On the use of LDA performance as a metric of feature extraction methods for a P300 BCI classification task

Iván Gareis[1], Yanina Atum[1], Gerardo Gentiletti[1], Rubén Acevedo[1], Verónica Medina Bañuelos[2], Leonardo Rufiner[3, 4]

[1] Laboratorio de Ingeniería en Rehabilitación e Investigaciones Neuromusculares y Sensoriales; Facultad de Ingeniería, Universidad Nacional de Entre Ríos, Argentina
[2] Laboratorio de Investigación en Neuro Imagenología; División Ciencias Básicas e Ingeniería, Universidad Autónoma Metropolitana (Iztapalapa), México.
[3] Centro de I+D en Señales, Sistemas e Inteligencia Computacional; Facultad de Ingeniería y Ciencias Hídricas, Universidad Nacional de Litoral, Argentina.
4 Laboratorio de Cibernética; Facultad de Ingeniería, Universidad Nacional de Entre Ríos, Argentina.

E-mail; ivangareis@bioingenieria.edu.ar

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 work the effect of several variables that affect the behaviour of linear discriminant analysis (LDA) has been studied when used to distinguish between electroencephalographic signals with and without the presence of event related potentials (ERPs). The error rate (ER) and the area under the receiver operating characteristic curve (AUC) were used as performance estimators of LDA. The results show that the number of characteristics, the degree of balance of the training patterns set and the number of averaged trials affect the classifier's performance and therefore, must be considered in the design of the integrated system.

1. 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].

Figure 1 shows the general architecture of a BCI proposed by Millán *et al* [4] where the functional blocks are described. In this figure it can be seen how a subject (user) could control a device (e.g. a motorized wheelchair).

Most BCI systems are based on electroencephalography (EEG), but there are several ways (or paradigms) to obtain the desired control signals. One of them is the oddball paradigm, which is based on event related potentials (ERP). The ERP are evoked potentials with latencies higher than 100 ms whose expression depends on psychological and behavioral processes. When infrequent visual or auditory stimuli are mixed with frequent stimuli, the former evoke a potential in the EEG in the parietal cortex with a peak located around the 300 ms called P300. In order to estimate or detect the

ERP, the initial signal to noise ratio (SNR) must be improved because many recorded epochs are immersed in high levels of noise; background brain activity and electromyogram are some examples.

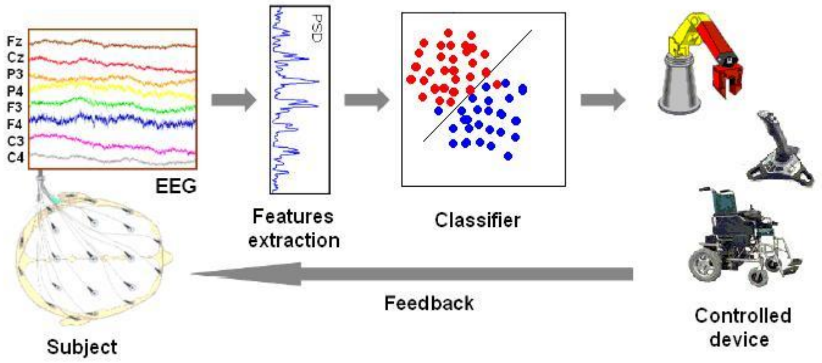


Figure 1. General architecture of BCI for device control.

With the recent advancement of machine learning algorithms and digital processing techniques, a part of the BCI research lies on exploration of feature extraction and classification techniques. The performance of a brain computer interface is highly dependent on these 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 that are being compared, 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 the system's performance caused by the particular properties of the feature extraction techniques that are being evaluated. Any other issues, like those 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, 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 is addressed.

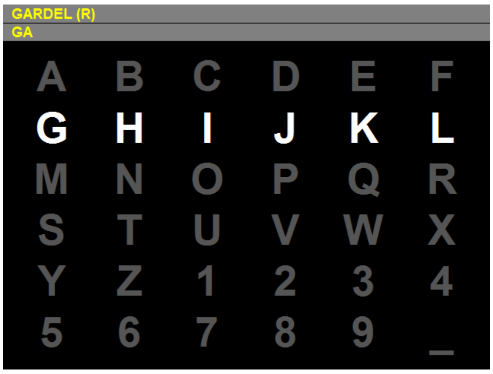
2. Methodology

2.1. 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. Figure 2 shows a matrix during the intensification of the second row.

