Determination of an optimal training strategy for a BCI classification task with LDA

Iván E. Gareis, Rubén C. Acevedo, Yanina V. Atum, Gerardo G. Gentiletti (*IEEE Member)*, Verónica Medina Bañuelos, Hugo L. Rufiner (*IEEE Member)*

***Abstract*—Brain computer interfaces (BCIs) translate brain activity into computer commands. To enhance the performance of a BCI, it is necessary to improve the feature extraction techniques being applied to decode the users' intentions. Objective comparison methods are needed to analyze different feature extraction techniques. One possibility is to use the classifier performance as a comparative measure. In this paper, we study the behavior of linear discriminant analysis (LDA) when used to distinguish between electroencephalographic (EEG) signals with and without the presence of event related potentials (ERPs).**

I. Introduction

BRAIN computer interfaces (BCI) are devices that provide a direct link between the brain and a computer [1]. Such interfaces can be considered as being the only way of communication for people affected by a number of motor disabilities [2].

The performance of a brain computer interface is highly dependent on the signal processing techniques used to extract the features that encode the BCI user intentions [3]. Therefore, there is a need for objective comparison methods to analyze different feature extraction techniques [4]. One straightforward solution is to feed a classifier with the features we wish to compare, and use its performance as a measure of the separation power of such features. In this case, care must be taken in order to consider only the variations in system performance caused by the particular properties of the feature extraction techniques that are being evaluated. Any other issues, like the ones that may result from changes in the parameters of the classifier, must be ignored. In order to do so, it is important to study the behavior of the classifier to be used in conditions and with data similar to the ones that will be presented when using it as a feature extraction techniques comparison method.

One of the most popular classifiers for BCI applications is the Fisher´s linear discriminant analysis (LDA) [5, 6]. Even though the LDA has been extensively studied [7-9], the effect of unbalanced training datasets using electroencephalographic (EEG) data and the number of patterns necessary to reach a performance plateau have not been tested. That is, the point at which no significant performance gain will exist when adding more training patterns has not been determined.

In this paper, we address the problem of studying the behavior of LDA when used to discriminate between EEG signals containing event related potentials (ERPs) and EEG signals without the presence of ERPs.

II. System overview

A. P300-Speller

When infrequent or particularly significant auditory, visual, or somatosensory stimuli, are mixed with frequent or routine stimuli, ERPs are typically evoked over the parietal cortex. This phenomenon can be used to implement a BCI commonly called P300 speller, which allows the user to select symbols from a matrix in a computer screen [10].

In the classical P300 speller the user faces a 6 x 6 matrix that contains all letters and characters. During the experiment a single row or column is intensified randomly with a predefined frequency; and, in a complete block of 12 intensifications, each row or column flashes once. To make a selection the user focuses on the character he/she desires to choose. As a result, assuming the intensification of one character of the matrix elicits ERPs, there will be two target trials and ten non target trials in each block. Typically the block of intensifications has to be repeated to effectively determine the character the user is focusing on.

To determine which intensification elicits an ERP the system has to be able to solve the binary classification problem (two possible classes: recordings with ERP and recordings without ERP).

B. Database

The Neuroimaging Research Laboratory at Universidad Autónoma Metropolitana (UAM) provided a database containing the recordings of 30 healthy subjects using the P300 speller on a BCI2000 platform [11, 12]. Ten channels of ERP (N_c) were recorded using a sample frequency of 256 Hz. Channels Fz, C3, Cz, C4, P3, Pz, P4, PO7, PO8, Oz

This work was supported by the Universidad Nacional de Entre Ríos (grants PID 6101 and 6106), the Universidad Nacional de Litoral (grant CAI+D 012-72), STIC-AMSUD program (09STIC01) and the Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) from Argentina

I. E. Gareis, R. C. Acevedo, Y. V. Atum and G. G. Gentiletti are with Laboratorio de Ingeniería en Rehabilitación e Investigaciones Neuromusculares y Sensoriales; Facultad de Ingeniería, Universidad Nacional de Entre Ríos, CC 47 suc.3, CP 3100 Paraná, Argentina

V. Medina Bañuelos is with Laboratorio de Investigación en Neuro Imagenología; División Ciencias Básicas e Ingeniería, Universidad Autónoma Metropolitana (Iztapalapa), México.

