Proceedings Paper:
Azab, A., Mihaylova, L., Arvaneh, M. et al. (2019) Robust common spatial pattern estimation using dynamic time warping to improve BCI systems. In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 12-17 May 2019, Brighton, UK. IEEE. ISSN: 0736-7791.

https://doi.org/10.1109/ICASSP.2019.8682689

ROUPST COMMON SPATIAL PATTERN (CSP) ESTIMATION USING DYNAMIC TIME WARPING TO IMPROVE BCI SYSTEMS

Ahmed M.Azab, Lyudmila Mihaylova, and Mahnaz Arvaneh

(Automatic Control and System Engineering, University of Sheffield)

ABSTRACT

Common spatial pattern (CSP) is one of the most popular feature extraction algorithms for brain-computer interfaces (BCI). However, CSP is known to be very sensitive to noise and prone to overfitting. This paper proposes a novel dynamic time warping (DTW) based approach to improve CSP covariance matrix estimation and hence improve features extracted. Dynamic time warping is a well-known technique to find an optimal alignment between two time-dependent sequences under certain conditions. The proposed approach align the available trials for each class to the mean signal of this class. The proposed DTW-based approach is applied to the SVM classifier and evaluated using one of the publicly available datasets. The results showed that the proposed approach when compared to normal CSP, improved the classification results from 78% to 83% on average. Importantly, some subjects gained improvement around 10%.

***Index Terms*—** Brain computer interface (BCI), Common spatial pattern (CSP), Dynamic time warping (DTW)

1. INTRODUCTION

Brain-computer interface (BCI) provides a direct communication between a person's brain and an electronic device without the need for any muscle control [1]. Electroencephalogram (EEG) is the most widely used brain signals in BCI since it is measured non-invasively with a high temporal resolution. In the EEG-based BCI system, User mental's state is decided by classifying the feature extracted from EEG. Common Spatial pattern is one of the most popular feature extraction methods for BCIs. The importance of spatial filtering arises due to the poor spatial resolution of EEG measurements. Thus the EEG pattern of interest is mixed with several irrelevant but concurrent neural activities. Using spatial filtering, signals from multiple electrodes are linearly combined to increase signal to noise ratio, leading to extract more discriminative EEG signals.

Despite being popular and effective, CSP is known to be very sensitive to noise and prone to overfitting. There are different reasons which can lead to poor CSP features.First, EEG signals are very easy to catch noisy during recording. So, it is very difficult to estimate the probability distributions for high dimensional noisy EEG signals where outliers will have a great negative effects. Second, EEG are highly non-stationary which may happen due to several factors such as: the variations of users' mental states, miss concentration and fatigue, which will also lead to inaccurate CSP features. So, poor performance is likely to happen when using the classifier trained on the features extracted from such EEG data.

To address CSP drawbacks, different algorithms have been proposed. Most of the proposed algorithms focus on improving common spatial patterns (CSP) through modification of either the covariance matrix estimation method or the CSP optimization objective function. Even to over come non-stationarity problem [2, 3, 4], or to overcome the problem of small training set [5]. Whereas, there are other techniques which tries to improve CSP by tuning the frequency band where CSP will be applied [6].

This paper proposes a novel DTW-based CSP approach to improve CSP covariance matrix estimation. In the proposed approach, the mean signal of the available trials for each class are calculated. To cope with the problem nos-stationarity of EEG signals, we hypothesize that alignment of EEG trials to a common signal might reduce the non-stationarties between these trials. Following this assumption, the available trials for each class are aligned to the calculated mean signal of this class to minimize the non-stationarity between the available EEG trials. At this stage the available trials from the same class get as close as possible to the mean of this class and also to each others. The new aligned trials are used to calculate the CSP covariance matrix. Based on our knowledge this the first time to use DTW with BCI and in a such way.

The proposed approach is evaluated using one of publicly available datasets with moderate number of subjects. Performance of the proposed approach is also compared with the results of the normal CSP state algorithm.

