One-Versus-the-Rest(OVR) Algorithm: An Extension of Common Spatial Patterns(CSP) Algorithm to Multi-class Case

Article *in* Conference proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference · February 2005

DOI: 10.1109/IEMBS.2005.1616947 · Source: PubMed

Proceedings of the 2005 IEEE
Engineering in Medicine and Biology 27th Annual Conference
Shanghai, China, September 1-4, 2005

One-Versus-the-Rest (OVR) Algorithm: An Extension of Common Spatial Patterns (CSP) Algorithm to Multi-class Case

Wei Wu, Xiaorong Gao, Shangkai Gao, *Senior Member*, *IEEE*
Department of Biomedical Engineering, Tsinghua University, Beijing, 100084, P. R. China

Abstract—**Extraction of relevant features that capture the invariant characteristics specific to each brain state is very important in order to implement a suitable Brain-Computer Interface (BCI) system. This paper presents an algorithm called One-Versus-the-Rest (OVR), which is an extension of a well-known method called Common Spatial Patterns (CSP) to multi-class case, to extract signal components specific to one condition from electroencephalography (EEG) data sets of multiple conditions. The algorithm was previously mentioned in [7], yet without an elaborate description. In this paper, detailed mathematical derivation of the algorithm is given, followed by a computer simulation. The computer simulation suggests that the algorithm is capable of reconstructing the actual specific part of each condition with high quality, even when the data are contaminated with considerable noise. We also hint future possible applications of the algorithm in the context of BCI at the end of the paper.**

Index terms—**Brain-computer interface (BCI), Common Spatial Patterns (CSP), multi-class, One-Versus-the-Rest (OVR)**

I. Introduction

A Brain-Computer Interface (BCI) transforms signals from the brain into control signals for conveying messages and commands to the external world [1]. By using that signal to, e.g., control a wheelchair or a neuroprosthesis, a BCI could become a valuable tool for paralyzed patients. Thus, an efficient discrimination of different brain states is important in order to implement a suitable system for human subjects. A well-known algorithm called Common Spatial Patterns (CSP) was proposed for this purpose [2][3]. The method of CSP was first suggested for use in binary case. Based on Principal Component Analysis (PCA) analysis, this approach can decompose the raw EEG signals into spatial patterns extracted from two classes of single trial EEG. An example can be found in motor imagery [4]. In motor imagery, a subject may be asked to imagine performing movements of the left or right hand. By using CSP, signal components in EEG specific to either left or right motor imagery can be extracted.

The use of BCIs is limited inherently by low information transfer rates in binary case. One possible way to enhance the speed of information transfer is to move from a binary decision to a choice among more options. But an increase in information transfer rates counts on the reliability of classification, as can be seen from the definition of the information transfer rate introduced by Shannon [5]. The percentage of correct classifications, in turn, is contingent on the extraction of features specific to each class.

The first multi-class extension of CSP was proposed in [6], which is based on pairwise classification and voting. Note, however, CSP in this approach is still in its binary version. In [7], two new algorithms for multi-class CSP were presented: One Versus the Rest CSP (OVR) and Simultaneous Diagonalization (SIM). The basic idea of SIM is the same as in binary CSP. The algorithm finds a simultaneous diagonalization of covariance matrices, each belongs to a certain class. The diagonal elements of the resultant diagonal matrices should sum to one. Although the idea is quite simple, simultaneous diagonalization can only be approximated for class number larger than 2. The performance of SIM therefore depends on which approximate simultaneous diagonalization algorithm is chosen. In OVR, spatial patterns are computed for each class against all others. Nevertheless, the authors in [7] didn't explain explicitly how to compare each class to to all others.

In this paper, a detailed derivation of the OVR algorithm for multi-class CSP is presented with a rigorous mathematical proof. Subsequently, a simulation study is given to demonstrate the validity of the algorithm. Finally, future possible applications of the algorithm in the context of BCI are introduced.