H. L. Rufiner is with Centro de I+D en Señales, Sistemas e Inteligencia Computacional; Facultad de Ingeniería y Ciencias Hídricas, Universidad Nacional de Litoral, Argentina and Laboratorio de Cibernética, Facultad de Ingeniería, Universidad Nacional de Entre Ríos.

were recorded using a right ear reference and a right mastoid ground. A complete description of the parameters used for the speller and the data are available on the database website: http://akimpech.izt.uam.mx/dokuwiki/doku.php.

Each subject in the database participated in four sessions with fifteen sequences per session. This yields a number N_t of labeled target trials equal to 630 and a number of labeled non target trials N_{nt} equal to 3150.

One of the premises of the creation of these database was to provide a realistic sample of the recordings, thus many of them present a significant number of outliers. A selection of ten subjects has been made among the ones without a large number of outlier samples, in order to prevent these variables to influence the results and to avoid using an artifact rejection block.

C. Preprocessing

As a first preprocessing stage an eighth-order forward-backward Chebysheb lowpass filter was used to filter the signals. The cutoff frequency was set to 7.0 Hz. The EEG was then downsampled from 256 Hz to 16 Hz by selecting each 16th sample from the lowpass filtered data.

The signals from each electrode were normalized independently as to have a zero mean and a unitary standard deviation.

Single trials, having one second duration, were extracted from the data and started at the beginning of the intensification of a character. Due to the trial duration and the downsampling rate the number of samples per trial or N_s is 16.

Finally the feature vectors (or patterns) were constructed by concatenating the single trials from the ten channels. Therefore the dimension of the feature vectors was N_c x N_s, or 160.

D. Classifier

The objective of LDA is to compute a discriminant vector $w \in \mathbb{R}^D$ (in this work D = 160) that, given a set of training patterns $x_j \in \mathbb{R}^D$, $j \in \{1...N\}$ with their corresponding class labels, separates the classes as well as possible. This is achieved in LDA by maximizing the criterion function represented by

$$J(w) = \frac{(\tilde{m}_1 - \tilde{m}_2)^2}{\tilde{s}_1^{\,2} + \tilde{s}_2^{\,2}} \tag{1}$$

where

$$\tilde{m}_k(w) = \frac{1}{N_k} \sum_{i \in Y_k} w^{\mathrm{T}} x_i, \quad \tilde{s}_k^{\,2} = \sum_{i \in Y_k} (w^{\mathrm{T}} x_i - \tilde{m}_k)^2, \tag{2}$$

Y_k is the set of indices i corresponding to class k and N_k is the number of training patterns corresponding to class k.

It can be proven that the w that maximizes (1) can be found by [13].

$$w \propto S_w^{-1}(m_1 - m_2) \tag{3}$$

where

$$m_k = \frac{1}{N_k} \sum_{i \in Y_k} x_i, \tag{4}$$

$$S_w = \sum_{k=1}^{2} \sum_{i \in Y_k} (x_i - m_k)(x_i - m_k)^{\mathrm{T}}. \tag{5}$$

In the LDA the between-class scatter matrix S_w can become singular, and the inverse of S_w can become ill defined. This happens when the number of features becomes larger than the number of training patterns, and is called the small sample size problem [14].

E. Performance evaluation

The error rate (ER), is the most widely used evaluation metric. However, as an average over all the observations that are classified, it favors the majority class, i.e., the class with higher prior probability [15].

For two-class discrimination of unbalanced data, the area under the receiver operating characteristic curve (AUC) is commonly used. The receiver operating characteristic (ROC) curve is a plot of true positive rate vs. false positive rate, and hence a higher AUC generally indicates a better classifier. In contrast to ER, AUC is invariant to the prior probabilities [16, 17].

