35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Approximation-based Common Principal Component for Feature Extraction in Multi-class Brain-Computer Interfaces

Tuan Hoang, Dat Tran and Xu Huang

Abstract— **Common Spatial Pattern (CSP) is a state-of-the-art method for feature extraction in Brain-Computer Interface (BCI) systems. However it is designed for 2-class BCI classification problems. Current extensions of this method to multiple classes based on subspace union and covariance matrix similarity do not provide a high performance. This paper presents a new approach to solving multi-class BCI classification problems by forming a subspace resembled from original subspaces and the proposed method for this approach is called Approximation-based Common Principal Component (ACPC). We perform experiments on Dataset 2a used in BCI Competition IV to evaluate the proposed method. This dataset was designed for motor imagery classification with 4 classes. Preliminary experiments show that the proposed ACPC feature extraction method when combining with Support Vector Machines outperforms CSP-based feature extraction methods on the experimental dataset.**

I. Introduction

Common Spatial Pattern (CSP) is one of state-of-the-art feature extraction methods in Brain Computer Interface (BCI) systems. It was originally proposed by Koles [5] to analyze abnormal components in clinic research and then successfully applied to 2-class BCI systems [2][8]. The idea of CSP is to map data of two classes onto the same dimension such that variance of one class is maximized while variance of the other one is minimized. Although CSP is very successful in 2-class BCI classification systems, applying it to multi-class BCI classification systems is still an open problem [11][4].

A current approach is to convert the multi-class classification problem to a set of 2-class classification problems. The two well-known methods are one versus the rest and combination of pairs of 2-class classification problems. These two methods have their own weakness. The first method assumes covariances of the rest classes are highly similar. However it is hard to observe this assumption in real-world applications. In the second strategy, it cannot guarantee that good common principal components of two particular classes are also good for other pairs of classes. This method can be viewed as forming common principal components for all classes by simply grouping common principal components of pairs of classes. Reduction techniques based on heuristics are applied to reduce number of dimensions in feature space. Consequently, these techniques cannot guarantee the above-mentioned idea of CSP.

T. Hoang, D. Tran and X. Huang are with Faculty of Education, Science, Technology and Mathematics, University of Canberra, ACT, Australia {Tuan.Hoang, Dat.Tran, Xu.Huang}@canberra.edu.au

From another different perspective, taking multi-class CSP methods under light of subspace method view as shown in an influential work [5][8], a subspace is formed for each class from the corresponding covariance matrix, then a union of these subspaces is performed to select a group of principal components based of some measure. We name this method Union-based Common Principal Components (UCPC) in this paper. However, the chosen principal components may have very little contribution from some data classes. To address this limitation, we propose a method that is called Approximation-based Common Principal Component Analysis (ACPC). In our method, after constructing subspaces derived from covariance matrices, we approximate a new subspace that resembles these subspaces and has the same number of dimensions. Principal angle between these subspaces is used as the metric for the subspace approximation. The idea of forming an approximate subspace from these subspaces is based on the work of Krzanowski [6] when dealing with problem of heterogeneous covariance matrices. Extended works such as of Fujioka et al. [3] and Rothman et al. [9] are applied to analyzing data with heterogeneous covariance matrices. Our proposed work is different from those, we focus on multi-class problems to derive the resembled subspace for feature extraction in multi-class BCI systems.

The remaining of the paper is organized as follows. In Section 2, we present theoretical foundation of Approximation-based Common Principal Component method. In Section 3, we describe model of using common principal components for feature extraction in BCI systems. Experimental protocols as well as methods for classification and validation are introduced in Section 4. Section 5 presents our results and related discussions. Finally, we present our conclusion and future work in Section 6.

II. Approximation-based Common Principal Components

Given a set of k symmetrical real matrices $C_1, C_2, \ldots, C_k$ size of $n \times n$. Instead of jointly diagonalizing the set of matrices as Union-based Common Principal Components, we diagonalize these k matrices separately resulting in set of eigenvectors V_i and eigenvalues λ_i satisfying

$$C_i = V_i \lambda_i V_i^T \tag{1}$$

for all matrix C_i with $i \in [1, k]$. In which, V_i is a $n \times n$ matrix whose rows representing principal components for the corresponding coordinate. It is easy to see that these problems are identical to conducting k principal component

978-1-4577-0216-7/13/$26.00 ©2013 IEEE

analysis (PCA) seperately on k matrices C_i. According to theory of principal component analysis, when mapping data on to new coordinates, first a few of principal components are enough for representing variance of original data. Let p be the number of selected principal components forming new subspaces for all V_i. Let L_i sized $p \times n$ be the matrix reprensenting p principal components taken from V_i. We then try to find a new subspace H which resembles all subspaces spanned by the set of eigenvectors L_i. We follow approach proposed by Krzanowski [6] using sum of principal angles between new subspace H and other original subspaces.

