Proceedings of the 29th Annual International
Conference of the IEEE EMBS
Cité Internationale, Lyon, France
August 23-26, 2007.

Comparison of EEG Pattern Classification Methods for Brain-Computer Interfaces

N. S. Dias, M. Kamrunnahar, P. M. Mendes, S. J. Schiff, J. H. Correia

Abstract— **The aim of this study is to compare 2 EEG pattern classification methods towards the development of BCI. The methods are: (1) discriminant stepwise, and (2) Principal Component Analysis (PCA) –Linear Discriminant Analysis (LDA) joint method. Both methods use Fisher's LDA approach, but differ in the data dimensionality reduction procedure. Data were recorded from 3 male subjects 20-30 years old. Three runs per subject took place. The classification methods were tested in 240 trials per subject after merging all runs for the same subject. The mental tasks performed were feet, tongue, left hand and right hand movement imagery. In order to avoid previous assumptions on preferable channel locations and frequency ranges, 105 (21 electrodes×5 frequency ranges) electroencephalogram (EEG) features were extracted from the data. The best performance for each classification method was taken into account. The discriminant stepwise method showed better performance than the PCA based method. The classification error by the stepwise method varied between 31.73% and 38.5% for all subjects whereas the error range using the PCA based method was 39.42% to 54%.**

I. Introduction

Brain-Computer Interface (BCI) enables people to control a device with their brain signals [1]. BCI is expected to be a very useful tool for impaired people both in invasive and non-invasive implementations. Because the electroencephalogram (EEG) does not have as much accuracy as invasive recordings to detect user movement intention from primary motor cortex, recent studies have tried to use 2 distinct approaches. In the *operant conditioning* approach, the training load is on the subject [1]. The subject must learn to control a specific rhythm in order to produce the desired result on the device that he is controlling. The *pattern recognition* approach is suitable for less trained subjects. The user is instructed to perform distinct mental tasks that should be identified by the BCI system [2]. The features selected to discriminate the mental tasks are usually based on previous assumptions on frequency ranges and electrode placements commonly used to distinguish such mental tasks.

A discriminant stepwise procedure to discriminate EEG spatiotemporal patterns, in response to mental tasks, was proposed by the authors in [3]. A stepwise procedure first selects the variables with the most discriminant information and then canonical functions based on Fisher's Linear Discriminant Analysis (LDA) were used for classification. The results in [3] were encouraging, but a comparison test with other common pattern classification methods was not presented. The objective of this work is to compare the proposed method with a common EEG pattern classification approach: Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) jointly. A 4 class classification test was done for that purpose. The pattern classification is intended to provide the subjects, in further sessions, the ability to control a device with a minor training load. Hence, the discriminant algorithm generates functions that will predict class group membership of observations during feedback sessions. The subjects that participated in this study had no previous BCI experience. Neither frequency ranges nor electrode locations typically used in motor imagery tasks were pre-selected. Power ratios for frequency ranges of interest were used as features. Because the available variables are likely to be much more than is necessary to obtain the best possible linear discrimination, data dimensionality was reduced through a stepwise [3] in the proposed method or a Principal Component Analysis (PCA) procedure in the control method [4].

Three subjects were submitted to 3 sessions each, conducting mental tasks about movement imagery. The discrimination quality and group prediction error for each method was evaluated in all subject datasets.

II. Experimental Design

Three subjects, 20 to 30 years old, were submitted to 3 sessions of motor imagery. Each session had 80 trials. Each subject was instructed to perform one of 4 tasks in each trial. The tasks were tongue, feet, left hand and right hand movement imageries. Each trial was 8 s long. After the first 2 s a cue warned the subject to be prepared and 1 s later, a cue

N.S. Dias, P. M. Mendes and J. H. Correia are with the Dept. of Industrial Electronics, University of Minho, Campus Azurem, 4800-058 Guimaraes, Portugal (phone: 351-253-604703; email: ndias@dei.uminho.pt).