Considering the different characteristics of ER and AUC, both were used to estimate the performance of the classifiers.

ER and AUC are not useful parameters to estimate the capacity a system has to accurately recognize one class, independently from its capacity to recognize the other [16]. Therefore the sensitivity and the specificity were also computed. The sensitivity is the fraction of correctly classified objects in the target class (in our case the target class is constituted by the patterns with ERP). The specificity is the fraction of non target objects that are not classified into the target class.

III. Experiments

Five sets of experiments were carried out each using different unbalance ratios to compute the classifiers, ranging from one to five target patterns per non target pattern. The different ratios were generated by random under-sampling of the non target patterns [15]. The first set of experiments represents the balanced situation, and the fifth set represents the situation when all the data are included. The other three sets represent less natural situations when using a 6 x 6 stimulation matrix, but when modifying the matrix size this target vs. non-target ratios can be present. The inclusion of these three sets also allows us to analyze trends.

In each set of experiments the classifiers were computed varying the number of training target patterns N_{ti} and the number of non target patterns N_{nti} according to

$$N_{ti} = \left[0.9^i \times N_t\right], \quad N_{nti} = \left[k \times 0.9^i \times N_t\right] \tag{6}$$

where i corresponds to the integers ranging from one to nineteen, k takes an integer value between one and five,

corresponding to the set of experiments considered, and N_t is the number of target patterns for each subject in the database. Care was taken to use more training patterns than features in the training sets, and thus to avoid dealing with the above mentioned small sample size problem. The patterns that were not used to train the classifiers constituted the validation datasets.

The performance was estimated by cross-validation [13]. With each experimental configuration, the classifiers were trained and tested thirty times with different randomly selected training and validation datasets and the results were averaged. Please notice the difference between this process and the m-fold cross validation, where the training set is randomly divided into m disjoint sets of equal size N_t/m.

IV. Results

Fig. 1 shows the performance results. It should be point out that the number of training patterns shown in the abscissas axis is different for each set of experiments; this is due to the inherent unbalance of the problem and to the balancing approach used. These values are the sum of N_{ti} and N_{nti} as given in (6) with the corresponding value of k. It is also important to notice, that even though the number of trials used for training is different for the corresponding points in the graphics (i.e., the points corresponding to the same i with different k values), the time the user should spend in the training session is the same.

V. Discussion

From the evaluation of the results it can be seen that the classifiers reach the performance plateau at around 1000 training patterns for all experimental sets.

The AUC has not shown any variations between the experimental sets attaining similar values for all cases when the classifiers training had been made with over 1000 trials. Regarding the ER, a variation for the different experimental sets could be observed, favoring the ones with larger unbalances.

The effect of unbalance over the LDA behavior can be clearly seen when analyzing the variations of specificity and sensitivity from Fig. 1. These performance measures have very similar values when the classes are balanced, but as the unbalance grows so does the specificity, while the sensitivity decreases accordingly. This type of behavior is also mentioned in [8].

An overall good performance was obtained with the proposed system, as the AUC values were over 0.9 and the ER below 0.1.

VI. Conclusions

In this study we have estimated the number of training trials that are necessary to reach a performance plateau using LDA to classify EEG with and without ERP. It is interesting to notice that this number was not dependent on the number of training trials corresponding to each class, but rather on the total number of training trials. However, other variables such as the number of features in the patterns and the difficulty of the classification problem were not considered here, so a more thorough analysis of their impact must be carried out.

The variations seen in specificity and sensitivity provide important information about the response of LDA. In addition of being significant factors when using the classifier to measure the discriminating power of a features set, they can be useful to analyze the performance of a system that uses this type of classifier.

In further work, the problem of determining the effect of variations in the number of features and the difficulty of the classification problem should be addressed. Also, an extension to other types of classifiers with different characteristics should be considered.

References

[1] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain Computer Interfaces for communication and control." Clin. Neurophysiol., vol. 113, no. 6, pp. 767-791, Jun 2002.

