Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Analysis of P300 Classifiers in Brain Computer Interface Speller

H. Mirghasemi, R. Fazel-Rezai, *Senior Member, IEEE*, and M. B. Shamsollahi, *Member, IEEE*

***Abstract*— In this paper, the performance of five classifiers in P300 speller paradigm are compared. Theses classifiers are Linear Support Vector Machine (LSVM), Gaussian Support Vector Machine (RSVM), Neural Network (NN), Fisher Linear Discriminant (FLD), and Kernel Fisher Discriminant (KFD). In classification of P300 waves, there has been a trend to use SVM classifiers. Although they have shown a good performance, in this paper, it is shown that the FLD classifiers outperform the SVM classifiers. FLD classifier uses only ten channels of the recorded electroencephalogram (EEG) signals. This makes them a very good candidate for real-time applications. In addition, FLD approach does not need any optimization similar to other methods. In addition, in this paper, it is shown that the efficiency of using Principal Component Analysis (PCA) for feature reduction results in decreasing the time for the classification and increasing the accuracy.**

I. Introduction

Brain computer interface (BCI) is a system that creates a direct channel between computer and brain [1]. Among various BCI systems, in this paper, we consider the P300 speller. The P300 speller paradigm is based on the nature of P300 component of electroencephalogram (EEG). The P300 component is a positive peak in EEG at about 300 ms after an unusual event or stimuli. Within a P300 speller, user is shown a 6 by 6 matrix, containing 36 symbols. All rows and columns of this matrix were successively and randomly intensified. Intensifications of each row and column for each character are repeated for 15 times (15 trails) as shown in Fig. 1 [2]. If row or column of desired character is highlighted, the related EEG (epoch) will contain the P300 component. Therefore, we are able to find a way to distinguish the epochs which have P300 component from the epochs which have not P300 component and to detect which row or column is related to desired character. Therefore, the target character can be identified and the subject can spell different characters.

Manuscript received April 24, 2006.

R. Fazel-Rezai is with the Brain Image and Signal Processing (BISP) Laboratory, Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, MB R3T 5V6 Canada (corresponding author: phone: 1-204-474-9490; fax: 1-204-261-4639; email: fazel@ee.umanitoba.ca).

H. Mirghasemi is with the BDP laboratory, Department of Electrical Engineering , Sharif University of Technology, Tehran, Iran (hmirghasemi@ee.sharif.edu)

M. B. Shamsollahi is with the BDP laboratory Department of Electrical Engineering , Sharif University of Technology, Tehran, Iran (mbshams@sharif.edu)

II. Materials and Methods

A. Data acquisition

In this work, we used the dataset from BCI 2003 competition. This dataset was selected, because the results of this paper can be compared with the results of other works which were presented in different papers. Dataset contains three sessions and each session contains several characters. Similar to competition condition, we used first two sessions as training data (i.e. the data used to train classifier) and the third session was used as testing data.
Data were recorded from 64 electrodes; however we did not use all of them. We applied our methods to two cases with different number of channels from EEG signals. For first case three channels (Cz, Pz, Fz), and for second case ten channels (Fz, Cz, Pz, Oz, P3, P4, C3, C4 PO7 and PO8) were used. The locations of channels are defined based on 10-20 standard [9] as shown in Fig. 2.

B. Preprocessing

All the data was filtered with a bandpass digital filtering (0.5-30 Hz) and normalized to an interval of [-1 1].