II. Methodology

In binary CSP, an optimal decomposition is found to transform two groups of measurements into a space that is common to them. In this common space, the two groups of transformed data have the same principal components, and their corresponding eigenvalues add up to one. Because these principal components are optimal in the proportion of variance they can account for in the common space, they could serve to maximally distinguish between the two groups.

In multi-class case, we will show, however, different optimal decompositions should be applied to different classes to extract their own specific signal components.

Without loss of generality, we will derive OVR algorithm for the three-class case. Extension to more classes is trivial, with the underlying theory being the same.

Denote raw common average referenced EEG signal matrices under three conditions as X_a, X_b, and X_c with dimension of N by T, where N is the number of channels and T is the number of samples in time. The spatial covariance of the EEG for these conditions can therefore be estimated by:

$$R_a=X_aX_a^T, \qquad R_b=X_bX_b^T, \qquad R_c=X_cX_c^T \tag{1}$$

where X^T denotes transpose of X.

As in binary CSP, we could still build the composite covariance matrix as:

$$R=R_a+R_b+R_c \tag{2}$$

The composite covariance matrix can be factored by eigendecomposition as:

$$R=U_0\wedge U_0^T \tag{3}$$

where U_0 is the $N\times N$ unitary matrix of principal components, and $\wedge$ is the $N\times N$ diagonal matrix of eigenvalues. The whitening transformation matrix is then formed as:

$$W=\wedge^{-1/2} U_0^T \tag{4}$$

where we take the components with non-zero eigenvalues into consideration.

To see how to extract common spatial patterns specific to condition a, let $R_a^{'}=R_b+R_c$. Then R_a and $R_a^{'}$ are individually transformed as:

$$S_a=WR_aW^T \tag{5}$$

and

$$S_a^{'}=WR_a{'}W^T \tag{6}$$

It can be demonstrated [8] that S_a and $S_a^{'}$ share common principal components, and that the sum of the corresponding eigenvalues for the two matrices will always be one. Consequently, if the eigendecomposition of S_a is:

$$S_a=U\wedge_aU^T \tag{7}$$

then $S_a^{'}$ can be factored as:

$$S_a^{'}=U\wedge_a^{'}U^T \tag{8}$$

and

$$\wedge_a+\wedge_a^{'}=I \tag{9}$$

where U is the matrix of common principal components.

Combine (5), (6), (7), and (8), we may write $\wedge_a$ and $\wedge_a^{'}$ as:

$$\wedge_a=(W^TU)^TR_a(W^TU)=SF_aR_aSF_a^T \tag{10}$$

and

$$\wedge_a^{'}=(W^TU)^TR_a{'}(W^TU)=SF_aR_a^{'}SF_a^T \tag{11}$$

where

$$SF_a=U^TW \tag{12}$$

and it can be seen as a spatial filter.

Because of (9), in the space spanned by U, the variance accounted for by the m common principal components corresponding to the m largest eigenvalues in $\wedge_a$ will be maximal for condition a. However, in order for these common principal components to be optimal in the least square sense for discriminating between condition a and all other conditions, the eigenvalues in $\wedge_a^{'}$ should represent variance accounted for by all the other conditions on these common principal components. This will ensure that when the variance of the signal components for condition a is maximized, the variance of the signal components for all other conditions is minimized. To see this, in light of $R_a^{'}=R_b+R_c$, we may rewrite (11) as:

$$SF_aR_bSF_a^T+ SF_aR_cSF_a^T=\wedge_a^{'} \tag{13}$$

we thus have:

$$\wedge_b+\wedge_c=\wedge_a^{'} \tag{14}$$

where $\wedge_b=SF_aR_bSF_a^T$ and $\wedge_c=SF_aR_cSF_a^T$

Clearly, the diagonal elements of $\wedge_b$ and $\wedge_c$ represent respectively the variance accounted for by condition b and condition c on the common principal components, although S_b and S_c are not diagonal matrices themselves. Therefore, (14) implies that the diagonal elements of $\wedge_a^{'}$ represent variance accounted for by both condition b and condition c on the common principal components.