M. Kamrunnahar is with the Dept. of Engineering Sciences and Mechanics, The Pennsylvania State University, University Park, PA 16802

S. J. Schiff is with the Depts. of Engineering Sciences and Mechanics, Neurosurgery, and Physics, The Pennsylvania State University, University Park, PA 16802 (sschiff@psu.edu).

about the required mental task was presented to the subject. The subject should perform the task during the last 4 s.

A subset of the 32 available electrodes was used for classification, due to the presence of noise in some electrode signals. Hence, 21 electrodes (F7, F3, Fz, F4, F8, FC5, FC1, FC2, FC6, C3, Cz, C4, CP5, CP1, CP2, CP6, P7, P3, Pz, P4 and P8) according to the standard 10-20 system were used for feature extraction. All electrodes were referenced to linked earlobes. Data was digitized at 250 Hz and passed through a 6^{th} order (48 dB per octave) 0.5-30 Hz band-pass Butterworth filter. Data were visually inspected for artifacts after amplitude threshold and gradient artifact detection was applied. The trials that contained artifacts in the 2 to 8 s interval were marked and were excluded from the discriminant function analysis. Five frequency bins (10 Hz, 14 Hz, 18 Hz, 22 Hz and 26 Hz bin central frequencies, 4 Hz width bins) were considered for each channel. The data sets were epoched from 1 s before the cue to 4 s after the cue (5 s length). Each epoch was subdivided in 2 s time windows with 1 s overlap (4 time windows). Thus we used 4 time windows for classification error evaluation. The time window central points are 0 s, 1 s, 2 s and 3 s, with respect to the trigger point. The feature matrix of each time window is the ratio of the pre-filtered EEG signal power in one of these frequency ranges to the power in the broadband frequency range 0.5-30 Hz. Since 21 channels and 5 frequency bins were selected, 105 variables (features) were available for discrimination.

Two classification error measures were used for the comparison of methods. The plug-in error rate (PIR) is the ratio of misclassified observations to the total observations when the discrimination functions are extracted from all the data observations. The leave-one-out error rate (LOOR) is the ratio of the misclassified observations to the total observations when one observation at a time is left out of the discrimination function generation and its group membership is predicted by those functions.

III. Classification Methods

A. Discriminant Stepwise

A discriminant stepwise method was used to decrease data dimensionality [5]. The original feature matrix of each subject has 160-220 multivariate observations (observations in rows). Each observation is described by 105 variables. Each feature is a power ratio of a specific channel (out of the 21 available) for one of the frequency ranges mentioned in the previous section. This method is based on a multivariate canonical discrimination technique that was first developed by Fisher [6] in order to quantify the static taxonomic classification of plant species. A more robust approach on spatiotemporal EEG patterns discrimination [7] was used.

The discrimination was performed on a feature matrix Y, which was previously formatted by the stepwise procedure. The canonical discrimination functions Z_i are the result of a linear transformation of original data Y according (1). The discrimination coefficients of each i^{th} canonical discrimination function are denoted by the columns of b_i^T .

$$Z_i = Y b_i^T \tag{1}$$

Covariance matrices of the Y matrix were calculated for the whole dataset Ψ_{total} and within each group Ψ_{within}. For any linear combination Z_i the separation between groups implies that the $\Psi_{between}$ in (2) should be emphasized with respect to Ψ_{within}.

$$\Psi_{between} = \Psi_{total} - \Psi_{within} \tag{2}$$

Upon normality assumption, each multivariate observation vector in Y has a transformed vector z with mean u and normal p-variate distribution $f(z)$. Prior probabilities π_j were determined by the ratio of observations in group j to the total observations (N). The group membership prediction was based on the posterior probability π_{jz} in (3) as the probability that the data of a given value z came from group j of n groups. The $exp[q(z)]$, for $q(z)=u_j^T z-1/2 u_j^T u_j+\ln\pi_j$, was used as a good approximation of $\pi_j f(z)$ [5]. The highest π_{jz} value for $j=1,\ldots,4$ was the predicted group membership for posterior calculations.

