Biomedical Signal Processing and Control 35 (2017) 70–78

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

Generalized sparse discriminant analysis for event-related potential classification

Victoria Peterson [a,*], Hugo Leonardo Rufiner [a,b], Ruben Daniel Spies [c]

[a] Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional, UNL, CONICET, FICH, Ruta Nac. 168, km 472.4, 3000 Santa Fe, Argentina
[b] Facultad de Ingeniería, Universidad Nacional de Entre Ríos, Ruta Prov. 11, km 10, 3100 Oro Verde, Argentina
[c] Instituto de Matemática Aplicada del Litoral, UNL, CONICET, FIQ, Predio Dr. Alberto Cassano del CCT-CONICET-Santa Fe, Ruta Nac. 168, km 0, 3000 Santa Fe, Argentina

ARTICLE INFO

Article history:
Received 26 September 2016
Received in revised form 14 February 2017
Accepted 8 March 2017

Keywords:
Brain–computer interface
Event-related potential
Kullback–Leibler divergence
Penalization
Sparse discriminant analysis

ABSTRACT

A brain–computer interface (BCI) is a system which provides direct communication between the mind of a person and the outside world by using only brain activity (EEG). The event-related potential (ERP)-based BCI problem consists of a binary pattern recognition. Linear discriminant analysis (LDA) is widely used to solve this type of classification problems, but it fails when the number of features is large relative to the number of observations. In this work we propose a penalized version of the sparse discriminant analysis (SDA), called generalized sparse discriminant analysis (GSDA), for binary classification. This method inherits both the discriminative feature selection and classification properties of SDA and it also improves SDA performance through the addition of Kullback–Leibler class discrepancy information. The GSDA method is designed to automatically select the optimal regularization parameters. Numerical experiments with two real ERP-EEG datasets show that, on one hand, GSDA outperforms standard SDA in the sense of classification performance, sparsity and required computing time, and, on the other hand, it also yields better overall performances, compared to well-known ERP classification algorithms, for single-trial ERP classification when insufficient training samples are available. Hence, GSDA constitute a potential useful method for reducing the calibration times in ERP-based BCI systems.

1. Introduction

A brain–computer interface (BCI) is a system that measures brain activity and converts it into an artificial output which is able to replace, restore or improve any normal output (neuromuscular or hormonal) used by a person to communicate and control his/her external or internal environment. Thus, BCI can significantly improve the quality of life of people with severe neuromuscular disabilities [1].

Communication between the brain of a person and the outside world can be appropriately established by means of a BCI system based on event-related potentials (ERPs), which are manifestations of neural activity as a consequence of certain infrequent or relevant stimuli. The main reason for using ERP-based BCI are: it is non-invasive, it requires minimal user training and it is quite robust (in the sense that it can be used by more than 90% of people) [2]. One of the main components of such ERPs is the P300 wave, which is a positive deflection occurring in the scalp-recorded EEG approximately 300 ms after the stimulus has been applied. The P300 wave is unconsciously generated and its latency and amplitude vary between different EEG records of the same person, and even more, between EEG records of different persons [3]. By using the “oddball” paradigm [4] the ERP-based BCI can decode desired commands from the subject by detecting those ERPs in the background EEG. From a pattern recognition point of view, the ERP-based BCI classification problem, in which two classes are involved (EEG with ERP or target class and EEG without ERP or non-target class), is highly complex. This is so mainly for two reasons: the presence of the high inter-trial variability and the unfavorable signal-to-noise ratio.

