Feature Selection for Brain-Computer Interface with Six Motor Imagery Tasks Using Orthogonal Forward Selection

Aryan Mobiny*, Ehsan Arbabi**, and Tooraj Abbasian Najafabadi***
School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran
*aryan.mobiny@ut.ac.ir
**earbabi@ut.ac.ir
***najafabadi@ut.ac.ir

Abstract: *In brain-computer interfaces (BCIs), finding the most effective features among all the extracted features plays an important role to reach a proper classification. In this study, orthogonal forward selection (OFS) method has been used to select the best features extracted from the brain signals of 4 subjects during performing 6 motor imagery tasks. The results show that kurtosis, skewness and Hjorth parameters are among the most effective features, comparing to the other features extracted from the brain signals. Finally, by using Bayesian classifier, in average 91.62% correct classification rate (CCR) has been achieved for separating 6 different classes.*

Keywords: Brain-computer interface (BCI), electroencephalogram (EEG), orthogonal forward selection (OFS), Bayes optimal classifier

1. Introduction

Spinal Cord Injury (SCI) and Amyotrophic Lateral Sclerosis (ALS) are among common diseases which cause motor disabilities. In severe cases, people with motor disabilities are unable to perform any movement by their bodies. These people may be extremely limited and meet severe problems for communicating with their environments. Therefore, providing an alternative way for those individuals, who have severe motor disabilities, is an important issue. Brain–computer interfaces (BCIs) offer such an alternative way [1]. In fact, in BCI systems, the recorded brain signals are processed in order to interpret the mental tasks done by a user. Later, based on this interpretation, a device, such as a wheelchair, can be controlled.

Brain computer interfaces are based on either invasive or noninvasive signal recording. In the invasive systems, the recording electrodes are implanted into the premotor or motor frontal areas or into the parietal cortex. On the other hand, the noninvasive systems are mostly working by using electroencephalograms (EEGs) recorded from the subject's scalp [2].

Many noninvasive BCI systems are based on recognizing different imagery movements during mental activities of users [3-5]. In most cases, the BCI systems have been designed to recognize two to three mental tasks [6-8], which can limits the number of distinguishable orders given by users. In this study, a noninvasive BCI system based on six imagery movements has been used. These six mental tasks have been selected in a way that they can easily be imagined, they do not lead to user fatigue, and also they have enough separability from each other.

Four subjects have performed the planed six mental activities and their brain signals (EEG) have been recorded. After noise removal, different features have been extracted from the recorded brain signals using different frequency and time-domain methods. The most effective features have been selected, using orthogonal forward selection method. The main characteristic of this method is that the features are decorrelated in the orthogonal space and they can be evaluated independently [9]. Finally, in order to separate the signals related to six different mental tasks, the extracted features have been classified using Bayesian classifier.

The remaining of the paper is as follows: In section 2, data collection and processing are discussed in detail. The results are reported and discussed in section 3. Finally, the conclusion is drawn in section 4.

2. Methods

2.1 Data Acquisition

The brain signals (EEG) used in this study have been recorded by NRSIGN3840 system (with sampling rate of 500 Hz) at School of Electrical and Computer Engineering, University of Tehran. These data are related to four subjects, including two males and two females, all right-handed and between 23 to 25 years old. Their EEG signals have been recorded during 6 imagery movements including imagery movements of right hand, left hand, face, feet, tongue and head. A 20-electrode EEG cap based on the international 10-20 system was used to collect the data, referenced to an electrode placed on the forehead.

The subjects, whose EEG was to be recorded, sat on a chair in front of a monitor and tried to perform the desired mental tasks in accordance with some cues given on the monitor. Ten runs were carried out and data collection from each subject lasted half an hour. The subjects could have one or two minutes break between

two consecutive runs. Each run included three trials and each trial took ten seconds (Fig. 1).

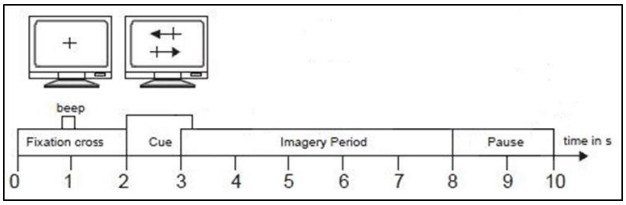

Fig. 1: Timing of each trial

As it is illustrated in Fig. 1, from second 0 to 2, a cross symbol (accompanying with a short tone of "beep") is shown on the screen to make the subject concentrate on the center of the screen. Then, a cue appears on the screen, telling the subject which mental activity he or she must think about. From second 3 to 8 the subject starts performing the relevant mental activity asked in the previous step.