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).


GARDEL (R)
GA

A B C D E F
G H I J K L
M N O P Q R
S T U V W X
Y Z 1 2 3 4
5 6 7 8 9 _

Figure 2. Classical P300 speller stimulation matrix.

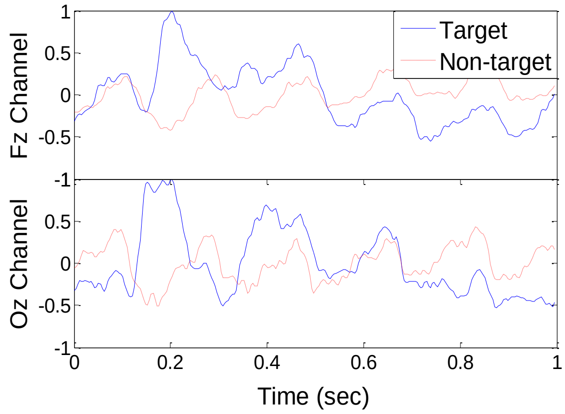


Figure 3. The coherent averages of 600 trials with and without ERP for channels Fz and Oz.

2.2. Recordings

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 (*Nc*) were recorded using a sample frequency of 256 Hz. Channels Fz, C3, Cz, C4, P3, Pz, P4, PO7, PO8, Oz 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/p300db. In figure 3 the coherent averages of 600 trials with and without ERP can be seen. The signals were taken from channels Fz and Oz from a subject of the database.

Each subject in the database participated in four sessions with fifteen sequences per session. This yields a number *Nt* of labeled target trials equal to 630 and a number of labeled non target trials *Nnt* equal to 3150.

One of the premises of the creation of this 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 outliers, in order to prevent these variables to influence the results and to avoid using an artifact rejection block.

2.3. Preprocessing

As a first preprocessing stage the data were filtered and downsampled. Lowpass eighth-order forward-backward Chebyshev filters were used and a downsampling step was performed by selecting each Nth sample from the lowpass filtered data. As it will be explained in section 3 three sets of experiments were performed, each at different cutoff (3.5, 7 and 14 Hz) and decimation (Fsi of 8, 16 and 32 Hz) frequencies.

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

Single trials were extracted from the data, having one second duration and starting at the beginning of the intensification of a character. Due to the trial duration and the downsampling rates, the numbers of samples per trial or Ns_i are 8, 16 and 32 for each set of experiments.

The feature vectors (or patterns) were constructed by concatenating the single trials from the ten channels. Therefore the dimension of the feature vectors was *Nc* x Ns_i, or 80, 160 and 320.

Finally coherent averaging was applied in some experiments as it will be explained in section 3. This technique improves the signal to noise ratio, but may diminish the bit-rate of the BCI and reduces the available number of training patterns.

2.4. Classifier

The objective of LDA is to compute a discriminant vector $w \in \mathbb{R}^D$ 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

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

where

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

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].

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

where

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

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

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]. For these cases several solutions have been proposed, one of which is to carry out the calculation of the inverse by the Moore-Penrose pseudo-inverse [18]. That was the alternative chosen in this work.

2.5. Performance evaluation

The error rate (ER), is the most widely used evaluation metric. However, as it is 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 that 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.

3. Experiments

Three sets of experiments were carried out, each having a different downsampling ratio according to $Fs_i = 2^i \times 4\text{Hz}$, where *i* defines the set.

For each set, five subsets of experiments were defined 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, while the fifth corresponds to the situation when all the data are included. The other three subsets 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 subsets also allows us to analyze trends.

In addition, the effect of coherent averaging was studied, by taking the average of the signals to train and test the classifiers. Averages from two to five trials were tested. In each subset of experiments the classifiers were computed by varying the number of training target patterns Nt_{jkl} and the number of training non target patterns Nnt_{jkl} according to

$$Nt_{jk} = \left[\frac{\text{[illegible]}}{l} \times k\right] \quad Nnt_{jk} = \left[\frac{j\,\text{[illegible]}}{l} \times k\right] \qquad (6)$$

where *j* takes an integer value between one and five corresponding to the subset of experiments considered, *k* corresponds to the integers ranging from one to twenty, *l* varies from one to five and depends on the amount of trials averaged and *Nt* is the number of target patterns for each subject in the database (630).