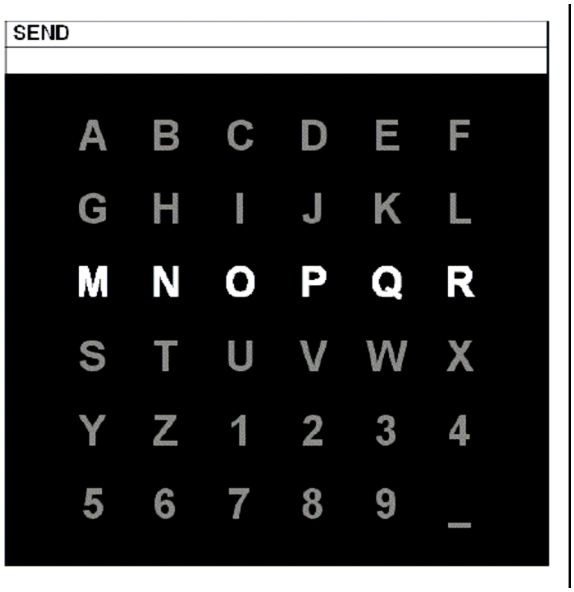

Fig. 1. The P300 speller paradigm [2]

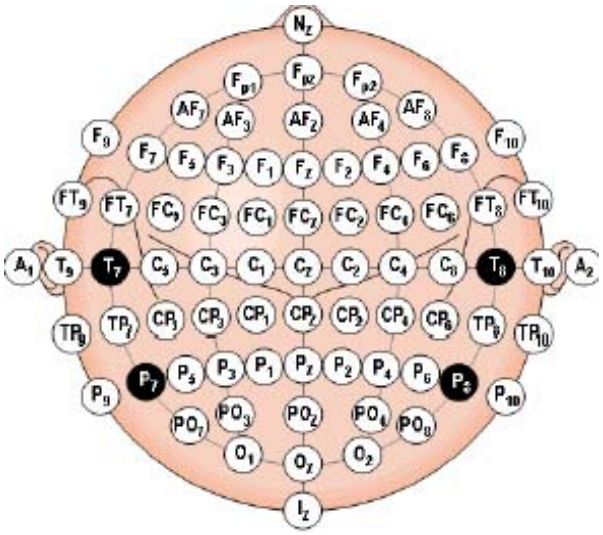

Fig2. Electrode designation in 10_20 system [9]

C. *Feature reduction*

Since the EEG signal with the P300 component has a distinct temporal pattern, we used the value of samples of filtered data as feature. However, we were interested to decrease the time of classification; therefore, we implemented feature reduction with Principal Component Analysis (PCA). We decreased the dimension of feature input from 144 (number of samples of each epoch during 600 ms after stimuli onset) to 21 features. The results when PCA was used were compared with those when PCA was not used.

D. Classification

We first trained our classifiers with labeled training data and then applied trained classifiers to unlabeled testing data. Training data are labeled with 1 and -1 for P300 absence and presence, respectively. Some of our classifiers need parameters which should be determined using optimization. These classifiers include: KFD, LSVM, RSVM, and NN. To determine these parameters, we changed these parameters in a specific range and assessed the accuracy of classification. Parameters were set equal to values which maximize the accuracy.

After training the classifiers, testing data should be applied to them. First, we applied the classifier to the first trail of each character. The values of decision function (decision values) for 6 rows and 6 columns are regarded as scores of each row and column. Then we performed this procedure for all of the trials (when 3 channels were used, all trials and when 10 channels are used, first four trials were used.). Scores of all trails for each row or column were accumulated. The row/column with the highest score was selected as row/column which contains the P300 wave.

III. CLASSIFIERS

The following classifiers were applied to BCI data.

A. Fisher Linear Discriminant (FLD)

Fisher linear discriminant is a linear classifier. If $X_1 = \{x_1^1, x_2^1, ..., x_{N_1}^1\}$ and $X_2 = \{x_1^2, x_2^2, ..., x_{N_2}^2\}$ are samples from two classes, FLD finds an optimal w to maximize the difference between two classes. Let $m_i = \frac{1}{N_i}\sum_{i=1}^{N_l} x_i^l$, $l = 1,2$ be the center of each class, we can define the $S_w = \sum_{l=1}^{2}\sum_{i=1}^{N_l}(x_i^l - m_l)(x_i^l - m_l)^T$ as the within class scatter matrix and $S_B = (m_1 - m_2)(m_1 - m_2)^T$ as the between-class scatter matrix. Finding w is equals to maximizing the class separability: $F(w) = \frac{w^T S_B w}{w^T S_w w}$. A classical approach is to set the w equals to $S_w^{-1}(m_1 - m_2)$. To find linear discriminant function $f(x) = \langle w.x \rangle + b$, we should determine b from the equation $f(m_1) = -f(m_2)$. Discriminant function is obtained after determining w and b [4].