$$\pi_{jz} = \frac{\pi_j f_j(z)}{\sum_{k=1}^{n} \pi_k f_k(z)}, k = 1,\ldots,n \tag{3}$$

Discrimination quality was accessed through 3 different tests. A robust method for quality testing is to leave one multivariate data point out of the discriminant function calculation and then test it for predicted group classification given its posterior probability. In order to test the significance of discrimination, we used a normal theory method that analyses the eigenvalues of the coordinate's transformation matrix [3]. After calculating the log likelihood ratio as $LLRS = N\sum_{i=1}^{m}\ln(1+\lambda_i)$ for m canonical discriminators, where λ_i are their eigenvalues, the Wilks' statistic was used as $W = \exp[-LLRS/N]$. A good discrimination yields large eigenvalues and W becomes small. Small eigenvalues and W values close to 1 are typical for poor discriminations. W is chi-square distributed and confidence limits were calculated for discrimination significance [5]. Since the W statistic is based on the assumption of normal distribution of data variables, which may not be the case, a bootstrap method was therefore used as an alternative method of testing discrimination quality. It randomly permutes the group labeling of each multivariate data point and re-tests the goodness of fit [7]. The permutation

number was limited to 1000.

Although every variable in the data set has between groups discriminative information, the preferred criterion for discrimination methods comparison was the LOOR. It gives a more robust measure of the real-time performance of our classification method. Additionally, a moderate small ratio of number of observations to variables can make the classification unstable in the case of over-fitting. A trade-off between number of variables and *W* discrimination value must be sought. In order to achieve good quality discrimination, we seek to optimize which of the 105 variables are best to include in the *Y* matrix.

The first step of this method is to select a first variable to start with and then add new variables in the order of decreasing discrimination ability. The function that best discriminates the multivariate data observations for all 105 variables is determined. The likelihood between the discriminant function and each variable is given by their correlation, also called a structure coefficient [8]. From (1), it can be calculated using the correlation between each column of *Y* and the transformed observations of *Z*'s first column. The largest absolute value of the correlation indicates the first variable to be selected and its observations vector will be the *Y* at this step. Then discrimination functions were determined as well as the starting *W* value. The second variable to be selected is the one that jointly with the first one promotes the largest decrease in *W* once new discrimination functions are calculated. Iteratively, it adds new variables according to their discrimination ability in decreasing order. The LOOR is calculated every time one variable is added. Once all variables were added into *Y* by discrimination ability decreasing order, the subset of variables that reached the lowest LOOR was selected if the discrimination between groups was significant by the bootstrap method. Once this procedure was finished, we have an optimized variable set and new canonical discriminant function available to predict group membership on training data as well as test data (feedback sessions).

B. PCA+LDA

This method uses PCA as a data dimensionality reduction step and then applies LDA on the selected components. The objective of PCA is to identify a small number of dimensions that provide a succinct and meaningful interpretation of the structure underlying the data [4]. The original feature matrix *Y* is composed of all the 160-220 (artifact free) available multivariate observations for all the 105 variables. A Singular Value Decomposition (SVD) of the *Y* outputs 3 matrices. *U, S* and *V*. *V* is the eigenvector orthogonal matrix. *S* is a diagonal matrix with the eigenvalues. *U*×*S* is the component matrix. The component matrix is obtained from *Y*×*V,* which is the projection of the original data over the eigenvectors dimensional space. The equality in (4) explains how the original data can be generated back from the SVD output matrices.

$$Y = U \times S \times V^T \tag{4}$$

Although the components extracted from the SVD (columns of U matrix) are already organized by decreasing order of total variance accounted for, those may not be the most important for population discrimination – indeed this decomposition is optimized for orthogonality rather than discrimination between groups. In order to try to compensate for this and take the data group structure into account, attention must be paid on the highest scores that account for across group variance (AGV) [4], instead of looking for the scores that account for the total variance (eigenvalues in *S*).