For the above reason, we may thus be able to extract signal components that are specific to condition a based on the diagonal elements of $\wedge_a$. When a diagonal element is close to 1, we choose its corresponding common principal component as belongs to condition a.

With

$$Z_a=SF_aX_a \tag{15}$$

The decomposition of X_a can be written as:

$$X_a=SP_aZ_a \tag{16}$$

where SP_a is the pseudoinverse of SF_a. SP_a can be seen as the matrix of common spatial patterns.

The covariance matrix of Z_a is:

$$R_{Za}=Z_aZ_a^T=SF_aX_aX_a^TSF_a^T=SF_aR_aSF_a^T=\wedge_a \tag{17}$$

Therefore, X_a is projected into the space of common spatial patterns as orthogonal components. Z_a can be seen as a new time series that is equivalent to X_a in the space of common spatial patterns.

After we have chosen common spatial patterns specific to condition a, we could use their corresponding rows in Z_a as feature vectors for X_a. Specifically, denote the matrix of the selected feature vectors as Z_a^s. Z_a^s can be written as:

$$Z_a^s=SF_a^sX_a \tag{18}$$

where SF_a^s consists of the rows in SF_a corresponding to Z_a^s.

Furthermore, we could obtain signal components X_a^s specific to condition a by back-projecting Z_a^s to the original signal space:

$$X_a^s=SP_a^sZ_a^s=SP_a^sSF_a^sX_a \tag{19}$$

where SP_a^s consists of the columns in SP_a corresponding to Z_a^s.

Repeat the above procedures, we could also find feature vectors Z_b^s and Z_c^s for condition b and c.

In the above algorithm, we compute spatial patterns for each class against all others, hence its name "One-Versus-the-Rest (OVR)".

Assume there are n conditions, the OVR algorithm can be summarized as follows:

1) Estimate covariance matrices R_1, R_2, $\cdots$, R_n for all conditions using (1) and then build the composite covariance matrix R using (2).
2) Construct whitening transformation matrix W according to (4).
3) i=1.
4) Transform covariance matrix R_i into S_i by (5) and factorize S_i according to (7).
5) Select m principal components of S_i as specific to condition i. Denote them as U_i^s.
6) The specific spatial filter SF_i^s for condition i is

then constructed as

$$SF_i^s = (U_i^s)^T W \tag{20}$$

The specific spatial patterns SP_i^s for condition i is computed as the pseudoinverse of SF_i^s.

7) Estimate the signal components specific to condition i using (19).
8) $i=i+1$.
9) Repeat the steps 4)-8) until $i>n$, i.e., until specific signal components for all conditions have been found.

III. Results

The following spatio-temporal source model was adopted to conduct the computer simulation.

$$X_a = [C_1\ C_6]\begin{bmatrix} S_1 \\ S_6 \end{bmatrix}, \qquad X_b = [C_2\ C_5\ C_6]\begin{bmatrix} S_2 \\ S_5 \\ S_6 \end{bmatrix},$$

$$X_c = [C_3\ C_5\ C_6]\begin{bmatrix} S_3 \\ S_5 \\ S_6 \end{bmatrix}, \quad X_d = [C_4\ C_5\ C_6]\begin{bmatrix} S_4 \\ S_5 \\ S_6 \end{bmatrix} \tag{21}$$

where X_a, X_b, X_c, and X_d are simulated measurements from 40 channels under four different conditions, with 500 time points each. C_i ($i=1,2,\cdots,6$) are spatial patterns with dimension of 40×1, in which C_1, C_2, C_3, and C_4 are specific to condition a, b, c, and d, respectively. S_i ($i=1,2,\cdots,6$) are source waveforms with dimension of 1×500. We could see the actual specific parts in X_a, X_b, X_c, and X_d are C_1S_1, C_2S_2, C_3S_3, and C_4S_4, respectively.

In our simulation, the spatial patterns were random Gaussian white noises generated independently. The source waveforms are described as follows:

S_1 is the real part of a linear frequency-modulated signal, whose initial and final normalized frequency is 0 and 0.3.