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. It is important to note 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 *Nt*/*m*.

4. Results

Figures 4, 5 and 6 show the averaged performance results, where all evaluation metrics were plotted against the total amount of training patterns. Each figure corresponds to a different set of experiments, and each row of graphics corresponds to a different subset.

It must be noted the logarithmic scale in the abscissas axis, and that the number of training patterns shown is different for each subset of experiments; this is due to the inherent unbalance of the problem and to the balancing approach used. These values are the sum of the number of target training patterns Nt_{jkl} and the number of non target training patterns Nnt_{jkl}. It is important to consider that even though the number of patterns used for training is different for the corresponding points in the graphics (*i.e.,* the points corresponding to the same *k* with different *j* values), the time the user should spend in the training session is the same for those cases when using a 6 x 6 stimulation matrix. Also the different scale in the ordinates axis in the sensitivity plot of the fifth subset in figure 4 must be noted.

The minimum peaks seen on all metrics when the number of features equals the number of training patterns might appear abnormal, but they are due to the transition between the LDA and the Moore-Penrose pseudo-inverse LDA. This phenomenon is explained on [18].

5. Discussion

In this work the effect of several variables that affect the behavior of LDA has been studied. The variations in the amount of averaged trials have influence on the difficulty of the problem, as does the subsampling rate (or number of features), but this is determined by the amount of discriminating information available on the higher frequencies of the EEG.

The analysis of results shows that when using fewer features the classifiers reach the performance plateau with fewer training patterns, as expected. It is interesting to notice that the number of training patterns at which the performance reach the plateaus is not dependent on the number of training trials corresponding to each class, but rather on the total number of training trials. Another variable that has influence over this is the coherent averaging, since it can be seen that rendering the problem easier, diminishes the amount of training patterns needed to reach the plateau. Although the results for the different subjects are not shown nor analyzed in the paper, there were also differences between them in this regard. Given all these variables affecting the point where the performance reaches a plateau, it is important to be very careful when using or defining a relationship among the necessary amounts of training patterns with the other variables; however the most significant dependence seen was on the number of features. Taking this into account, it can be said as a general rule that, in order to reach the point at which no significant performance gain will exist when adding more training patterns, more than ten training patterns per feature will be needed.

The AUC has not shown any variations between the experimental subsets (i.e. between different unbalance ratios), stabilizing at similar values after reaching the plateau, when averaging the same amount of trials. Regarding the ER, a variation for the different experimental sets could be observed, favoring, as expected, the ones with larger unbalances. On the other hand in some cases it was not possible to reach the plateaus with the balanced datasets, and so for these cases a better performance was obtained using unbalanced datasets. Regarding the other metrics the effect of unbalance over the LDA behavior can be clearly seen when analyzing the variations of specificity and sensitivity from figures 4, 5 and 6. 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].

As expected, the performance was superior when applying coherent averaging if the results having the same amount of training patterns are compared, being the biggest difference the one between the average of two trials and the single trial case. However this is not always true when comparing points that require training sessions of the same duration.

Even though it was not the objective of this work, an overall good performance was obtained with the proposed system, considering its simplicity, since the AUC values were over 0.9 and the ER below 0.1.

6. Conclusions

In this work the effect of several variables that affect the behavior of LDA has been studied. It is important to consider these factors when using the LDA as a tool for evaluating the feature extraction stage, as these strategies tend to influence these variables. These techniques usually change the number of features and in some case the available amount of training patterns, which can affect the system's performance without changing the inherent discrimination power of these features, or in some cases disguising an actual improvement in the quality of the features by a detriment in the classifier performance.

In this study the number of training trials that are necessary to reach a performance plateau using LDA to classify EEG with and without ERP were estimated. However, other variables may influence these results.

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 feature set, it can prove being useful information when designing a BCI based on LDA.

In further work, the problem of evolution of the density distribution of the mappings of the classes as the amount of patterns in each dataset is increased should be studied, as it could give a more complete explanation of the obtained results. Also, an extension to other types of classifiers with different characteristics could be considered.

Acknowledgments

This work was supported by the Universidad Nacional de Entre Ríos (grant PID 6101), 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.