From (2) and $V^T \Psi_{Total} V = S^2$ where the columns of *V* are the component loadings of Ψ_{Total} corresponding to the eigenvalues of the diagonal values of *S,* the total amount of information provided by the i^{th} component is given by (5).

$$\lambda_i = v_i^T \left(\Psi_{Within} + \Psi_{Between} \right) v_i \tag{5}$$

The AGV accounted for by the i^{th} component is given by

$$AGV_i = \frac{v_i^T \Psi_{Between} v_i}{\lambda_i} \tag{6}$$

This AGV measure was used to rank every component. The original data in *Y* were projected over the eigenvectors axes corresponding to the components with best AGV ranking. The component selection criterion was 99% of the total AGV. The time series resulting from *Y*×*V* (where the columns of *V* are truncated, keeping the eigenvectors matching the previously selected components) was used for LDA discriminant function calculation and tested for classification according to the previous subsection details. The classification error for group membership prediction, using each subject's selected components as predictors, was used for comparison.

IV. RESULTS

The classification results for method comparison in each subject data are presented as PIR and LOOR values for each time window. The t-Test for both methods' classification error mean equality (H_0 null hypothesis) is presented on Table I. PCA and Stepwise methods are the t-Test sample groups and their time window classification errors are the observations. Each column in Table I have the results of the t-Test for H_0 hypothesis (PIR or LOOR) in a specific subject data. *Bonferroni* corrections were used on confidence intervals

(α=0.5) of multiple comparisons for the three subjects.

Although, the null hypothesis probability for PIR error is quite significant for 2 out of the 3 subjects (0.003, 0.452 and 0.115 for EM, FF and JC respectively), the LOOR values are significantly different in both methods for all subjects (0.004, 0.015 and 0.003 respectively).

Fig. 1 depicts both method classification errors for PIR and LOOR measures. The lowest error rates were the PIR with 11% misclassified observations for EM subject, 11.06% for FF subject and 15.67% for JC subject. The lowest LOOR error rates were 31.73%, 38.5% and 35.02% for FF, EM and JC subjects respectively.

TABLE I
PAIRED T-TEST FOR EQUALITY OF METHODS' CLASSIFICATION ERROR MEANS

	PIR			LOOR		
	EM	*FF*	*JC*	*EM*	*FF*	*JC*
Pearson Correlation	0.57	0.76	0.76	0.54	0.92	0.99
t Stat	-8.506401	-0.862011	-2.198467	8.013944	5.028126	8.778204
P(T<=t) two-tail	0.003412	0.452077	0.115337	0.004056	0.015157	0.003114
t Critical two-tail	4.856657	4.856657	4.856657	4.856657	4.856657	4.856657

The hypothesis of classification error (PIR and LOOR) equality for both PCA and Stepwise methods was tested with a paired t-Test (tow-tail approach) for each subject (EM, FF and JC). Multiple comparison tests were done with *Bonferroni* corrections α=0.05.

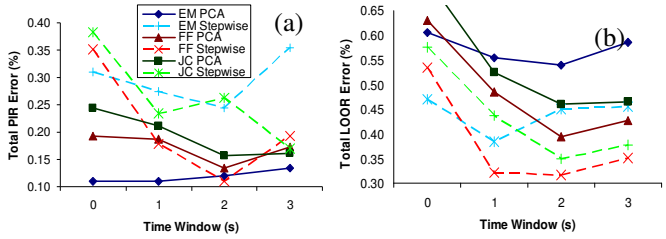

Fig. 1. (a) Plug-in classification error rates for all the 3 subjects (EM, FF and JC) in both PCA and Stepwise Discriminant Analysis at each time window. (b) Leave-one-out classification error rates for all the 3 subjects (EM, FF and JC) in both PCA and Stepwise Discriminant Analysis at each time window.

