Neural Networks 22 (2009) 1278–1285

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

2009 Special Issue

A linear feature space for simultaneous learning of spatio-spectral filters in BCI

J. Farquhar *

Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, The Netherlands

ARTICLE INFO

Article history:
Received 14 October 2008
Received in revised form 2 June 2009
Accepted 26 June 2009

Keywords:
Brain computer interfaces
Kernel methods
Spatial filtering
Spectral filtering
Event-related-desynchronisation

ABSTRACT

It is shown how two of the most common types of feature mapping used for classification of single trial Electroencephalography (EEG), i.e. spatial and frequency filtering, can be equivalently performed as linear operations in the space of frequency-specific detector covariance tensors. Thus by first mapping the data to this space, a simple linear classifier can directly learn optimal spatial + frequency filters. Significantly, if the classifier's loss function is convex, learning these filters is a convex minimisation problem. It is also shown how to pre-process the data such that the resulting decision function is robust to the biases inherent in EEG data. Further, based upon ideas from Max Margin Matrix Factorisation, it is shown how the *trace norm* can be used to select solutions which have *low rank*. Low rank solutions are preferred as they reflect prior information about the types of EEG signals we expect to see, i.e. that the classifiable information is contained in only a few spatio/spectral pairs. They are also easier to interpret. This feature-space transformation is compared with the Common-Spatial-Patterns on simulated and real Imagined Movement Brain Computer Interface (BCI) data and shown to give state-of-the-art performance.

1. Introduction

The aim of Brain Computer Interface (BCI) research (Birbaumer et al., 1999; Wolpaw, Birbaumer, McFarland, Pfurtscheller, & Vaughan, 2002) is to provide a direct link from human intentions, as observed in brain signals, to control of computers. One remarkable feature of current BCI systems is the high complexity of their feature extractors in comparison to their simple (usually linear) classifiers. Thus, correctly tuning the feature extractor, by for example identifying the best spatial filter or spectral band, is critical to the BCI's performance. This tuning is generally based upon either prior knowledge, such as the selection of the 7–30 Hz band for detecting Event Related Desynchronisation (ERD) in imagined-movement BCIs, or some approximate measure of the classifiability of the resulting features, such as r-scores, the derived feature 'independence' used in ICA (Hyvärinen, Karhunen, & Oja, 2001), or ratios of class variances used by CSP (Koles, 1991). This leads to a complex heterogeneous training process with different learning algorithms using different objective functions to learn different parts of the problem.

The contribution of this paper is a simpler unified approach which integrates the two most common types of feature extraction used in BCI, i.e. spatial and spectral filter estimation, within a single well-regularised *convex* objective function. It is also shown that by defining an appropriate kernel function such combined spatial + frequency filters can be learned efficiently with conventional kernel methods.

2. Problem setting

BCI Signal extraction can be seen as a type of supervised Source Separation or beam-forming problem. That is, a region of the brain generates a class-specific source signal which due to signal propagation and volume conduction effects (van den Broek, Reinders, Donderwinkel, & Peters, 1998) is (linearly) mixed with many other non-class-specific noise signals before being detected at multiple spatially distributed detectors. That is;

$$X = A[\mathbf{s}_y^{(+/-)}; \mathbf{s}_1; \ldots; \mathbf{s}_{m-1}] = AS, \tag{1}$$

where $X \in \mathcal{R}^{d\times T}$ is the signal from d detectors over T sampled time-points, $\mathbf{s}_y^{(+/-)} \in \mathcal{R}^{1\times T}$ is the class-dependent source signal, $\{\mathbf{s}_1, \ldots, \mathbf{s}_{m-1}\}$ are the noise sources, and $A \in \mathcal{R}^{d\times m}$ is the source mixing matrix.