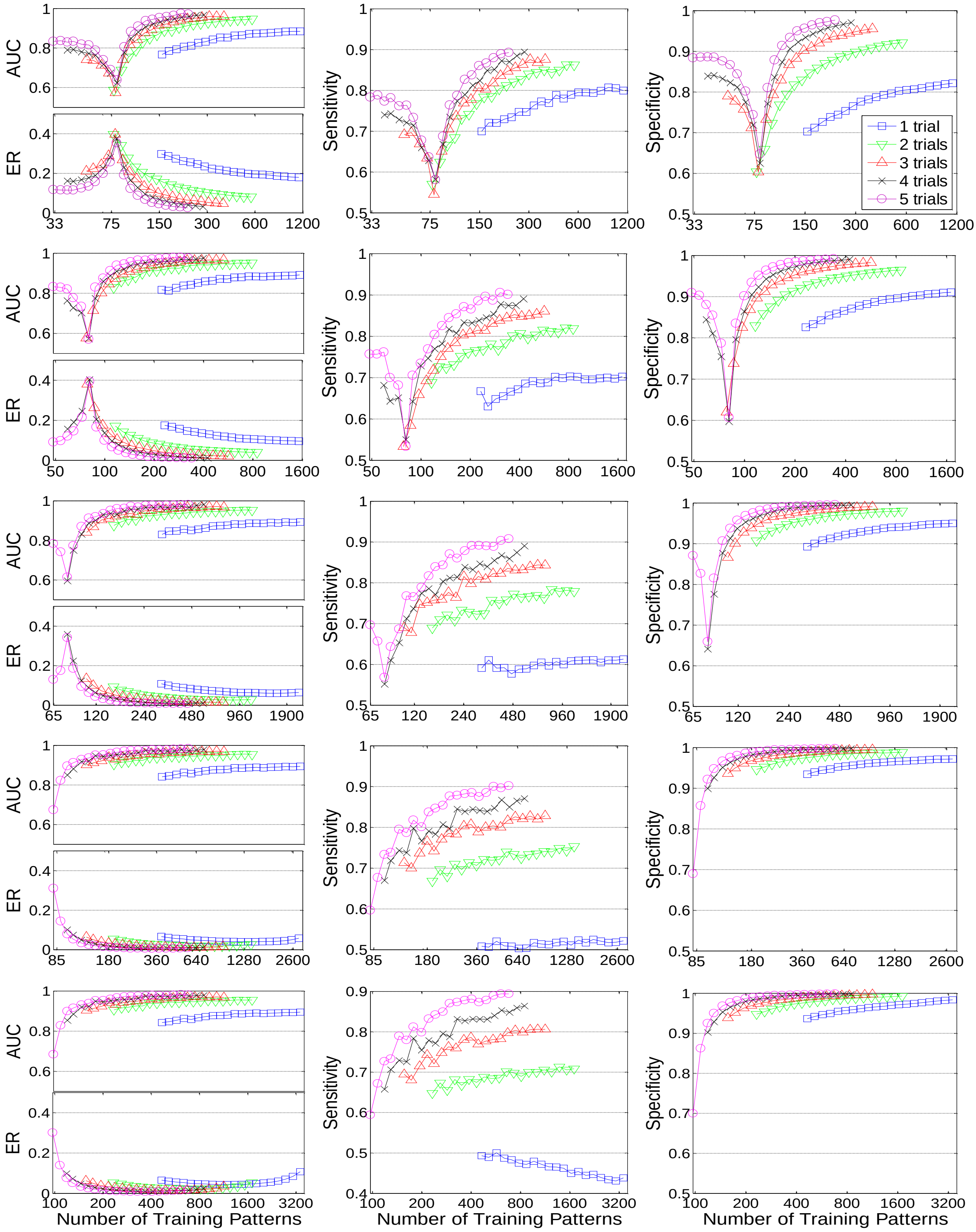


Figure 4. Graphics of performance estimates for the first set of experiments averaged over subjects vs. number of total training patterns obtained from the different subsets of experiments. From top to bottom: using one non target per target pattern, using two non target per target pattern, using three 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).