Let h be an arbitrary vector in the original n-dimensional data space. Let $\theta_i \in [0, \frac{\pi}{2}]$ be the angle between vector h and the vector most nearly parallel to it in the space spanned by L_i. We define the Δ function as follows

$$\Delta = \sum_{i=1}^{k} cos^2\theta_i \tag{2}$$

Theorem 1. Let h be an arbitrary vector in the original n-dimensional data space. Let θ_i be the angle between vector h and the vector most nearly parallel to it in the space spanned by L_i. Then we have

$$cos\theta_i = \frac{\sqrt{h^T L_i^T L_i h}}{\|h\|} \tag{3}$$

Proof: We can see that by the definition, the angle θ_i is the angle between vector h and its project on the subspace spanned by L_i. Let p be the project of h on subspace spanned by L_i. Because L_i is the basic of the subspace we can then rewrite vector p as

$$p = L_i^T x \tag{4}$$

The projected vector of h on the subspace perpendicular to subspace spanned by L_i is $h - L_i^T x$. Therefore,

$$\begin{aligned} L_i(h - L_i^T x) &= 0 & (5)\\ L_i h - L_i L_i^T x &= 0 & (6)\\ x &= (L_i L_i^T)^{-1} L_i h & (7) \end{aligned}$$

Substitute (7) to (4), we have

$$p = L_i^T (L_i L_i^T)^{-1} L_i h \tag{8}$$

The orthogonality of L_i gives us

$$p = L_i^T L_i h \tag{9}$$

Cosine of angle between two vectors by definition is

$$cos(\theta_i) = \frac{h^T L_i^T L_i h}{\|h^T\| \|L_i^T L_i h\|} \tag{10}$$

Rewriting norm of vector in dot product operator, we have

$$\begin{aligned} \|L_i^T L_i h\| &= \sqrt{(L_i^T L_i h)^T (L_i^T L_i h)} & (11)\\ &= \sqrt{h^T L_i^T L_i L_i^T L_i h} & (12)\\ &= \sqrt{h^T L_i^T L_i h} & (13) \end{aligned}$$

Substitute (13) to (10), we have (3). The theorem is proven.

Theorem 2. Let h be an arbitrary vector in the original n-dimensional data space. Let θ_i be the angle between vector h and the vector most nearly parallel to it in the space spanned by L_i. Then the value of h is given by the eigenvector h_1 corresponding to the largest eigenvalue λ_1 of the matrix $L = \sum L_i^T L_i$ will maximize the value of Δ.

Proof: According to Theorem 1, we have

$$\begin{aligned} \Delta &= \sum_{i=1}^{k} cos^2\theta_i = \sum \frac{h^T L_i^T L_i h}{\|h\|^2} & (14)\\ &= \frac{h^T \sum_{i=1}^{k} (L_i^T L_i) h}{h^T h} & (15)\\ &= \frac{h^T L h}{h^T h} & (16) \end{aligned}$$

We then convert the original optimal problem of finding arbitrary vector h that $max_h \frac{h^T L h}{h^T h}$ to the simpler optimal problem of finding normal vector h that $max_{\|h\|=1} h^T L h$. Let V be the matrix used for diagonalizing matrix L as follows

$$L = V D V^T \tag{17}$$

in which

$$D = diag(\lambda_1, \lambda_2, , \lambda_n) \tag{18}$$

and

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \tag{19}$$

We will prove that $sup_{\|h\|=1} h^T L \leq \lambda_1$. Let $y = V^T h$. Since $\|h\| = 1$, we have $\|y\| = 1$. Then $max_{\|h\|=1} h^T L h = max_{\|y\|=1} y^T D y = max_{\|y\|=1} \sum_{i=1}^{n} \lambda_i \|y_i\|^2 \leq max_{\|y\|=1} \sum_{i=1}^{n} \lambda_1 \|y_i\|^2 = \lambda_1$. The equality happens only when vector h is h_1 which is an eigenvector of unit norm associated with the largest eigenvalue λ_1. The theorem is proven.