It is well-know that in any BCI classification scheme two main difficulties must be dealt with: the curse-of-dimensionality and the bias-variance trade-off [5]. While the former is a consequence of working with a concatenation of multiple time points from multiple channels, the latter refers to the generalization capability of the classifier. Several works have proposed different feature extraction methods for reducing the dimension of the feature space and capturing the most discriminative information in a single-trial ERP

* Corresponding author.
E-mail addresses: vpeterson@sinc.unl.edu.ar (V. Peterson), lrufiner@sinc.unl.edu.ar (H.L. Rufiner), rspies@santafe-conicet.gov.ar (R.D. Spies).

http://dx.doi.org/10.1016/j.bspc.2017.03.001
1746-8094/

[6–8]. For instance, the common spatial patterns (CSP) method introduced in [9] is a supervised feature extraction technique which is widely used in motor imagery BCI [10–12]. A Fisher's criterion (FC)-based on spatial filtering for ERP classification, which has shown stronger denoising capability than CSP for ERP-based BCI, was presented in [13].

The feature extraction step is usually follow by the design of an appropriate classification technique. In this regard, although many classification strategies have been proposed, it is widely accepted that linear discriminant analysis (LDA) is a very good classification scheme, resulting most of the times in optimal performances while keeping the solution simple [14]. As a drawback, effective training of a LDA classifier usually requires a number of samples between five and ten times the dimensionality of the patterns [15], resulting in very long system calibration times. Several regularized LDA schemes within the BCI context have been proposed [4,16,14,17]. It has been shown that a regularized version of LDA can significantly increase the classification performance obtained by standard LDA. This improvement is due to the fact that regularization helps avoiding: (i) the influence of outliers and strong noise, (ii) the complexity of the classifier and (iii) the raggedness of the decision surface [16].

One of the main disadvantages of current BCI systems is the fact that they require long calibration times to achieve a reliable and stable communication. Hence, the design of a scheme capable of providing good classification performance in small sample scenarios is highly desirable in order to enhance the practicability of an ERP-based BCI system. As an effort in this direction, for the case of high dimensional data with small training samples, the shrinkage LDA (SKLDA) method presented in [14] seeks to improve the usual estimation of the ill-conditioned covariance matrix used in LDA by a shrinkage covariance estimator.

Also, it has been claimed in [18] that data preprocessing, feature extraction and classification should not be regarded as isolated processes, since attacking each of these tasks separately and ignoring the inter-relationship between them might result in sub-optimum performances. Other works [19,20] also suggest that an unified discriminative approach might provide a better overall performance. In line with the above philosophy, in this article we propose a method in which feature selection and classification are made in an interleaved and integrated process. A well-known and widely used method in which classification and feature selection are jointly made is the so-called stepwise LDA (SWLDA), originally introduced in ERP classification problems by Farwell and Donchin in [4]. The SWLDA method is a combination of forward and backward stepwise regression with statistical testing in which features are automatically selected by adding the most significant variables and removing the least significant ones. This process is iterated until a predetermined number of coefficients are included, or until no additional coefficients satisfy the given entry nor the removal criteria.

More recent classification schemes [21,22,17,23,24] make use of ℓ_1-regularized least squares regression techniques which induce sparse solutions and therefore result in very robust classifiers.

Following the above research direction, in this work we propose a model which combines and makes simultaneous use of regularization, sparse feature selection and a-priori discriminative information. More precisely, we develop a new penalized version of the sparse discriminant analysis (SDA) [25], which we call generalized sparse discriminant analysis (GSDA), with the main objective of solving the binary ERP classification problem. As far as we know SDA has never been used before in ERP-based BCI classification problems. The performance of the GSDA method will first be compared with that of SDA and then, in small training sample scenarios, with those of LDA, SWLDA, SKLDA and FC + LDA. These comparison results will clearly show that our GSDA method has a high potential for reducing calibration times in BCI systems.

The organization of this article is as follows. In Section 2 we make a brief review on discriminant analysis from the statistical literature. Our proposed new approach is presented in Section 3. In Section 4 the two ERP-EEG databases used in the experiments are described. Section 5 contains details on all the experiments and results. Discussions are given in Section 6. Finally, concluding remarks and future works are presented in Section 7.