2.2 Pre-processing

Generally, brain signals have their major energy in the frequency range between 0.5-100 Hz (especially below 60 Hz) [10]. This frequency range includes alpha and beta bands which are among the most important bands for imagery movements [2]. Hence, raw signals have been filtered in a 0.5-100 Hz band by 50^{th} order linear-phase FIR filter. It should be noticed that the sampling rate of the recording system (i.e. 500 Hz) is large enough to prevent aliasing effects. Moreover, a 50 Hz Notch filter has been used to remove the power line interference.

We have used Independent Component Analysis (ICA) for removing the artifacts caused by eyes (i.e. eyes movements or blinking). Several algorithms have been developed to implement ICA method. In this work, we applied the FastICA algorithm. FastICA works based on a fixed-point iteration scheme to maximize the non-gausssianity (measured by approximating the negentropy), and estimate the individual independent components [11].

After removing noises and artifacts, data segmentation has been applied on the signal. Each segment of the signal is 1 second long (given that sampling frequency of the signal is 500Hz, each 500 samples considered as a segment) and overlaps by 50% with the adjacent segment. Thus, the stationary problem of EEG signals can be ignored by choosing short segments [1].

2.3 Feature Extraction

In this research, by using 19 channels devoted to data recording and applying different methods, totally 72 features have been extracted from each channel. These features have been explained as follows.

Spectral analysis can provide important features and they are widely used in BCI systems. In order to estimate the power spectral density (PSD) of the signals, we have applied an autoregressive method solved by the Yule-Walker algorithm [12]. Then, we have divided the spectrum into ten frequency sub-bands as represented in TABLE I. The relative spectral power (RSP) is the ratio between the sub-band spectral power (*BSP)* and the total spectral power (considering all *BSP* sub-bands). This normalization can be helpful for improving the classification robustness [13]. We have computed RSP for each sub-band separately.

Slow wave index, defined by the following ratios, can also help us to highlight some of the spectral bands over slow wave bands:

$$DSI = BSP_{Delta} / (BSP_{Theta} + BSP_{Alpha}) \quad (1)$$

$$TSI = BSP_{Theta} / (BSP_{Delta} + BSP_{Alpha}) \quad (2)$$

$$ASI = BSP_{Alpha} / (BSP_{Delta} + BSP_{Theta}) \quad (3)$$

where *DSI*, *TSI* and *ASI* stand for delta-slow-wave index, theta-slow-wave index and alpha-slow-wave index, respectively [13].

TABLE I: Spectral sub-bands used in RSP computation [13]

Bands	Sub-bands	Bandwidth $\{f_L , f_H\}$ (Hz)
Delta	Delta 1	{0.5,2.0}
	Delta 2	{2.0,4.0}
Theta	Theta 1	{4.0,6.0}
	Theta 2	{6.0,8.0}
Alpha	Alpha 1	{8.0,10.0}
	Alpha 2	{10.0,12.0}
Sigma	Sigma 1	{12.0,14.0}
	Sigma 2	{14.0,16.0}
Beta	Beta 1	{16.0,25.0}
	Beta 2	{25.0,35.0}

Harmonic parameters include center frequency (f_c), bandwidth (f_σ) and spectral value at center frequency (S_{fc}). These parameters are defined as follows [14]:

$$f_c = \sum_{f_L}^{f_H} f P_{xx}(f) \Big/ \sum_{f_L}^{f_H} P_{xx}(f) \quad (4)$$

$$f_\sigma = \Big(\sum_{f_L}^{f_H} (f - f_c)^2 P_{xx}(f) \Big/ \sum_{f_L}^{f_H} P_{xx}(f)\Big)^{1/2} \quad (5)$$

$$S_{f_c} = P_{xx}(f_c) \quad (6)$$

where $P_{xx}(f)$ denotes the PSD calculated for the desired frequency band (see TABLE I).