The BCI system's task is therefore to find some transformation of X which *unmixes* class-specific sources, $\mathbf{s}_y$, from the irrelevant noise signals such that the class specific signal ($\mathbf{s}_y^+$ or $\mathbf{s}_y^-$) the user was generating can be identified with maximal accuracy. How to find such an unmixing transformation depends on which features of the source signal $\mathbf{s}_y$ are class-dependent. In BCI applications these features are of two main types,

- *temporal* where the signal amplitude itself of the source is class-dependent
- *spectral* where the signal's power in particular *frequency bands* is class-dependent.

* Tel.: +31 243611938; fax: +31 243616066.
E-mail addresses: J.Farquhar@donders.ru.nl, jdrf@zepler.org.

0893-6080/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2009.06.035

3. Classifying temporal sources

Temporal sources form the basis of many successful exogenous BCI systems. Here an external stimulus evokes a characteristic temporal response in the detected brain signal which is modulated by the level of user attention to the stimulus (Ross, Herdmann, & Pantev, 2005) — making it class specific. Perhaps the most famous and successful evoked response BCI is the (Farwell & Donchin, 1988) style p300 visual speller. Here a grid of symbols are displayed on the screen and flashed in parallel in such a way that each symbol has a unique flash sequence. Thus if the user attends to a specific character, the brain's response to its flash sequence can be used to identify the attended symbol.

Given a particular flash sequence, the classification problem is one of determining if X contains an attended, $\mathbf{s}_y^+$, or unattended, $\mathbf{s}_y^-$, manipulation response. If a linear classifier is used to classify the unmixed, then due to the linearity of the unmixing process (it is just the inverse of the linear mixing) the combination of unmixing and classification is also linear. Mathematically this means,

$$g(S) = \langle S, W^{(S)} \rangle \tag{2}$$

$$= \mathrm{Tr}(UX(W^{(S)})^\top) \tag{3}$$

$$= \mathrm{Tr}((W^{(S)})^\top UX) \tag{4}$$

$$= \mathrm{Tr}(W^{(X)}X) = \langle W^{(X)}, X \rangle \tag{5}$$

where $g(S)$ is the classifier's score function over unmixed sources, U is the matrix which undoes the mixing process, $W^{(S)}$ is the weight matrix for the linear classifier over unmixed sources, and $W^{(X)}$ is the equivalent weighting over X. $\langle \cdot, \cdot \rangle$ is used to denote the Euclidean inner product and its generalisations to matrices and higher order tensors, in all cases it denotes the sum element-wise products, e.g. $\langle x, y \rangle = \sum_i x(i)y(i)$ for vectors, $\langle X, Y \rangle = \sum_{i,j} X(i,j)Y(i,j)$ for matrices.

Thus, the linearity of the classifier over the unmixed source features means that for temporal sources, instead of learning the unmixing matrix and the classifier weights as separate steps, we can equivalently learn $W^{(X)}$ directly. Furthermore, since $(W^{(S)})^\top U = W^{(X)}$, the pairs of unmixing matrix rows (spatial filters) and source classification weights (temporal filters) can be extracted from $W^{(X)}$ by singular-value-decomposition.

Note that in general A is not full rank so the mixing process is not exactly invertible — thus one cannot extract the signal from a single source but only a mixture of a few sources. In practice this is not a problem as detectable sources tend to be large enough that extracting them is not a problem. In this paper we do not consider this ill-posedness problem further as our goal is to learn $W^{(X)} = (W^{(S)})^\top U$ directly without estimating (or inverting) A.