V. DISCUSSION AND CONCLUSIONS

PIR has lower values than LOOR in all cases. Additionally, the PIR classification showed very similar results for both Stepwise and PCA methods. These facts can be due to over-fitting of the classification algorithms. PIR is more susceptible to this effect. Furthermore, the leave-one-out error rates were significantly better for the Discriminant Stepwise method than for the PCA+LDA method.

The FF subject achieved lower classification errors than EM and JC subjects. Although the best time window for classification differed from subject to subject, the classification error in the first time window was generally the worst. That is supported by the fact that the task order was randomized along data recording runs. Hence, the subject had no clue of the task to perform during the first 2 s time window. Most of the best classification time points for LOOR values were in the 1 s to 3 s after trigger time window, which is a time period free of evoked potentials and gives a reasonable delay for task performance initiation.

The group membership prediction error values presented above were intended for classification methods comparison. However, the error rates are not low enough as required for effective 4 groups BCI online operation. In future work, to address this issue, session data recorded with more than one EEG cap placement should not be used on discriminant functions extraction. Slightly different electrode locations may induce data variability. A larger number of electrodes and different data filtering other than frequency ratios (e.g. event-related synchronization) should be considered in order to get a lower error rate. A larger subject population should be considered to double check the hypothetically better performance of the Discriminant Stepwise method than the PCA+LDA method.

ACKNOWLEDGMENTS

This work was supported by Center Algoritmi, N. S. Dias is supported by the Portuguese Foundation for Science and Technology under Grant SFRH/BD/21529/2005. The authors would like to thanks the participation in the tests of Fernando Ferreira, Eurico Martins and Jorge Cardoso, Biomedical Engineering students. S. J. Schiff and M. Kamrunnahar were supported by a Keystone Innovation Zone Grant from the Commonwealth of Pennsylvania, and S. J. Schiff by NIH grant K02MH01493.

REFERENCES

[1] J.R.Wolpaw, D.J.McFarland, and T.M.Vaughan, "Brain–Computer Interface Research at the Wadsworth Center," *IEEE TRANSACTIONS ON REHAB. ENGINEERING*, vol. 8, no. 2, pp. 222-226, June 2000.

[2] C.Guger, H.Ramoser, and G.Pfurtscheller, "Real-Time EEG Analysis with Subject-Specific Spatial Patterns for a Brain-Computer Interface (BCI)," *IEEE TRANSACTIONS ON REHABILITATION ENGINEERING*, vol. 8, no. 4, pp. 447-456, 2000.

[3] N.S.Dias, M.Kamrunnahar, P.M.Mendes, S.J.Schiff, and J.H.Correia, "Customized Linear Discriminant Analysis for Brain-Computer Interfaces," *CNE '07 IEEE/EMBS*, Vol., Iss, pp. 430-433, 2-5 May 2007.

[4] W.R.Dillon, N.Mulani, and D.G.Frederick, "On the Use of Component Scores in the Presence of Group Structure," *JOURNAL OF CONSUMER RESEARCH*, vol. 16, pp. 106-112, June 1989.

[5] B.Flury, *A First Course in Multivariate Statistics*. Springer, 1997.

[6] R.A.Fisher, "The use of multiple measurements in taxonomic problems,", *ANNALS OF EUGENICS*, 7 ed 1936, pp. 179-188.

[7] S.J.Schiff, T.Sauer, R.Kumar, and S.L.Weinstein, "Neuronal spatiotemporal pattern discrimination: The dynamical evolution of seizures.,", *NEUROIMAGE*, 28 ed 2005, pp. 1043-1055.

[8] Bowling J., "The importance of structure coefficients as against beta weights," *ANNUAL MEETING OF MID-SOUTH EDUCATION RESEARCH ASSOCIATION*, New Orleans, 1993.