B. Kernel Fisher Discriminant (KFD)

Using Mercer kernels, any linear algorithm in the form of dot product can be performed directly in high dimensional feature space to drive a non linear discriminant. Using this technique, FLD can be generalized to its nonlinear form, i.e., KFD. This results in a nonlinear discriminant function:

$$f(x) = \sum_{i=1}^{N} \alpha_i K(x, x_i) + b \qquad (1)$$

Coefficients α_i can be obtained by solving an eign-problem.

In our work, we used kernel Gaussian function as the kernel:

$$k(x, x_i) = \exp\left(\frac{\|x - x_i\|^2}{2\sigma^2}\right) \qquad (2)$$

C. Support Vector Machine (SVM)

SVM tries to minimize an upper bound of generalization error, as opposed to other classifiers which try to minimize the empirical error. Let the input is a set $\Gamma = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ of binary labeled $y_i \in \{-1, 1\}$ training vector. Then the hyperplane separates the data if and only if:

$$y_i((w.x_i) + b) \geq 1 \quad \forall i \qquad (3)$$

SVMs maximize the distance between two classes in order to find the optimal hyperplane with the best generalization capabilities. In linearly separable case, this equals to

minimizing $\frac{1}{2}\|w\|^2$ subject to (3). However, for nonlinear separable case, we should introduce the slack-variables ε_i to modify constraints to looser constraints. Also, a penalty (regularization parameter C) is incurred for misclassification:

$$\min \quad \frac{1}{2}\|w\|^2 + C\sum_i \varepsilon_i \tag{4}$$

$$\text{subject to:} \quad y_i(w.x_i + b) \geq 1 - \varepsilon_i,\ \varepsilon_i > 0 \ \ \forall i \tag{5}$$

By forming the lagrangian and solving the dual problem, it can be interpreted as following:

$$\max \quad \sum_i \alpha_i - \frac{1}{2}\left(\sum_{i,j} \alpha_i \alpha_j y_i y_j x_i . x_j\right) \tag{6}$$

$$\text{subject to:} \quad 0 \leq \alpha_i \leq C \tag{7}$$

The α_i is lagrangian multiplier, and for each of training samples there is one lagrangian multiplier. The training sample whose lagrangian multiplier is not zero is called support vector. Solving this problem (QP problem) [5] results:

$$w = \sum_i^{N_s} y_i \alpha_i x_i \tag{8}$$

where N_s is the number of support vectors [5]. Similar to the FLD, the SVM can be generalized to nonlinear decision surfaces to make separation more likely for nonlinear separable data. This can be achieved implicitly by using the different type of symmetric functions $K(x, y)$ instead of the ordinary scalar product. In the SVM, when linear discriminant function is used, SVM is called LSVM and if kernel Gaussian function is used as the kernel, SVM is called RSVM. We applied both LSVM and RSVM to our data.

D. Neural Network (NN)

A NN is composed of simple elements which operating in parallel and its function is determined by network structure. NNs can be trained to perform a particular function by adjusting the values of the connections between elements.

Perceptron is the simplest NN, and it can classify linearly separable data. To deal with non-separable data, we can extend the simple perceptron to a Multi-Layer Perceptron NN (MLPNN), which includes at least one hidden layer of neurons [6]. In this work, 2-Layer Perceptron Neural Network (2LPNN) and 3-Layer Perceptron Neural Network (3LPNN) were used.