Dynamic temporal information of EEG signal can be provided by Hjorth parameters. For the epoch x of length N, the Hjorth parameters includes activity, mobility and complexity which are computed from variance of x and its first and second derivatives (x' and x'' respectively) [15]:

$$Activity(x) = \mathrm{var}(x) = \frac{1}{N}\sum_{n=1}^{N} (x(n) - \bar{x})^2 \quad (7)$$

$$Mobility = \sqrt{\mathrm{var}(x') / \mathrm{var}(x)} \quad (8)$$

$$Complexity = \sqrt{\mathrm{var}(x'') \times \mathrm{var}(x) / \mathrm{var}(x')^2} \quad (9)$$

In statistics and probability theory, skewness and kurtosis measure some characteristics of data probability distribution. Skewness is used to measure the symmetry of probability distribution and kurtosis is used to measure whether the probability distribution of the data is peaked or flat, relative to a normal distribution. Defining the k^{th} order moment m_k as:

$$m_k = \frac{1}{N}\sum_{i=1}^{N}(x(i) - \bar{x})^k \quad (10)$$

where N is the number of samples of an epoch and $\bar{x}$ is the mean of these samples, skewness and kurtosis are given by [16]:

$$skewness = m_3 / m_2 \times \sqrt{m_2} \quad (11)$$

$$kurtosis = m_4 / m_2 \times m_2 \quad (12)$$

Frequency transforms are also among the common methods for feature extraction from brain signals. In this work, we have used Discrete Cosine Transform (DCT), calculated as follows [17]:

$$y(k) = w(k)\sum_{n=1}^{N} x(n)\cos\left(\frac{\pi(2n-1)(k-1)}{2N}\right) \quad (13)$$

where $k=1, 2, \ldots, N$ and

$$w(k) = \begin{cases} 1/\sqrt{N} & k = 1 \\ \sqrt{2/N} & 2 \le k \le N \end{cases} \quad (14)$$

The last features that have been used are coefficients of Autoregressive model. The Autoregressive (AR) model of an order p for the one-dimensional signal $x(n)$ is written as:

$$x(n) = \sum_{i=1}^{p} a_p(i)x(n-i) + e(n) \quad (15)$$

where $a_p(i)$ and $e(n)$ represents the AR coefficients and the zero mean error with a finite variance, respectively. $e(n)$ is assumed to be a random process independent of previous values of the signal x [1]. For each channel, the AR coefficients of the segments have been calculated using the Burg algorithm. In this algorithm, the coefficients are estimated at successive orders (forward and the backward directions) [18].

2.4 Feature Selection

Naïve feature extraction from all EEG channels results in a large number of parameters, without considering the effectiveness of each feature in discriminating different classes. The aim of feature selection is to identify and choose the most effective features. In this study, we used orthogonal forward selection algorithm (OFS).

Based on the desired evaluation criterion, feature selection methods can be divided into filter and wrapper strategies [19]. In the wrapper strategies, feature selection and pattern classification are considered as a whole. In these methods, feature subsets are directly evaluated based on classification results. On the other hand, the filter strategies employ intrinsic characteristics of data (e.g. class separability measures) as the criterion for feature subset selection. Since the filter strategies are independent from the classification algorithms, any classifier can use the feature subsets selected by these methods. In the present study, we have evaluated and selected feature subsets based on Mahalanobis class separability measure (see, for example, [20]).

After selecting the evaluation criterion, we need to apply a suitable searching algorithm on the features. Although exhaustive searching method finds the optimal solution, it needs to evaluate all possible subsets of the features. Generally, exhaustive searching method is not applied when we are dealing with a large number of features (due to its extremely low speed). In practice, suboptimal searching methods such as sequential forward selection (SFS) algorithm are often employed [20]. We have attempted to reduce the redundancy problem of the features by employing orthogonal decompositions, while using SFS algorithm for finding the most effective features. The strength of employing orthogonal decomposition is that it decorrelates the features in the orthogonal space, which can be then evaluated and selected independently. The orthogonal transform used in our study is Gram–Schmidt transform. This process uses a set of vector as input and generates an orthogonal set, in an inner product apace, with the same dimension of input [9].

2.5 Classification

Each EEG segment corresponds to one of the six mental activities done by a subject. We want to design a system which is capable of receiving the selected features as input (extracted from an EEG segment), and predicting the corresponding mental activity. In this work, Bayesian classifier has been used for classifying EEG segments. Although Bayesian optimal classifier has the least error probability, it is not easy to accurately calculate the needed probability functions. Thus, a special parametric structure, such as Gaussian distribution, is usually used [21]. We have assumed Gaussian distribution of our input vectors (which is a very common assumption) for training our Bayesian classifier. It should be stated that in order to avoid dealing with ill-conditioned covariance matrix, Linear Discriminant Analysis (LDA) [22] has been applied on the selected features, before using them by the Bayesian classifier. In our study, 70 percent of the data for any of the classes was chosen randomly for training the classifier, whereas 30 percent left was chosen for testing.

3. Results and Discussion

In this study, an experiment has been designed in order to investigate the performance of feature groups and EEG channels when discriminating six different mental activities. In this test, 700 features were chosen out of all

of the extracted features of all the channels using OFS algorithm. Number of selected features from each channel using OFS algorithm has been presented in TABLE II.