[2] A. Kübler, B. Kotchoubey, J. Kaiser, J.R. Wolpaw and N. Birbaumer, "Brain–computer communication: unlocking the locked in." Psychol. Bull. 127 358–75, 2001.

[3] A. Bashashati, M. Fatourechi, R. K. Ward and G. E. Birch, "A survey of signal processing algorithms in brain-computer interfaces based on electrical brain signals." Journal of Neural Engineering, 4:32-57, 2007.

[4] Atum, Y.; Gareis, I.; Gentiletti, G.; Acevedo, R.; Rufiner, L, "Genetic feature selection to optimally detect P300 in brain computer interfaces," Engineering in Medicine and Biology Society (EMBC), Annual International Conference of the IEEE, Page(s): 3289 – 3292, 2010.

[5] F. Lotte, M. Congedo, A. Lecuyer, F. Lamarche and B. Arnaldi, "A review of classification algorithms for EEG-based brain-computer interfaces," Journal of Neural Engineering, 4:R1-R13, 2007.

[6] DJ Krusienski, EW Sellers, F Cabestaing, S Bayoudh, DJ McFarland, TM Vaughan, JR Wolpaw, "A Comparison of Classification Techniques for the P300 Speller," Journal of Neural Engineering, 3:299-305, 2006.

[7] Fisher R. A. "The use of multiple measurements in taxonomic problems," Ann. Eugenics 7 179–88, 1936.

[8] Xue, J.H. and Titterington, D.M. "Do unbalanced data have a negative effect on LDA?," Pattern Recognition, 41 (5). pp. 1575-1588. ISSN 0031-3203, 2008.

[9] J.G. Xie, Z.D. Qiu, "The effect of imbalanced data sets on LDA: a theoretical and empirical analysis," Pattern Recognition 40 (2) 557–562, 2007

[10] L.A. Farwell, E. Donchin "Talking off the top of your head: toward a mental prothesis utilizing event-related brain potentials." Electroenceph. clin. Neurophysiol. 70:510–523, 1988.

[11] http://www.bci2000.org/BCI2000/Home.html.

[12] Claudia Ledesma Ramírez, Erik Bojorges Valdez, Oscar Yañez Suárez, Carolina Saavedra, Laurent Bougrain, Gerardo Gentiletti. "An open-access P300 speller database," Fourth international BCI meeting, Paper L-12, Monterrey California, 2010.

[13] Richard O. Duda and Peter E. Hart and David G. Stork, "Pattern Classification (2nd Edition)," Wiley-Interscience, 2000.

[14] K. Fukunaga, "Introduction to Statistical Pattern Recognition," Academic Press, 1990.

[15] G.M. Weiss, "Mining with rarity: a unifying framework," SIGKDD Explor. 6 (1), 7–19, 2004.

[16] A.P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," Pattern Recognition 30 (7) 1145–1159, 1997.

[17] MH Zweig, G Campbell, "Receiver-Operating Characteristic (ROC) Plots: A Fundamental Evaluation Tool in Clinical Medicine" Clinical Chemistry, 1993.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
I. E. Gareis, R. C. Acevedo, Y. Atum, G. Gentiletti, V. Medina Bañuelos & H. L. Rufiner; "Determination of an optimal training strategy for a BCI classification task with LDA"
Proc. of the 5th International IEEE EMBS Conference on Neural Engineering, 2011.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
I. E. Gareis, R. C. Acevedo, Y. Atum, G. Gentiletti, V. Medina Bañuelos & H. L. Rufiner; "Determination of an optimal training strategy for a BCI classification task with LDA"
Proc. of the 5th International IEEE EMBS Conference on Neural Engineering, 2011.