For 2LPNN, tan-sigmoid transfer and linear transfer function were used for first hidden layer and second layer. To train the network, we used gradient descent learning approach. Also, to improve the performance of learning, we allowed change in the learning rate during learning.

For 3LPNN, we used tan-sigmoid transfer function for first two layers and linear transfer function was used for last layer. The same learning approach was chosen for 3LPNN.

IV. Results

Our work consists of three steps. First, we filtered data using a bandpass filtering. Then, we used PCA for feature reduction and finally we classified testing data with the trained classifier. Also, we did above steps without feature reduction to monitor the effect of feature reduction on the final results. All of these steps were conducted at two runs. In the first run, we used 3 channels (Cz, Pz, Fz), and in second run, ten channels (Fz, Cz, Pz, Oz, P3, P4, C3, C4 PO7 and PO8) of the EEG data were used.

Table I shows the results of the first run. As it can be seen the highest accuracy was achieved with FLD classifier and when PCA was used to reduce feature space dimension. Using this approach, 30 characters of the 31 characters were predicted correctly.

In table II, the results of using ten cannels are shown. The perfect performance (the 100% accuracy) was obtained with several methods. In comparison to winners of BCI 2003 competition which achieved the 100% accuracy after five trails, our approaches yielded prefect performance after four trials. This increases the transfer rate.

In Fig.3 and Fig.4, we illustrated that how increasing the numbers of channel can affect the accuracy for different classification methods.

V. Conclusions

A. Trade- off between number of channels and bit rate

EEG acquisition is a delicate work which needs high accuracy. The difficulty and cost of EEG recording has a direct relationship with number of channels. As we can see in Fig. 2 and Fig. 3, with increasing the number of channels we can achieve the prefect accuracy and higher bit rate. The bit rate is important for online situation. Also, we should note that by increasing the number of channels, the required time for data classification is increased. In FLD method, we used only 10 channels to achieved 100% accuracy.

B. Feature reduction

Feature reduction is a method to reduce the dimension of feature space to decrease the time of classification. Since feature reduction reduces the dimension of feature space, it sometimes reduces performance. Therefore, it is important to use a feature extraction method which does not decrease the accuracy. In the all of the cases that we examined, the PCA increased the speed of classification and the accuracy.

C. Classifier

In many P300 speller systems, the SVM has been proposed as the best classification method. For example both winners of BCI 2003 and BCI 2005 used SVM for classification. Although, in our work the SVM shows good

performance, but we showed that the FLD can achieve better performance. For example, with using FLD as classifier and the PCA, with only three channels we could achieve an acceptable accuracy in classification (30 of 31 characters were predicted correctly). Another point which should be considered is that the FLD is the only classifier among these classifiers which does not need optimization. The SVM, KFD and neural network have parameters which should be determined with optimization. Since the result of optimization is based on the data, these parameters should be obtained for each dataset. Also, among these classifiers, the time needed to train and test a classifier is minimum in FLD classifier. Multilayer perceptron yields worse result. In addition, multilayer perceptron needs more time to be trained and unknown parameters are more than SVM and KFD. Increasing the number of layer form 2 to 3 not only increased time for training but also decreased the accuracy.

In conclusion, the results from the FLD are very encouraging when used with PCA for feature reduction. This method can bring the P300 speller to realm of practicality for many real time applications.

D. Comparison with results of BCI2003

We could achieve the 100% accuracy with only ten channels and after 4 trials. The winners of BCI2003 [3,8] used 5 trials to achieve this accuracy.

TABLE I. THE NUMBER OF WRONG PREDICTED CHARACTER AMONG 31 CHARACTERS FOR THREE-CHANNEL DATA. THE FIRST ROW SHOWS WHEN THE PCA WAS USED FOR FEATURE REDUCTION AND THE SECOND ROW SHOWS WHEN NO METHOD WAS USED FOR FEATURE REDUCTION.