S_2 is a real sinusoidal signal, whose normalized frequency is 0.06.

S_3 is simulated as the membrane potential of a neuron, generated by the famous Hindermarsh-Rose (HR) neuronal model in accordance to the parameters used in [9].

S_4 is the sum of three real sinusoidal signals, with normalized frequencies being 0.1, 0.2, and 0.28. The ratio of the three components' amplitudes is 1:2:7.

S_5 is the sign function of S_6.

S_6 is a randomly generated Gaussian white noise.

These source waveforms are presented in Fig. 1.

Because there exists noise in actual EEG measurement, Gaussian white noise was added as sensor measurement noise to the model (21) to simulate a more realistic situation. In this simulation we increased the SNR from 30 dB to 0 dB in 5 dB decrements. The SNR was defined as:

$$10\log\left(\frac{\mathrm{var}(X_{exact})}{\sigma^2}\right) \tag{22}$$

where $\mathrm{var}(X_{exact})$ is the variance of the simulated noise-free measurements, and σ^2 is the variance of the added noise, which was independently and identically distributed on each channel.

The resultant X_a, X_b, X_c, and X_d were then subjected to OVR analysis. The OVR-extracted specific part for each condition was calculated by (19). After that, we calculated the correlation coefficients between the actual specific part and the OVR-extracted specific part for each condition. For each condition, the correlation coefficient was defined as:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{< X_i^s, \widehat{X_i^s} >}{\| X_i^s \| \cdot \| \widehat{X_i^s} \|} \tag{23}$$

where X_i^s and $\widehat{X_i^s}$ are the actual and OVR-extracted specific parts for the condition in the i th channel, respectively. $<a,b>$ denotes the inner product of vector a and vector b and $\|a\|$ denotes the Euclidean norm of vector a. N is the channel number, which equals 40 in our simulation.

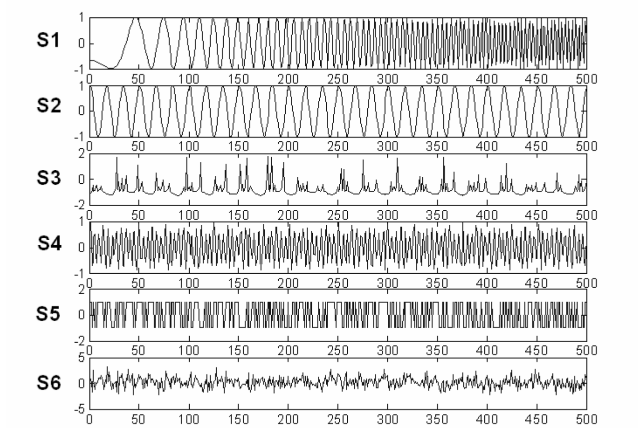

Fig. 1. Waveforms of six sources. S_1, S_2, S_3, and S_4 correspond to the specific spatial patterns of the four conditions. S_5 and S_6 are simulated as noisy sources.

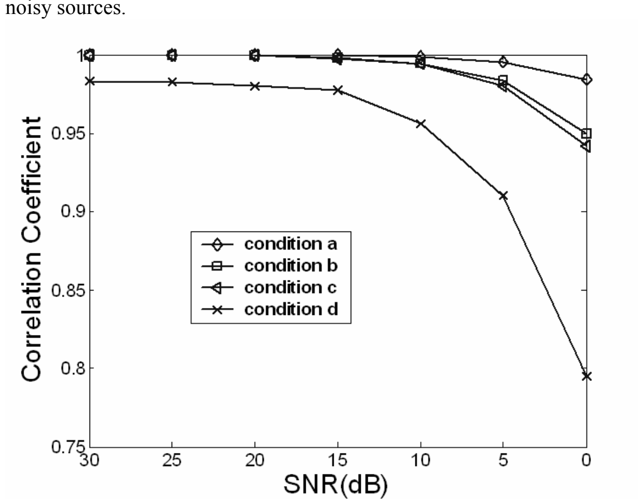

Fig. 2. Effects of noise on OVR for four conditions. Four curves correspond to four conditions. The horizontal axis gives the SNR, which is decreased from 30 dB to 0 dB in 5 dB decrements. The vertical axis represents the correlation coefficient between OVR-extracted specific part and the actual specific part.