4. Classifying spectral sources

Spectral sources are also widely used in BCI. They come in 2 main types, naturally occurring neural oscillations, and externally evoked oscillations. *Imagined movement* (Peters, Pfurtscheller, & Flyvjerg, 2001) BCIs use naturally occurring oscillations in the μ-(8–14 Hz) and β-(14–30 Hz) bands. By performing real or imagined body movements, the power in these bands can be modulated in a spatially localised way, with a power decrease (Event Related Desynchronisation) during movement and a power increase (Event Related Synchronisation) afterwards (Pfurtscheller & Silva, 1999). The main change in power is also localised to the cortical region responsible for controlling that region of the body. Thus the user can generate class-dependent source signals by, for example, imagining moving their left hand for the positive class and right hand for the negative class. Externally evoked oscillations are used in the Steady State Evoked Potential (SSEP) BCIs. These use a high frequency (20–80 Hz) stimulus to induce a neurological response at the same frequency. As with the temporal BCIs, the amplitude of this response is modulated by the user's attention to the stimulus, allowing them to generate a class-dependent source. SSEPs have been used in many BCIs across many different modalities, including visual (Fries, Reynolds, Rorie, & Desimone, 2001), tactile and auditory (Moller, 1974; Picton, John, Dimitrijevic, & Purcell, 2003; Ross et al., 2005).

Combining source unmixing and classification for temporal sources is simple because the class-dependent source characteristic, i.e. its time course, is discriminable with a linear weighting over the extracted source signal. Thus the classification and unmixing can be combined into a single linear operation. Unfortunately, when the class-dependent information is contained in a more general non-linear characteristic of the source signal, such as its power, such a direct merging of unmixing and classification is no longer possible. The rest of this paper shows how for second order statistics of the source signals, such as band power, unmixing and classification can still be performed as a linear operation in a transformed feature space. We start by showing how to unmix and linearly classify signal variances, and then extend this result to unmixing and classifying band-powers.

4.1. Classifying signal variances

Consider the case where the feature of interest is simply the variance of the source. Given the source unmixing matrix, U, the signal variance can be computed for source k as, $\sigma_k = \mathbf{s}_k\mathbf{s}_k^\top = [U_kX][U_kX]^\top = U_k\Sigma U_k^\top$, where $\Sigma = XX^\top$ is the signal covariance, and U_k is the kth *row* of U. Thus, the unmixed signal variance, σ_k, can be computed as a quadratic transformation of the signal covariance. The key problem for classification is thus to find the spatial filters which equal the source rows of the unmixing matrix.

4.1.1. Common spatial patterns (CSP)

Directly finding spatial filters which provide good classification performance for signal variance features is the core problem addressed by the CSP algorithm (Blankertz, Tomioka, Lemm, Kawanabe, & Müller, 2008; Koles, Lazar, & Zhou, 1990). CSP finds its spatial filters, $\mathbf{w}^{(\sigma)}$, by simultaneously maximising the signal variance for trials in one class, $\mathbf{w}^{(\sigma)}\Sigma_+\mathbf{w}^{(\sigma)}$, whilst minimising the variance in the other class(es), $\mathbf{w}^{(\sigma)}\Sigma_-\mathbf{w}^{(\sigma)}$. $\Sigma_{+/-}$ is the total data covariance over all positive (resp. negative) class trials. CSP does this by maximising the *ratio* of class variances. This is equivalent to maximising a *Rayleigh quotient* which in turn is equivalent to solving the generalised eigenvalue problem,

$$\Sigma_+\mathbf{w}^{(\Sigma)} = \lambda\Sigma_-\mathbf{w}^{(\Sigma)}, \tag{6}$$

where, λ is the eigenvalue of this solution.[1] Solving this generalised eigenvalue problem one finds a full class-focused unmixing matrix, W. Further, CSP has the nice property that the eigenvalue λ equals the spatial filters per-class variance ratio. This is one measure of classification performance and is commonly used to select which subset of CSP's spatial filters should be used.

CSP is very widely used because of its simplicity and high performance. The biggest problem with CSP is that its ratio-of-variances objective function can be sensitive to outliers leading to over-fitting. Many variants of CSP have been developed to address this problem (Blankertz et al., 2007; Farquhar, Hill, Lal, & Schölkopf, 2006). However, a better solution is to develop alternative methods which merge the unmixing process with classification in such a way that any robust classification objective can be used. In the next section we show how this can be done using a linear feature-space transformation for variance features.

[1] There are many equivalent formulations of a binary CSP.