	FLD	KFD	LSVM	RSVM	MLP2	MLP3
PCA	**1**	**2**	**2**	**2**	**2**	4
None	3	3	3	2	2	5

TABLE II. THE NUMBER OF WRONG PREDICTED CHARACTER AMONG 31CHARACTERS FOR TEN-CHANNEL DATA. THE FIRST ROW SHOWS WHEN THE PCA WAS USED FOR FEATURE REDUCTION AND THE SECOND ROW SHOWS WHEN NO METHOD WAS USED FOR FEATURE REDUCTION.

	FLD	KFD	LSVM	RSVM	MLP2	MLP3
PCA	**0**	**0**	**0**	**0**	**1**	2
None	0	0	2	2	1	3

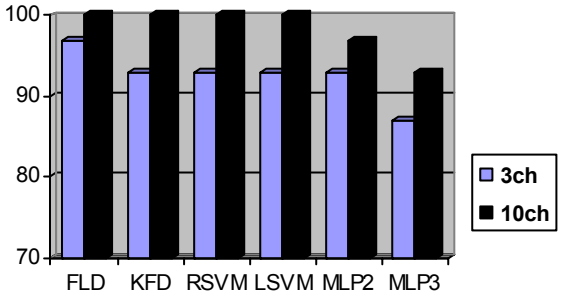

Fig3. The effect of increasing number of channels when feature reduction is used

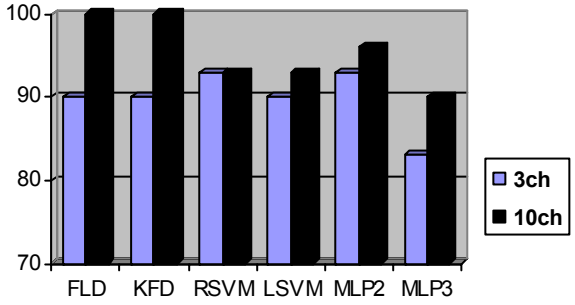

Fig 4.The effect of increasing number of channels when feature reduction is not used

REFERENCES

[1] J. R. Wolpaw, N.Birbaumer, D. J. MackFarland, G. Pfurstcheller and T. M. Vaugan, "Brain Computer Interface for communication and control", *Clin. Neurophyiol.,* vol..11 3, pp.767-791, 2002..

[2] L. A. Farwell and E. Donchin, "Talking of top of your head: Toward a mental prostethis utilizing event-related brain potentials, *Electroencephalogr and Clin Neurophy,* vol. 70, pp.510-523, 1988.

[3] M. Kaper, P. Meinicke, U. Grosskathoefer, T. Lingner, H. Ritter, "Support Vector Machines for the P300 speller paradigm" , *IEEE Trans. Biomed. Eng.* Vol. 51, no. 6, pp. 1073-1076, 2004.

[4] S. Mika, G. Ratsch , and K.-R. Muller, "Fisher Discriminant Analysis with Kernels" , *Neural Networks for Signal Processing,* vol. 4, pp. 41-48, 1999.

[5] C. J. C. Burges. (1998) A tutorial on support vector machines for pattern recognition. *Data Mining Knowledge Discovery* [Online], vol. 2, pp. 121-167. Available: citeseer.nj.nec.com/burges98tutorial.html.

[6] S. Theodiris and K. Koutrombas, *Pattern Recognition.* Academic Press, pp. 77-88, 1999

[7] V. N. Vapnic, *The nature of Statistical Learning Theory.* New York: Springer 2005.

[8] Xu N, Gao X, Hong B, Miao X, Gao S and Yang F, "BCI Competition 2003_Data Set IIb: Enhancing P300 Wave Detection Using ICA-Based Subspace Projections for BCI Applications", *IEEE Transaction On Biomedical Eng,* vol. 51, pp. 1067-1072, 2004.

[9] J. Malmivuo and R. Plonesy, *Bioelectromagneticm,* Oxford, Newyork: Oxford University Press, 1995, [online]. Available: http://butler.cc.tut.fi/~malmivuo/bem/bembook/.