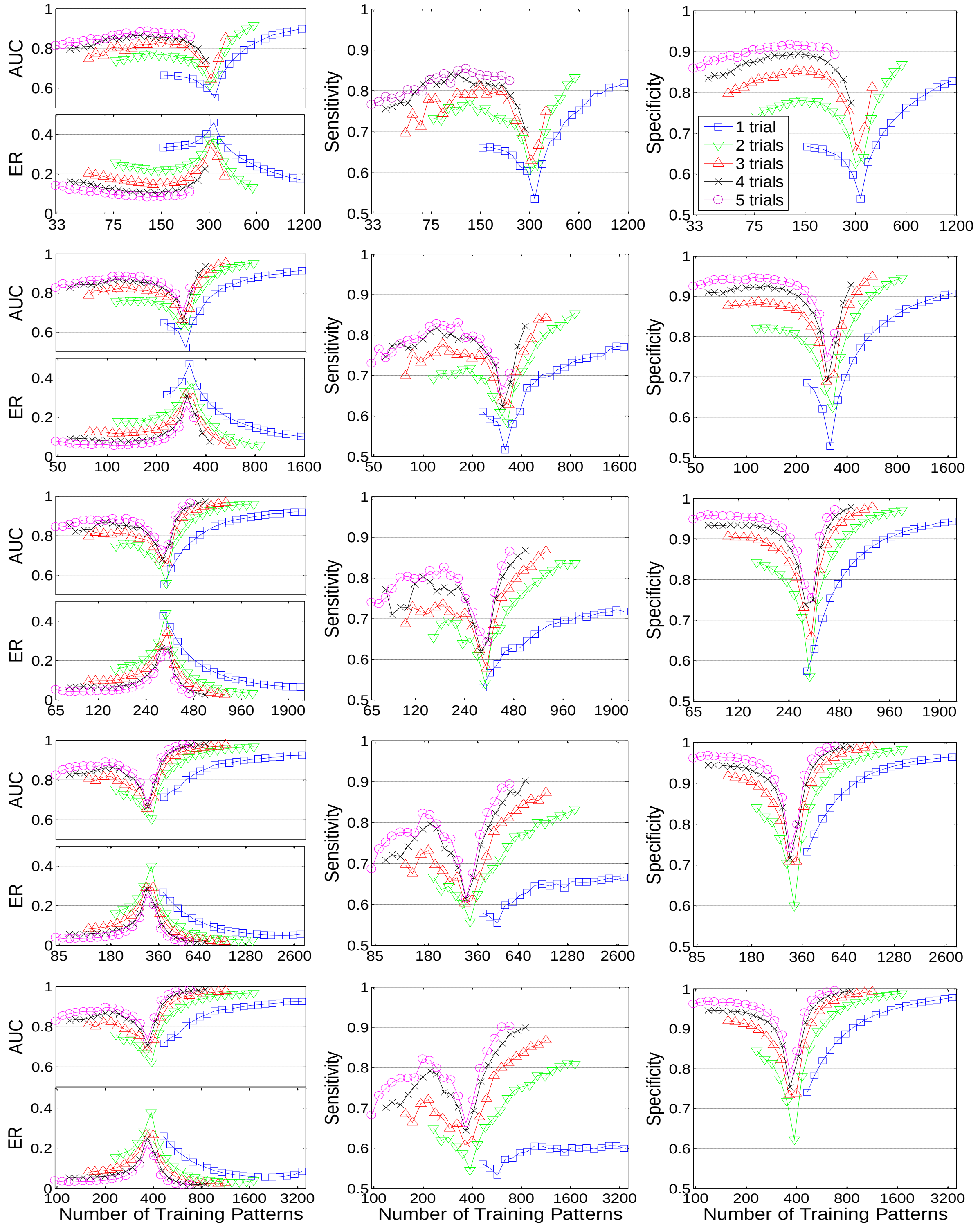


Figure 5. Graphics of performance estimates for the second set of experiments averaged over subjects vs. number of total training patterns obtained from the different subsets of experiments. From top to bottom: using one non target per target pattern, using two non target per target pattern, using three 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).