The effect of noise on OVR is presented in Fig. 2. When the SNR is high (SNR>10 dB), OVR-extracted specific part is very similar to the actual specific parts, with correlation coefficients greater than 0.95 for all conditions. As the SNR decreases, the result becomes worse. However, even when the SNR deteriorates to 0dB, the correlation coefficients are still quite high.

IV. Discussion

In this paper, we have explained in detail OVR algorithm that can extract signal components specific to one condition from EEG datasets of multiple task conditions. The algorithm is a generalization from binary CSP to multi-class problems. Our computer simulation has demonstrated the validity and effectiveness of the method. The robustness of this method to noise shows it is an accurate tool for reconstructing the actual specific part of a condition.

As binary CSP can be used in binary BCI systems, OVR algorithm can be applied to multi-class BCI systems. For example, the features we extracted can be based on event-related desynchronization (ERD) [10]. ERD is an event-related physiological phenomenon presenting as changes in EEG power over certain frequency bands. It is commonly pronounced in the contralateral hemisphere. Motor imagery is typically accompanied by ERD in mu and beta rhythms [11]. This observation leads us to find spatial structures of ERD specific to each kind of motor imagery by using OVR algorithm in future applications of multi-class BCI system.

Acknowledgement

This work was supported by National Natural Science Foundation of China (#60318001) and the Key Project of Chinese Ministry of Education (NO.1041185). The authors would like to thank colleagues in the lab for helpful discussions.

References

[1] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T.M. Vaughan, “Brain-computer interfaces for communication and control,” *Electroencephalogr. Clin. Neurophysiol.*, vol. 113, pp. 767-791, 2002.

[2] Z. J. Koles, “The quantitative extraction and topographic mapping of the abnormal components in the clinical EEG,” *Electroenceph. Clin. Neurophysiol.*, vol. 79, pp. 440-447, 1991.

[3] Z. J. Koles, M. S. Lazar, and Z. Zhou, “Spatial patterns underlying population differences in the background EEG,” *Brain Topogr.*, vol. 2, pp. 275-284, 1990.

[4] H. Ramoser, J. Müller-Gerking, and G. Pfurtscheller, “Optimal spatial filtering of single trial EEG during imagined hand movement,” *IEEE Trans. Rehab. Eng.*, vol. 8, pp. 441-446, 2000.

[5] C. E. Shannon and W. Weaver, *Mathematical theory of communication*, Champaign, IL: Univ. Illinois Press, 1964.

[6] J. Müller-Gerking, G. Pfurtscheller, and H. Flyvbjerg, “Designing optimal spatial filters for single-trial EEG classification in a movement task,” *Clin. Neurophysiol.*, vol. 110, pp. 787-798, 1999.

[7] G. Dornhege, B. Blankertz, G. Curio, and Klaus-Robert Müller, “Boosting bit rates in noninvasive EEG single-trial classifications by feature combination and multiclass paradigms,” *IEEE Trans. Biomed. Eng.*, vol.51, pp. 993-1002, 2004.

[8] K. Fukunaga, *Introduction to statistical pattern recognition*, second ed., Boston: Academic Press, 1990.

[9] Hindmarsh JL and Rose RM, A model of neuronal bursting using three coupled first order differential equations. *Proc R Soc Lond B.*, vol. 221, pp. 87-102, 1984.

[10] G. Pfurtscheller and F. Lopes da Silva, “Event-related EEG/MEG synchronization and desynchronization: Basic principles,” *Clin. Neurophys.*, vol. 110, pp. 1842-1857, 1999.

[11] G. Pfurtscheller and C. Neuper, “Motor imagery and direct brain-computer communication,” *Proc. IEEE*, vol. 89, pp. 1123-1134, 2001.