2. Discriminant analysis: a brief review

The LDA criterion is a well-known dimensionality reduction tool in the context of supervised classification. Its popularity is mainly due to its simplicity and robustness which lead to very high classification performances in many applications [26].

Let $\mathbf{W}_1, \ldots, \mathbf{W}_K$ be p-dimensional random vectors whose distributions uniquely characterize each one of the K classes of a given classification problem. In addition, let $\mathbf{X}$ be an $n \times p$ data matrix such that each one of its rows, $\mathbf{x}_i$, is a realization of one and only one of the aforementioned random vectors, and let $\mathbf{z} \in \{1, 2, \ldots, K\}^n$ be a categorical variable accounting for class membership, i.e. such that if pattern $\mathbf{x}_i$ is a realization of $\mathbf{W}_k$, then $z_i = k$.

The LDA method consists of finding $q < K$ discriminant vectors (directions), $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q$ such that by projecting the data matrix $\mathbf{X}$ over those directions, the "classes" will be well separated one from each other. It is assumed that the random vectors $\mathbf{W}_1, \ldots, \mathbf{W}_K$ are independently and normally distributed with a common covariance matrix $\boldsymbol{\Sigma}_t$. The procedure for finding the vectors $\boldsymbol{\beta}_j$ requires of estimates of the within-class, the between-class and the total covariance matrices, $\boldsymbol{\Sigma}_w$, $\boldsymbol{\Sigma}_b$ and $\boldsymbol{\Sigma}_t$, respectively. These estimates are given by:

$$\hat{\boldsymbol{\Sigma}}_w = \frac{1}{n}\sum_{k=1}^{K}\sum_{i \in I_k}(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T,$$

$$\hat{\boldsymbol{\Sigma}}_b = \frac{1}{n}\sum_{k=1}^{K} n_k(\boldsymbol{\mu}_k - \boldsymbol{\mu})(\boldsymbol{\mu}_k - \boldsymbol{\mu})^T,$$

$$\hat{\boldsymbol{\Sigma}}_t = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T,$$

where I_k and n_k are the set of indices and the number of patterns belonging to class k, respectively, $\boldsymbol{\mu}_k \doteq \frac{1}{n_k}\sum_{i \in I_k}\mathbf{x}_i$ is the k-class sample mean and $\boldsymbol{\mu} \doteq \frac{1}{n}\sum_{k=1}^{K}\boldsymbol{\mu}_k$ is the common sample mean. Note that $\hat{\boldsymbol{\Sigma}}_t = \hat{\boldsymbol{\Sigma}}_w + \hat{\boldsymbol{\Sigma}}_b$.

The LDA method seeks to find the vectors $\boldsymbol{\beta}_j$ in such a way that they maximize separability between classes, which is achieved by simultaneously maximizing $\hat{\boldsymbol{\Sigma}}_b$ and minimizing $\hat{\boldsymbol{\Sigma}}_w$, or equivalently, by simultaneously maximizing $\hat{\boldsymbol{\Sigma}}_b$ and minimizing $\hat{\boldsymbol{\Sigma}}_t$. Since the rank of $\hat{\boldsymbol{\Sigma}}_b$ is at most $K-1$, there are at most $K-1$ non-trivial solutions $\boldsymbol{\beta}_j^*$. Usually $q = K-1$.

In the particular case $K = 2$ (and therefore $q = 1$), the solution to the LDA problem has the following explicit formulation:

$$\boldsymbol{\beta}^* = \hat{\boldsymbol{\Sigma}}_t^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2). \tag{1}$$

This special case is known as Fisher linear discriminant analysis (FLDA) [27]. The FLDA approach can be formulated as a linear regression model [27,26]. Let $\mathbf{X}$ be as before and let $\mathbf{y}$ be a n-dimensional vector such that $y_i = \frac{n_2}{n}$ or $y_i = -\frac{n_1}{n}$, depending on whether the ith observation belongs to class 1 or to class 2,