 trials per class. The sample frequency is 250Hz. The EEG signals are already band-pass filtered between 0.5Hz and 100Hz with a 50Hz notch filter enabled. We only classify the trials of left hand and right hand in the experiments. According to [6], the time segments from 0.5s to 2.5s after the visual cue are extracted as samples. We preprocess the EEG segments using a fifth order Butterworth filter with cutoff frequencies 8 Hz and 35Hz. For all the CSP algorithms, six spatial filters are used as suggested in [11], which means p=q=3.

Similar to [11], outlier vectors are added into the training data to demonstrate the robustness of the proposed method. The outliers are generated by the 22-dimensional Levy alpha-stable distribution with characteristic exponent ($0 < \alpha \leq 2$), skewness (-1 $\leq \beta \leq$ 1), scale parameter ($0 < \gamma < \infty$) and location parameter (-$\infty < \delta < \infty$). The time positions to add outliers are selected randomly. Every experiment is independently repeated for 10 times with the average accuracy being evaluated.

For CSP-L_1 and CSP-CIM, in each iteration, different values of η between 1e-5 and 2.5 are used and the one resulting in the largest objective value is employed to update $\omega(t+1)$. For CSP-CIM, the kernel bandwidth σ is chosen between 0.001 and 1.0 by five-fold cross-validation.

In Fig.2, we present the average classification accuracies across all subjects of conventional CSP, CSP-L_1 and the proposed CSP-CIM with increasing occurrence frequencies of outliers. The four parameters of alpha-stable distribution are set as $[\alpha, \beta, \gamma, \delta] = [1.6, 0, 0.001, 0]$. The number of outliers added in the training dataset varied from 0 to $0.5(m+n)$ with step $0.1(m+n)$. When no outliers appear, all methods show almost the same performance. However, as the frequency of outliers increases, the classification accuracies of the

α	1.6	1.5	1.4	1.3	1.2	1.1	1	avg
Subject A01E								
CSP	61	57	55	52	49	50	49	53
CSP-L_1	**85**	82	78	71	61	55	52	69
CSP-CIM	83	**85**	**81**	**85**	**86**	**85**	**81**	**84**
Subject A02E								
CSP	52	52	53	53	51	50	50	51
CSP-L_1	58	**57**	**54**	53	51	**51**	51	53
CSP-CIM	**59**	**57**	53	**54**	**54**	**51**	**54**	**55**
Subject A03E								
CSP	69	65	61	57	56	51	57	59
CSP-L_1	**94**	91	89	82	71	59	56	77
CSP-CIM	**94**	**95**	**93**	**92**	**94**	**93**	**91**	**93**
Subject A04E								
CSP	51	52	51	50	50	50	49	51
CSP-L_1	63	63	59	60	58	53	51	58
CSP-CIM	**67**	**66**	**64**	**66**	**63**	**63**	**63**	**64**
Subject A05E								
CSP	54	54	51	49	50	49	51	51
CSP-L_1	**56**	52	51	51	50	50	49	51
CSP-CIM	**56**	**58**	**57**	**54**	**53**	**54**	**51**	**55**
Subject A06E								
CSP	53	52	52	53	53	51	50	52
CSP-L_1	**67**	64	64	57	53	50	50	58
CSP-CIM	66	**65**	**65**	**65**	**64**	**65**	**62**	**65**
Subject A07E								
CSP	50	50	51	52	51	50	50	50
CSP-L_1	62	61	55	53	51	51	51	55
CSP-CIM	**63**	**63**	**61**	**60**	**60**	**58**	**57**	**60**
Subject A08E								
CSP	77	70	68	66	61	57	53	65
CSP-L_1	**94**	**95**	94	89	87	71	55	83
CSP-CIM	**94**	**95**	**95**	**94**	**94**	**94**	**95**	**94**
Subject A09E								
CSP	58	58	62	60	51	50	53	56
CSP-L_1	**91**	**90**	90	89	88	70	58	82
CSP-CIM	90	**90**	**92**	**90**	**91**	**91**	**91**	**91**

Table 1. Classification accuracies (%) of all subjects of CSP, CSP-L_1 and CSP-CIM. The best accuracies in every case are shown in bold.