Applying the Graham-Smith method, it is seen that the eigenvector h_2 associated to the second largest eigenvalue of matrix L will be orthogonal to vector h_1 and leads to the second largest value of Δ. Similarly, we form the remaining n vectors h_i which will span the new subspace H. Moreover, we can see that finding new subspace H is actually to conduct another Principal Component Analysis on the matrix L. Therefore we can then select q components from n components of subspace H which is enough to represent all data points.

The following is pseudo code of the algorithm in finding common principal components by using two times of principal component analysis.

Algorithm 1: Finding approximate-based common principal components.

Data: A set of k real symmetrical matrices C_i of size n

Result: Common principal components V and their accompanied eigenvalues E

Step 1: For each matrix C_i, do Principal Component Analysis on $C_i = V_i \lambda_i V_i^T$

Step 2: For each V_i, select k_i components from n components of V_i whose sum of eigenvalues exceeds 90% of total eigenvalue.

Step 3: Set k = $\max k_i$

Step 4: Set $L = \sum_i^n (L_i^T L_i)$

Step 5: Do Principal Component Analysis on $L = H\lambda H^T$

Step 6: Select q components from n components of H

Step 7: Return q selected components of H and their corresponding eigenvalues.

III. Feature extraction with Common Principal Component analysis in BCI systems

In BCI systems, data is acquired by multi-channel devices. Let $X_i = x^p(t)$ be the i^{th} trial of the dataset. All trials consist of n channels $x^p(t) = (x_1^p, x_2^p, \ldots, x_n^p)$ where p is the channel index, t is the time index of the signal, and T is its length or number of samples. Each trial belongs to a class $L(X_i)$ of mental action or motor imagery. Assume that there are k classes in the dataset $L(X_i) \in [1, k]$. Let $Cov(X_i) = X_i X_i^T$ be the covariance matrix of trial X_i and C_i be the estimate covariance matrix of class i^{th} . We use empirical method to estimate these covariance matrices.

$$C_i = \frac{1}{|X_q : L(X_q) = i|} \sum_{X_q : L(X_q)=i} Cov(X_q) \quad (20)$$

$$= \sum_{X_q : L(X_q)=i} X_q X_q^T \quad (21)$$

From these covariance matrices, we derive common principal components V of data by applying the above-described algorithm $ACPC$. Original data will then be mapped on new subspace as shown in Equation (22).

$$X_i^{CPC} = V^T X_i \quad (22)$$

$$\text{Feature vector} = (log(var(X_i^{CPC}))) \quad (23)$$

Due to orthogonal property of new subspace, data on common principal components are de-correlated. Moreover, eigenvalues represent variance degree of corresponding principal components. Feature vector of a trial as shown in Equation (23) is formed by combining variances of all channels from mapped data. To remove nonlinear property of variances, logarithm function is then applied. So the feature vector of a trial X_i is $log(var(X_i^{CPC}))$ and its size is q where q is number of selected components and is independent of length of the trial. This property is useful in allowing us to flexibly determine length of trials, especially in real time or online BCI systems.

IV. Experimental methods and validations

Our main purpose is to compare our proposed method $ACPC$ with two popular methods in applying 2-class Common Spatial Pattern on multi-class BCI systems. They are one-versus-the-rest ($1vsN\_CSP$) and pair-wise ($pair\_CSP$) methods. The Dataset 2a from BCI Competition IV [1] which is a well-known dataset for multi-class BCI systems is chosen for conducting experiments. The dataset was acquired by Graz University of Technology, Austria using Electroencephalography (EEG) technology with 22 channels at sampling frequency 250Hz. Nine subjects were asked to perform 4 classes of motor imagery tasks to move cursor left, right, down or up corresponding with imagination of movement of the left hand, right hand, both feet and tongue. There are 576 trials in total in both training and testing sets of the competition. For each trial, there are 2 seconds to help participants prepare themselves. After that, there is a cue appearing and staying on screen in 1.25s. The subjects were asked to carry out motor imagery tasks until 6^th second.