TABLE II: Number of selected features from each channel within the 700 most discriminative features for all subjects.

		Subjects				
		1	2	3	4	Mean
Channels	C3	34	38	48	41	40.25
	C4	45	38	30	35	37.00
	Cz	38	34	35	37	36.00
	F3	46	34	32	37	37.25
	F4	35	40	43	40	39.50
	F7	38	38	45	33	38.50
	F8	39	33	41	36	37.25
	FP1	33	37	37	21	32.00
	FP2	36	42	57	43	44.50
	Fz	41	32	34	39	36.50
	O1	34	33	25	37	32.25
	O2	31	35	36	45	36.75
	P3	33	34	26	29	30.50
	P4	37	36	39	45	39.25
	Pz	29	36	40	45	37.50
	T3	32	34	28	27	30.25
	T4	43	38	21	38	35.00
	T5	31	46	32	32	35.25
	T6	45	42	51	40	44.50

As it can be seen in TABLE II, not only all of the EEG channels have been used for feature selection, but also nearly equal number of features has been selected from each channel. Hence, we cannot choose a channel as the best one exclusively and all of the channels provide useful features. But if we take the mean number of selected features for all subjects, it could be said that T6, FP2 and C3 play the most important roles in separating six classes. Moreover, T3 and P3 seems to be the least important channels.

We have considered 7 different groups of features related to Relative Spectral Power (RSP), harmonic parameters (HP), Slow Wave Index (SWI), Hjorth parameters (PHj), kurtosis and skewness (KS), coefficients of Autoregressive model (AR), and coefficients of Discrete Cosine Transform (DCT). Number of selected features from each feature groups has been presented in TABLE III. As TABLE III shows, features related to discrete cosine transform (DCT) has been selected the most. However, it cannot necessarily mean that this group of features is the most important one. It is because of the fact that the total number of features in the considered groups of features is not the same. Hence, we must normalize the number of selected features in each group according to the total number of features in each one. Normalized number of selected features of each group has been illustrated in Fig. 2.

TABLE III: Number of feature groups within the 700 most discriminative features for all subjects.

		Subjects				
		1	2	3	4	Mean
Features	RSP	110	113	119	113	113.75
	HP	168	193	180	193	183.5
	SWI	37	34	34	36	35.25
	PHj	40	41	29	43	38.25
	KS	27	25	26	28	26.50
	AR	60	77	83	73	73.25
	DCT	258	217	229	214	229.50

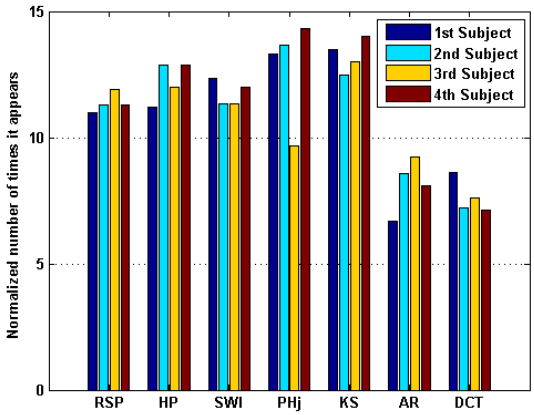

Fig. 2: Normalized Number of selected features from each group

As it can be seen in Fig. 2, results have been completely changed by normalizing the number of selected features. Now, kurtosis and skewness (KS group) which are related to the signal probability distribution are the most important features. After this group, Hjorth parameters including activity, mobility and complexity are selected the most (except for the third subject). Also, features in AR and DCT groups are the least important features.

Finally correct classification rate based on the selected features, using Bayesian classifier, has been presented in TABLE IV.

TABLE IV: CCRs and their standard deviation for all subjects

Subject	1	2	3	4
CCR (%)	93.89	89.07	92.41	91.11
Std	1.45	1.10	0.73	0.45

As it can be seen, the resulted CCRs for all of the subjects are pretty high with acceptance standard deviations. It should be noticed that the random classification rate for six classes is about 16.67%.

4. Conclusion

In this study, we have collected the brain signals of four subjects, during imagination of six imagery movements, in order to design a Brain-Computer Interface system. The main goal of this paper was to select the most discriminative and non-redundant features among a pool of features. For this reason, we have used Orthogonal Forward Selection.

The results show that almost all channels of EEG seem to have an equivalent importance for discriminating the six classes from each other. However, when evaluating the extracted features, kurtosis, skewness and Hjorth parameters seem to be among the most effective features for classifying the desired six mental tasks.