conventional CSP decrease rapidly. On the contrary, CSP-L_1 and CSP-CIM peform much better owing to their robustness to outliers. When a small number of outliers appear, CSP-L_1 performs a little better than CSP-CIM. But as the frequency of outliers increases CSP-CIM behaves relatively better. In Fig.3, we show the classification accuracies when the parameter α is decreasing from 1.6 to 1.0 with step 0.1 where the number of outliers in the training dataset of each subject is $0.3\,(m+n)$. Note that when α becomes smaller, the outliers will be more impulsive. One can see that the classification accuracies of conventional CSP method and CSP-L_1 decrease

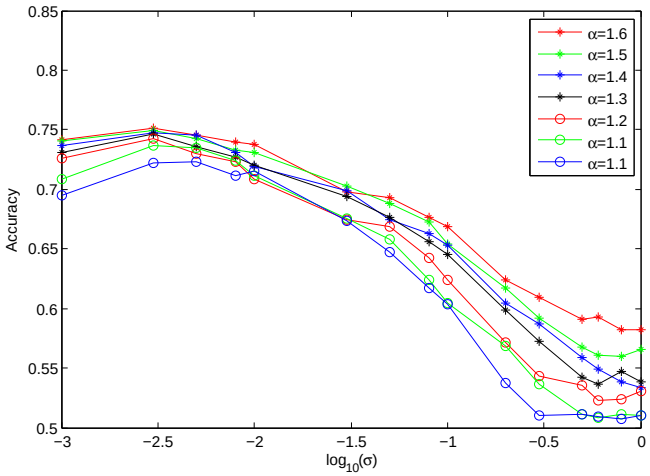

Fig. 4. Average classification accuracies across nine subjects with the kernel bandwidth increasing from 0.001 to 1.

rapidly with α decreasing. But the accuracies of the proposed CSP-CIM decrease very slowly, showing strong robustness to impulsive noises. This observation suggests that our algorithm is more robust to large outliers than CSP-L_1. In Table 1, we present the classification accuracies(%) of every subject and the best accuracies are shown in bold.

Finally, we investigate how the kernel bandwidth σ influences the performance of CSP-CIM. The classification accuracies with different kernel bandwidths are shown in Fig. 4. We see that when the kernel bandwidth σ is large, the performance is poor because CIM behaves like an L_2-norm which amplifies the negative effects of outliers. With a smaller kernel bandwidth (say σ = 0.003), the CSP-CIM achieves very good performance. However, if the kernel bandwidth is too small (e.g. smaller than 0.001), the performance of CSP-CIM will degrade. Thus it is very important to select a proper kernel bandwidth to achieve desirable performance. But how to select the best parameter is a challenging problem and this is an interesting topic for future study.

6. CONCLUSION

We propose a new algorithm CSP-CIM which is an improvement of the conventional common spatial patterns method. The correntropy induced metric (CIM) which can approximate the L_2, L_1 and L_0 norms in different dynamic regions, is used to build a robust objective function. The new objective function is optimized by using a gradient-based iterative algorithm. This new approach is proposed mainly to deal with motor imagery EEG data with outliers in the training dataset. We use two datasets, a toy example and dataset IIa of BCI competition IV to demonstrate the desirable performance of the new method.

7. REFERENCES

[1] Jonathan R Wolpaw, Niels Birbaumer, Dennis J McFarland, Gert Pfurtscheller, and Theresa M Vaughan, "Brain–computer interfaces for communication and control," *Clinical neurophysiology*, vol. 113, no. 6, pp. 767–791, 2002.

[2] Ali Bashashati, Mehrdad Fatourechi, Rabab K Ward, and Gary E Birch, "A survey of signal processing algorithms in brain–computer interfaces based on electrical brain signals," *Journal of Neural engineering*, vol. 4, no. 2, pp. R32, 2007.

[3] Zoltan J Koles, Michael S Lazar, and Steven Z Zhou, "Spatial patterns underlying population differences in the background eeg," *Brain topography*, vol. 2, no. 4, pp. 275–284, 1990.