We segmented data into lengths of 2 seconds from the time point 2.5 second. We moved segment window by half of second timeframe. All these segments were bandpass filtered with frequency cut-off at 8Hz and 30Hz before were extracted features as in Equation (23). Support Vector Machine (SVM) and its popular kernel function RBF $K(x, x') = e^{-\gamma \|x - x'\|^2}$, a state of the art method for classifying in BCI systems [7], was chosen to classify data. We applied grid search to get the optimal classifiers. The parameter γ was searched in range $2^k : k = -10, -9, \ldots, 19, 20$. The trade-off parameter C was searched over the grid $2^k : k = 0, 1, \ldots, 12, 13$.

To evaluate accuracy of classification on the dataset, we divided it into training and testing data sets by ratio 8:2. The test data was normalized based on distribution parameters extracted from the training data set. We performed a 5-fold cross validation test on the training data to find the optimal parameters γ and C. These optimal parameters were then used to build classifiers for the entire training dataset. Finally, the classifiers were applied on the testing dataset to get accuracy results. To reduce randomness due to division of data into training and testing data, we ran this process five times. The reported accuracy results were calculated by taking average of accuracies of five times running this process.

V. Results and Discussion

We conducted two experiments. In the first experiment Our purpose is to compare our proposed method $ACPC$ with $1vsN\_CSP$ and $pair\_CSP$ methods on Dataset 2a of BCI Competition IV. In the second experiment, we analyzed effect of chosen number of components on classification accuracy.

A. Comparison with $1vsN\_CSP$ and $pair\_CSP$ methods

In this experiment, we select all components which is 22. For other two methods, as in other work, we select two components for each class for each CSP problem. Therefore, for one-versus-the-rest CSP, its feature vector has size $4 \times 2 \times n = 8 \times n$ while for the pair-wise CSP, its feature vector has size $6 \times 4 \times n = 24 \times n$ (6 is the number of CSP problems in this pair-wise strategy.) Table 1 shows results of classification.

It can be seen that our $ACPC$ method outperforms others significantly in eight of nine subjects in classification accuracy. The only exception is subject 7 where $1vsN\_CSP$ and $pair\_CSP$ are better than $ACPC$. Another finding is that results of $1vsN\_CSP$ and $pair\_CSP$ methods are highly similar. Figure (1) shows this clearly. The light green line and the red line are nearly identical. The reason can be both methods use 2-class CSP as their cores. Therefore,

TABLE I

COMPARISON OF $ACPC$ WITH $1vsN\_CSP$ AND $pair\_CSP$ METHODS (IN PERCENTAGE). BOLD NUMBERS ARE THE BEST RESULT OF SUBJECTS.

Method	S1	S2	S3	S4	S5	S6	S7	S8	S9
$ACPC$	**84**	**73**	**82**	**67**	**65**	**65**	74	**80**	**75**
$1vsN\_CSP$	73	48	77	50	36	40	75	76	69
$pair\_CSP$	71	46	78	46	36	40	**77**	76	70

differences between them if any is very small in multi-class BCI systems.

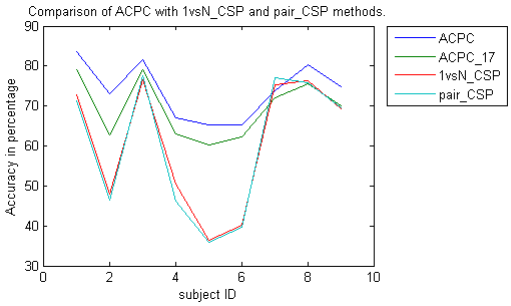

Fig. 1. Comparison of $ACPC$ with $1vsN\_CSP$ and $pair\_CSP$ methods. $ACPC$: all components. $ACPC\_80$: 80% total number of components.

B. Effect of number of selected common principal components on classification accuracy

We used the same dataset as in the first experiment for analysis. However instead of selecting all components, we varied the number of selected components. It can help us to see effect of number of selected components on classification accuracy. Basically, the more is number of selected components the more information do we have. However, the more is number of selected components the more time do we need for training and testing the classifiers. The percentage of sum of eigenvalues varies from 20% to 100% with step of 20%. Figure (2) shows classification accuracy of the experiments.