By considering six different classes, achieving a low rate of classification can be expected. However, the classification rate resulted by applying Bayesian classifier on the selected features is in average 91.62%, which is pretty high.

As a future work, we suggest to go more deeply into feature selection process by considering other types of searching methods and/or selecting smaller number of features.

Acknowledgement

The authors would like to thank Nima Salehi Ahangar and Mina Mirjalili for their assistance during data collection.

References

[1] F. Faradji, R. K. Ward, and G. E. Birch, "A brain-computer interface based on mental tasks with a zero false activation rate," in *Neural Engineering, 2009. NER'09. 4th International IEEE/EMBS Conference on*, 2009, pp. 355-358.

[2] B. Graimann, B. Allison, and G. Pfurtscheller, *Brain-computer interfaces: Revolutionizing human-computer interaction*: Springer, 2010.

[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*, vol. 4, p. R32, 2007.

[4] J. R. Wolpaw, D. J. McFarland, and T. M. Vaughan, "Brain-computer interface research at the Wadsworth Center," *Rehabilitation Engineering, IEEE Transactions on*, vol. 8, pp. 222-226, 2000.

[5] B. Blankertz, G. Dornhege, M. Krauledat, K.-R. Müller, and G. Curio, "The non-invasive Berlin brain–computer interface: fast acquisition of effective performance in untrained subjects," *NeuroImage*, vol. 37, pp. 539-550, 2007.

[6] G. Pfurtscheller, C. Neuper, A. Schlogl, and K. Lugger, "Separability of EEG signals recorded during right and left motor imagery using adaptive autoregressive parameters," *Rehabilitation Engineering, IEEE Transactions on*, vol. 6, pp. 316-325, 1998.

[7] K.-M. Ong and R. Paramesran, "A Study of Power Asymmetry Ratio of Single Trial Electroencephalogram During Finger Movement," in *TENCON 2005 2005 IEEE Region 10*, 2005, pp. 1-4.

[8] R. G. Rasmussen, S. Acharya, and N. V. Thakor, "Accuracy of a Brain-Computer Interface in Subjects with Minimal Training," in *Bioengineering Conference, 2006. Proceedings of the IEEE 32nd Annual Northeast*, 2006, pp. 167-168.

[9] K. Mao, "Orthogonal forward selection and backward elimination algorithms for feature subset selection," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 34, pp. 629-634, 2004.

[10] J. Webster, *Medical instrumentation: application and design*: John Wiley & Sons, 2009.

[11] R. Vigário, J. Sarela, V. Jousmiki, M. Hamalainen, and E. Oja, "Independent component approach to the analysis of EEG and MEG recordings," *Biomedical Engineering, IEEE Transactions on*, vol. 47, pp. 589-593, 2000.

[12] A. S. Yilmaz, A. Alkan, and M. H. Asyali, "Applications of parametric spectral estimation methods on detection of power system harmonics," *Electric Power Systems Research*, vol. 78, pp. 683-693, 2008.

[13] H. Simões, G. Pires, U. Nunes, and V. Silva, "Feature Extraction and Selection for Automatic Sleep Staging using EEG," in *ICINCO (3)*, 2010, pp. 128-133.

[14] W.-C. Tang, S.-W. Lu, C.-M. Tsai, C.-Y. Kao, and H.-H. Lee, "Harmonic Parameters with HHT and Wavelet Transform for Automatic Sleep Stages Scoring," *International Journal of Biomedical Sciences*, vol. 2, 2007.

[15] K. Ansari Asl, G. Chanel, and T. Pun, "A channel selection method for EEG classification in emotion assessment based on synchronization likelihood," 2007.

[16] L. Zoubek, S. Charbonnier, S. Lesecq, A. Buguet, and F. Chapotot, "Feature selection for sleep/wake stages classification using data driven methods," *Biomedical Signal Processing and Control*, vol. 2, pp. 171-179, 2007.

[17] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-time signal processing* vol. 2: Prentice-hall Englewood Cliffs, 1989.

[18] J. P. Burg, "A new analysis technique for time series data," *NATO Advanced Study Institute on Signal Processing with Emphasis on Underwater Acoustics*, vol. 1, 1968.

[19] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial intelligence*, vol. 97, pp. 273-324, 1997.

[20] P. A. Devijver and J. Kittler, *Pattern recognition: A statistical approach* vol. 761: Prentice-Hall London, 1982.

[21] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*: John Wiley & Sons, 2012.

[22] I. Jolliffe, *Principal component analysis*: Wiley Online Library, 2005.