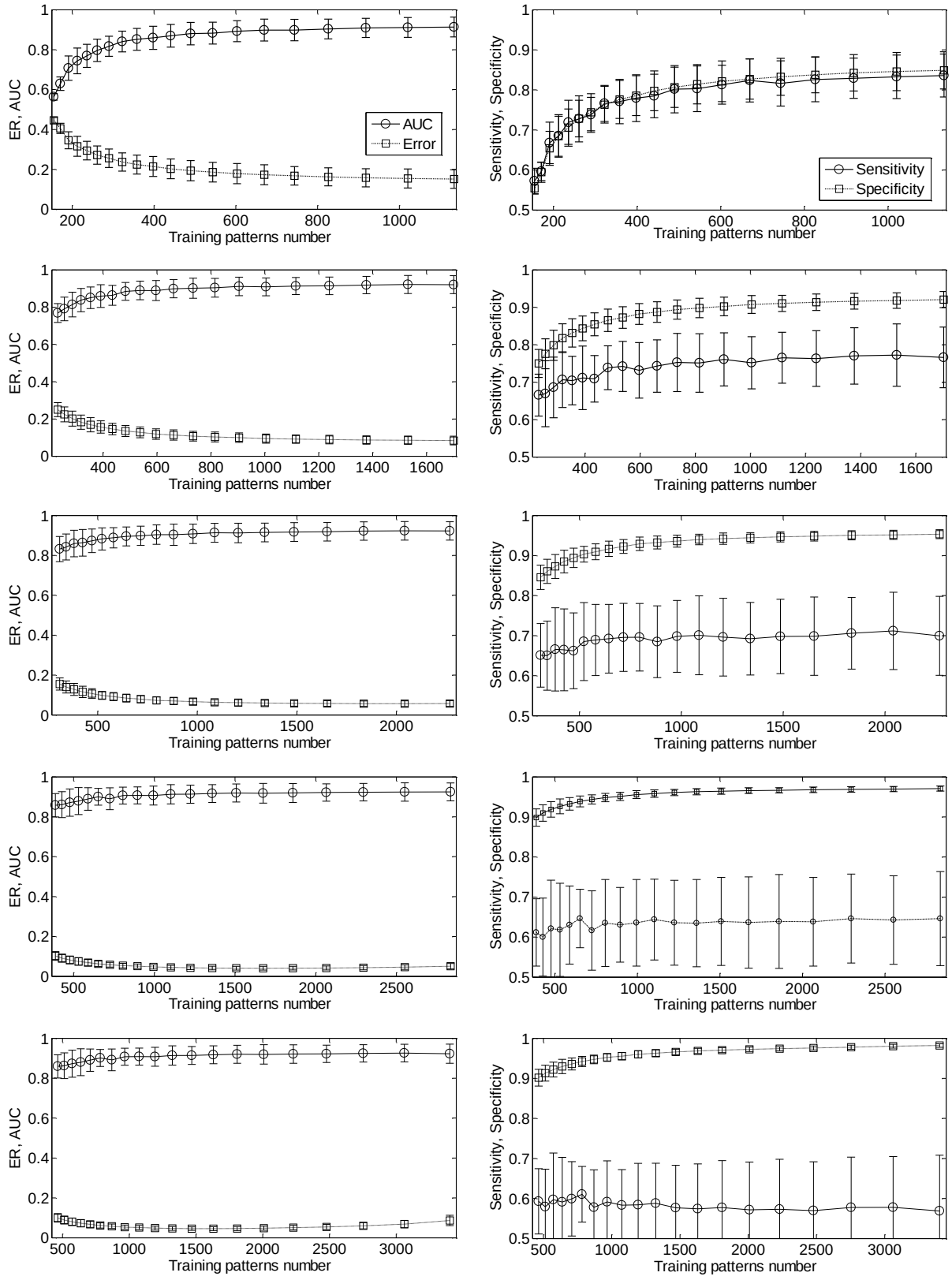

Fig. 1. Graphics of performance estimates averaged over subjects with the corresponding standard deviations vs. number of total training patterns obtained from the different sets of experiments. From top to bottom: using one non target per target pattern, using two non target per target pattern, using three non target per target pattern, using four non target per target pattern, using five non target per target pattern. Legend acronyms: ER (error rate), AUC (area under ROC curve).