The remainder of this paper is organized as follows. In Section II, we will describe our proposed transfer learning approach. The experimental setup is shown in Section III. Evaluation results are discussed and analyzed in Section IV. Finally, conclusions are drawn in Section V.

Thanks to XYZ agency for funding.

2. METHODOLOGY

2.1. Common Spatial Patterns (CSP) Algorithm

CSP linearly transforms the data from the original EEG-channels into new channels to better differentiate between two conditions by maximizing the variance of one condition while minimising it for the other. The CSP filters are calculated based assigning a new weight for each channel depend on the projection matrix . This projection matrix will have as much filters as the number of channels and the columns of the matrix will carry the weights to make linear combinations of the original EEG channels to decide which EEG-channels carry the most useful information. The first half of the projection matrix will maximize the variance for class one and minimize it for class two, while the second half of the projection matrix will maximize the variance for class two and minimize it for class one under the assumption that the signal is band-pass filtered [7].

One of the major limitation of CSP filters, is the non-robust covariance matrices estimation problem due to its sensitivity to artifacts and noise, which will negatively affect the CSP filters. Based on the amount of features needed an amount of CSP-channels, also called filter pairs, are selected. When selecting a pair of filters, the outermost channels are chosen, which correspond to the highest and lowest eigenvalues by construction, and thus contain most information, the following equations will show how feature extraction based on CSP works.

Let consider, $X_i \epsilon n \times t$ is the i^{th} band passed signal trial and $Z \epsilon t \times n$ is the signal after spatial filtration with $W \epsilon n \times n$ projection matrix of CSP.

$$Z = X_i^T W \tag{1}$$

Here, for each trial n and t are the numbers of EEG channel and time points respectively. Let $C_1 \epsilon R_n \times n$and $C_2 \epsilon R_n \times n$ are covariance matrix of EEG signal X for the two classes. C_1 and C_2 can be computed by [8]:

$$C_{(c)} = \frac{1}{n_c} \sum_{i \epsilon I_c} X_i \times X_i^T \qquad c = [1, 2] \tag{2}$$

Here, all trials corresponding to class $c$are denoted by I_c , and the total number of trials for each class c is n_c. CSP filter matrix W can be computed by:

$$C_1 \times W = (C_1 + C_2) \times WD \tag{3}$$

where, eigenvalues for C_1 formed the D diagonal matrix. Generally, classification is done using m pairs of filters from W. In this report, the first three and last three rows of W are used to acquire spatial filtered signal $Z^* \epsilon t \times m$ [8].

$$Z^* = X^T W^* \tag{4}$$

Finally, If all EEG-data points would be used, the dimensionality of the data would be too high to be used by the classifier, so the most relevant features are extracted so the feature vector $F \epsilon R^{2m}$ can be computed by calculating logarithm of variance of Z [8].

$$F = log(var(Z^*)) \tag{5}$$

3) Support Vector Machines(SVM):it is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyper-plane which categorizes new examples. The decision boundary can be found by solving the following constrained optimization problem :

$$Min \quad \frac{1}{2}||W||^2 \tag{6}$$

The aim is set to estimate a transformation matrix W where the scattering of between class is maximized and within class scattering is minimized of.

$$subject \quad to \quad y_i(W^T x_i + b) \tag{7}$$

Where, b is the bias value, which can be calculated empirically, and the class label will be depending on the sign of f(x) will give.

2.2. Dynamic Time Warping (DTW)

Dynamic time warping (DTW) was initially proposed to solve the time deformation problem between a two patterns in speech recognition problems [9]. Subsequently, DTW has been applied to other problems such as object recognition, classification and clustering of time domain signals such as:EEG, ECG, subject identification, and motion analysis. For EEG, DTW is typically used as a measure of dissimilarity between two patterns after being optimally aligned. In this paper, DTW is used in an unconventional way as goal is not to find the DTW distance between two trials but to align a collection of measured trials from one class in the average of these class trials. Based on our knowledge this is the first time to use DTW in such a way and with motor imagery signals.