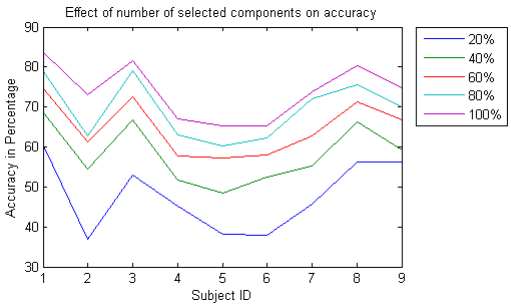

Fig. 2. Effect of number of selected components on classification accuracy.

The result shows that the more number of selected components there are, the more accurate the classification is. Moreover, our proposed method $ACPC$ still outperforms others with only 80% total number components, which is 17 components, selected in most subjects as shown in Figure (1).

VI. CONCLUSIONS AND FUTURE WORK

We have proposed Approximation-based Common Principal Component ($ACPC$) as a new method for direct feature extraction in multi-class BCI systems. It directly targets to the multi-class BCI problem instead of converting it to multiple 2-class BCI problems. Comparing with current approach of converting a multi-class BCI problem to multiple 2-class BCI problems by employing two popular strategies One-versus-The-Rest and pairs of 2-class BCI problems, our proposed method can provide a theoretical framework. We have also conducted experiments to compare $ACPC$ with two state-of-the-art methods for feature extraction which are One-versus-The-Rest CSP ($1vsN\_CSP$) and pair-wise CSP ($pair\_CSP$). Experimental results on the Datset 2a of BCI Competition IV show that our proposed method outperforms these two others even with 80% number of components selected. We also found that there is a relatively small difference between two strategies of converting a multi-class BCI problem into multiple 2-class BCI problems in classification accuracy on the same dataset.

Future work includes exploring error bound of $ACPC$ when resembling original subspaces and conducting more experiments on other well-known multi-class datasets.

REFERENCES

[1] BCI Competition IV website, URL: http://www.bbci.de/competition/iv/.

[2] Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., and Muller, K. R.: Optimizing spatial filters for robust eeg single-trial analysis, IEEE Signal Proc. Magazine, vol. 25, np. 1, pp. 41-56, 2008.

[3] Fujioka, T.: An approximate test for common principal component subspaces in two groups, Ann. Inst. Statist. Math., vol. 45, no. 1, pp. 147-158, 1993.

[4] Grosse-Wentrup, M., and Buss, M.: Multi-class common spatial patterns and information theoretic feature extraction, IEEE Trans. Biomedical Eng 01/2008, vol. 55, pp. 1991-2000, 2008.

[5] Koles, Z. J.: The quantitative extraction and topo-graphic mapping of the abnormal omponents in the clinical EEG, Electroencephalogr. Clin. Neurophysiol., vol. 79, no. 6, pp. 440-447, 1991.

[6] Krzanowsli, W. J.: Between-Groups comparisons of principal components, Journal of the American Statistical Association, vol. 74, no. 367, pp. 703-707, 1979.

[7] Lotte, F., Congedo, M., Lcuyer, A., Lamarche, F., and Arnaldi, B.: A review of classification algorithms for EEG-based brain-computer interfaces, Journal of Neural Engineering, vol. 4, no. 2, pp. 1-13, 2007.

[8] Ramoser, H., Muller-Gerking, J., and Pfurtscheller, G.: Optimal spatial filtering of single trial EEG during imagined hand movement, IEEE Transactions on Rehabilitation Engineering, vol. 8, no. 4, pp. 441-446, 2000.

[9] Rothman, J. A., Levina, E., and Zhu, J.: Generalized thresholding of large covariance matrices, Journal of the American Statistical Association, vol. 104, no. 485, pp. 177-186, 2009.

[10] Schlogl, A., Kronegg, J., Huggins, J. E., and Mason, S. G.: Evaluation criteria in BCI research, In: Dornhege, G., Millan, J. del R., Hinterberger, T., McFarland, D. J., Muller, K.-R. (Eds.), Toward brain-computer interfacing, MIT Press, 327-342, 2007.

[11] Wei, Q., Ma, Y., and Chen, K.: Application of quadratic Optimization to Multi-class Common Spatial Pattern Algorithm in Brain-computer Interfaces, In proceedings of The 3rd International Conference on Biomedical Engineering and Informatics, 2010.