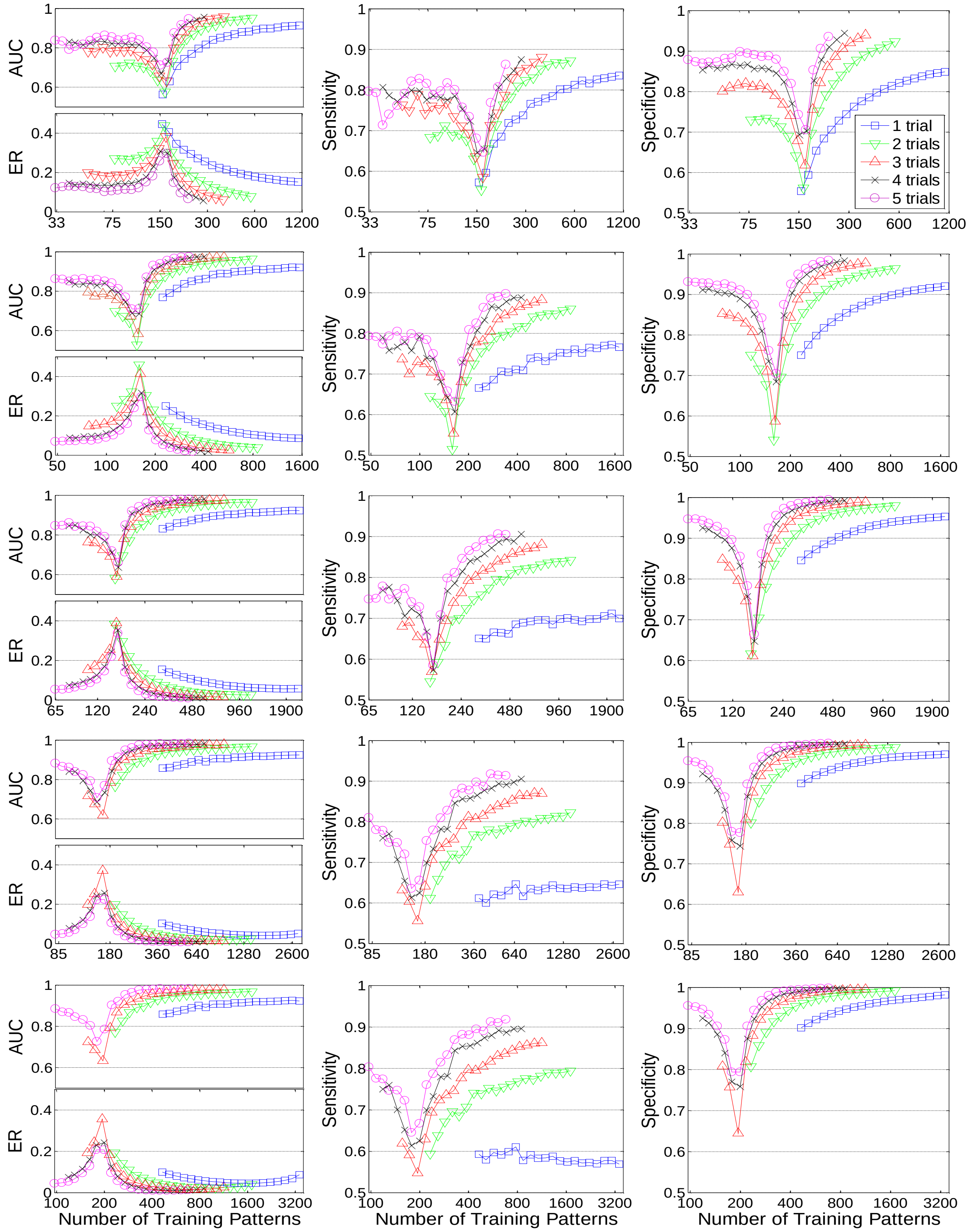


Figure 6. Graphics of performance estimates for the third set of experiments averaged over subjects vs. number of total training patterns obtained from the different subsets of experiments. From top to bottom: using one non target per target pattern, using two non target per target pattern, using three 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).

References

[1] Wolpaw J. R., Birbaumer N., McFarland D. J., Pfurtscheller G., and Vaughan T. M., 2002 Brain Computer Interfaces for communication and control. Clin. Neurophysiol., **113**(6): 767-791.

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

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

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

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

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

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

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

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

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

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

[12] Ramírez C. L., Bojorges Valdez E., Yañez Suárez O., Saavedra C., Bougrain L. and Gentiletti G. 2010 An open-access P300 speller database, *Fourth international BCI meeting*, Paper L-12, Monterrey California.

[13] Duda R., Hart P. and Stork D., 2000 *Pattern Classification* (2nd Edition), Wiley-Interscience.

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

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

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

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

[18] Raudys S., Duin R. P., 1998 Expected classification error of the Fisher linear classifier withpseudo-inverse covariance matrix. *Pattern Recognition Letters* archive vol. **19** Issue 5-6.