[4] Steven Lemm, Benjamin Blankertz, Gabriel Curio, and K-R Muller, "Spatio-spectral filters for improving the classification of single trial eeg," *IEEE transactions on biomedical engineering*, vol. 52, no. 9, pp. 1541–1548, 2005.

[5] Haixian Wang and Wenming Zheng, "Local temporal common spatial patterns for robust single-trial eeg classification," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 16, no. 2, pp. 131–139, 2008.

[6] Fabien Lotte and Cuntai Guan, "Regularizing common spatial patterns to improve bci designs: unified theory and new algorithms," *IEEE Transactions on biomedical Engineering*, vol. 58, no. 2, pp. 355–362, 2011.

[7] Eunho Noh and Virginia R de Sa, "Canonical correlation approach to common spatial patterns," in *Neural Engineering (NER), 2013 6th International IEEE/EMBS Conference on*. IEEE, 2013, pp. 669–672.

[8] Benjamin Blankertz, Ryota Tomioka, Steven Lemm, Motoaki Kawanabe, and K-R Muller, "Optimizing spatial filters for robust eeg single-trial analysis," *IEEE Signal processing magazine*, vol. 25, no. 1, pp. 41–56, 2008.

[9] Motoaki Kawanabe, Carmen Vidaurre, Simon Scholler, and Klaus-Robert Muuller, "Robust common spatial filters with a maxmin approach," in *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*. IEEE, 2009, pp. 2470–2473.

[10] Wojciech Samek, Duncan Blythe, Klaus-Robert Müller, and Motoaki Kawanabe, "Robust spatial filtering with beta divergence," in *Advances in Neural Information Processing Systems*, 2013, pp. 1007–1015.

[11] Haixian Wang, Qin Tang, and Wenming Zheng, "L1-norm-based common spatial patterns," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 3, pp. 653–662, 2012.

[12] Nojun Kwak, "Principal component analysis based on l1-norm maximization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 9, pp. 1672–1680, 2008.

[13] Yanwei Pang, Xuelong Li, and Yuan Yuan, "Robust tensor analysis with l1-norm," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 2, pp. 172–178, 2010.

[14] Jangwoo Park and Wonzoo Chung, "Common spatial patterns based on generalized norms," in *Brain-Computer Interface (BCI), 2013 International Winter Workshop on*. IEEE, 2013, pp. 39–42.

[15] Weifeng Liu, Puskal P Pokharel, and José C Príncipe, "Correntropy: Properties and applications in non-gaussian signal processing," *IEEE Transactions on Signal Processing*, vol. 55, no. 11, pp. 5286–5298, 2007.

[16] Sohan Seth and José C Príncipe, "Compressed signal reconstruction using the correntropy induced metric," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference On*. IEEE, 2008, pp. 3845–3848.

[17] Badong Chen, Xi Liu, Haiquan Zhao, and José C Príncipe, "Maximum correntropy kalman filter," *Automatica*, vol. 76, pp. 70–77, 2017.

[18] Badong Chen, Lei Xing, Haiquan Zhao, Nanning Zheng, and Jos C. PrNcipe, "Generalized correntropy for robust adaptive filtering," *IEEE Transactions on Signal Processing*, vol. 64, no. 13, pp. 3376–3387, 2015.

[19] Songlin Zhao, Badong Chen, and Jose C Principe, "Kernel adaptive filtering with maximum correntropy criterion," in *Neural Networks (IJCNN), The 2011 International Joint Conference on*. IEEE, 2011, pp. 2012–2017.

[20] Nachman Aronszajn, "Theory of reproducing kernels," *Transactions of the American mathematical society*, vol. 68, no. 3, pp. 337–404, 1950.

[21] Max Welling, "Fisher linear discriminant analysis," *Department of Computer Science, University of Toronto*, vol. 3, pp. 1–4, 2005.

[22] Muhammad Naeem, Clemens Brunner, Robert Leeb, Bernhard Graimann, and Gert Pfurtscheller, "Seperability of four-class motor imagery data using independent components analysis," *Journal of neural engineering*, vol. 3, no. 3, pp. 208, 2006.