In order to develop the proposed CSP approach based DTW alignment, the general formulation of DTW algorithm and how it is used to generate a pair of aligned responses is described first.

Suppose we have two time series Q and C, of length k and v respectively, where:

$$Q = q_1, q_2, ..., q_i, ..., q_k \tag{8}$$

$$C = c_1, c_2, ..., c_j, ..., c_v \tag{9}$$

To align two sequences using DTW we construct an $n-by-m$ matrix where the (i_{th}, j_{th}) element of the matrix contains the distance $d(q_i, c_j)$ between the two points q_i and c_j Typically the Euclidean distance is used

$$d(qi, cj) = (qi - cj) \tag{10}$$

Each matrix element (i, j) corresponds to the alignment between the points q_i and c_j. A warping path W, is a contiguous set of matrix elements that defines a mapping between Q and C. The k_{th} element of W is defined as: $w_k = (i, j)_k$ so we have:

$$W = w_1, w_2, ..., w_k, ..., w_K \qquad max(m, n) \leq K < m+n-1 \tag{11}$$

The warping path is typically subject to several constraints.
Boundary conditions: $w_1 = (1, 1)$ and $w_K = (m, n)$.
Continuity: Given $w_k = (a, b)$ then $w_{k-1} = (a, b)$ where $aa' \leq 1$ and $b - b' \leq 1$.
Monotonicity: Given $w_k = (a, b)$ then $w_{k-1} = (a', b')$ where $aa' \leq 0$ and $b - b' \geq 0$.
There are exponentially many warping paths that satisfy the above conditions, however we are interested only in the path which minimizes the warping cost. This path can be found very efficiently using dynamic programming to evaluate the following recurrence which defines the cumulative distance $\gamma(i, j)$ as the distance $d(i, j)$ found in the current cell and the minimum of the cumulative distances of the adjacent elements:

$$\gamma(i, j) = d(q_i, c_j)+min\{\gamma(i-1, j-1), \gamma(i-1, j), \gamma(i, j-1)\} \tag{12}$$

2.3. The proposed MI alignment based DTW transfer learning method

In this paper, we assume that a large numbers of labeled trials EEG trials are available from each subjects. The set of labeled EEG trials for each subject can be presented as $d = (x^i, y^i)_{i=1}^n$, where x^i and y^i respectively denote the instances matrix and the class label, $y^i \in \{0, 1\}$, of the i^{th} trial, and n refers to the number of the trials. Each trial $\mathrm{x}^i \subset R^{h\times v}$ where h is the number of instances and v is the number of channels per trial. Typically, the classifier, is trained using the available subject-specific training features to predict the labels of the unlabeled trials. The commonly used subject-specific (SS) BCI model uses CSP algorithm to extract features. However, one of the major limitation of CSP filters, is the non-robust covariance matrices estimation problem due to its sensitivity to artifacts and noise which will negatively affect the CSP filters. To overcome this problem this paper proposes CSP algorithm using DTW. The proposed algorithm tries to align the available trials from each class to be as much similar to the average of the available trials of this class. Now, we have a big amount of trials that are greatly similar and can be used to estimate the CSP features instead of the originally scattered trials.

At first the average of the available trials per each class c of is computed as follows:

$$t_c = (1/n_c) \sum_{i=1}^{n_c} x^i, \tag{13}$$

where c refers to the class label, t_c refers to the average of the available trials of class c, and n_c is the number of trials available from class c. After that a similarity matrix between each available trial and the average signal from the same class is computed using (12). Then the warping path for these two trials is calculated in a way to minimize the following cost function under the previously mentioned constraints:

$$D(t_c, x^i) = min\big(1/k \sum_{i=1}^{k} d\{w_k\}\big), \tag{14}$$

where D is the accumulated distance between the average of class c and each individual trial from the same class. Then the indices of the warping path that minimize the previous cost function for each source domain trial are used to construct the new aligned trial as follows:

$$x^i(aligned) = \{x^i(w_1), ..., x^i(w_k)\} \tag{15}$$

where () is the indices of this trial instance that have the minimum with the related instances from the reference signal. Those reflected instances are the instances that will make this trial to be similar to the reference average signal. Subsequently the covariance matrix of the new aligned raw EEG trail is calculated as follows:

$$\Sigma_i(aligned) = \frac{(x^i(aligned))(x^i(aligned))^T}{trace((x^i(aligned))(x^i(aligned)))}. \tag{16}$$

Finally, the average of the calculated covariance matrices of the aligned trials for each class c is computed as follows:

$$\Sigma_c = (1/n_c) \sum_{i=1}^{n_c} \Sigma_i(aligned). \tag{17}$$

3. EXPERIMENTS

In order to validate the proposed algorithms and compare them with the baseline algorithms, all the algorithms are applied to data set 2a BCI Competition IV 2008 [10]. This data set consists of EEG data from 9 subjects performing 4 classes of motor imagery task. In this paper only data from right and left hand motor imagery are used. Two sessions on different days were recorded for each subject. Each session is comprised of 6 runs, each run consists of 12 trials for each class.

EEG signal was recorded using 22 electrodes. EEG signals were sampled at 250 Hz, and were bandpass-filtered between 0.5 Hz and 100 Hz. Moreover, a 50 Hz notch filter was applied to remove power line noise. The proposed algorithms and the baseline algorithms are applied only on the trials recorded on the second day by dividing it to two sessions one for training (consists of the first 42 trials recorded per class) and one for testing (consists of the last 30 trials recorded per class). This was done to establish a practical case that new subject data is coming from the same session. For the new subject, different training sizes were examined (i.e. 10, 20 and 42 trials per class). It is note that in each multitask learning algorithm, the train data of each new subject and the other 8 other subjects were used for calculating classification parameters.

This average covariance matrix will replace the covariance matrix used to calculate the common spatial filters.

4. 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] 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.

[3] Wojciech Samek, Carmen Vidaurre, Klaus-Robert Müller, and Motoaki Kawanabe, "Stationary common spatial patterns for brain–computer interfacing," *Journal of neural engineering*, vol. 9, no. 2, pp. 026013, 2012.

[4] Mahnaz Arvaneh, Cuntai Guan, Kai Keng Ang, and Chai Quek, "Optimizing spatial filters by minimizing within-class dissimilarities in electroencephalogram-based brain–computer interface," *IEEE transactions on neural networks and learning systems*, vol. 24, no. 4, pp. 610–619, 2013.

[5] Haiping Lu, How-Lung Eng, Cuntai Guan, Konstantinos N Plataniotis, and Anastasios N Venetsanopoulos, "Regularized common spatial pattern with aggregation for eeg classification in small-sample setting," *IEEE transactions on Biomedical Engineering*, vol. 57, no. 12, pp. 2936–2946, 2010.

[6] Kai Keng Ang, Zheng Yang Chin, Haihong Zhang, and Cuntai Guan, "Filter bank common spatial pattern (fbcsp) in brain-computer interface," in *Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*. IEEE, 2008, pp. 2390–2397.

[7] Gerwin Schalk, Dennis J McFarland, Thilo Hinterberger, Niels Birbaumer, and Jonathan R Wolpaw, "Bci2000: a general-purpose brain-computer interface (bci) system," *IEEE Transactions on biomedical engineering*, vol. 51, no. 6, pp. 1034–1043, 2004.

[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] Hiroaki Sakoe and Seibi Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE transactions on acoustics, speech, and signal processing*, vol. 26, no. 1, pp. 43–49, 1978.

[10] C Brunner, R Leeb, G Müller-Putz, A Schlögl, and G Pfurtscheller, "BCI competition 2008–Graz data set A," .