FOUNDATIONS OF COMPUTING AND DECISION SCIENCES
Vol. 37 (2012) No. 4

DOI: 10.2478/v10209-011-0016-7

EEG FEATURE SELECTION FOR BCI BASED ON MOTOR IMAGINARY TASK

Izabela REJER*

Abstract. The greatest problem met when a Brain Computer Interface (BCI) based on electroencephalographic (EEG) signals is to be created is a huge dimensionality of EEG feature space and at the same time very limited number of possible observations. The first is a result of a huge amount of data which can be recorded during the single trial, the latter - the result of individuality of EEG signals, which can significantly differ in different frequency bands determined for different subjects. These two reasons force the brain researches to reduce the huge EEG feature space to only some features, those which allow to build a BCI of a satisfactory accuracy. The paper presents the comparison of two methods of feature selection – blind source separation (BSS) method and method using interpretable features. The comparison was carried out with the data set recorded during EEG session with a subject whose task was to imagine movements of right and left hand.

Keywords: BCI, EEG classification, feature selection, neural classifier.

1. Introduction

A BCI (Brain Computer Interface) is defined as a communication system in which messages or commands that a user sends to the external world do not pass through the brain's normal output pathways of peripheral nerves and muscles [14]. Instead, the messages and commands are encoded in the brain activity and in order to be propagated to the external world, they have to be read directly from the brain via a dedicated device (e.g. EEG device).

BCIs can be classified according to several criteria, among which two are most frequently mentioned – BCI dependency and invasiveness. According to the first criterion, two kinds of BCIs can be pointed out: dependent BCIs and independent BCIs. The characteristic feature of dependent BCIs is that they require some amount of motor control from the user. This control is not needed, however, to send messages to the external world but to support the brain to generate the activity encoding these messages (e.g. control of muscles responsible for changing the gaze direction). On the contrary, in case of

* West Pomeranian University of Technology of Szczecin, Faculty of Computer Science, ul. Żołnierska 52, 71-210 Szczecin, Poland

independent BCIs, no motor control is needed because the brain activity encoding the message is generated by the brain itself. According to the second criterion, a BCI system can be classified as an invasive or a non-invasive one [6]. While in an invasive BCI sensors used for measuring the brain's activity are placed directly on the cortex, in a non-invasive BCI the sensors are placed outside the head.

Regardless of BCI type, in order to execute a command via a BCI, the brain activity has to be read, decoded and then translated into right command. Reading mind is not an easy task but there are some methods capable of monitoring brain activity like: electroencephalography (EEG), magnetoencephalography (MEG), positron emission tomography (PET), functional magnetic resonance imaging (fMRI) and others. At the moment, however, due to high costs and very limited mobility, most methods (apart from EEG) cannot be considered as a tool for establishing a reasonable BCI, it is a BCI which is not very expensive and possible to use in different environments – not only in laboratory conditions.

While the role of EEG is well established in the first step of the process of BCI designing, the second step (signal decoding) can be performed with a huge variety of mathematic techniques. In fact, all the time new papers presenting improved methods for brain activity decoding are published in top scientific journals. Methods used in this step can be divided into four categories, according to four stages of information decoding process which are: preprocessing, feature extraction, feature selection and classification. All four stages are very important to design a BCI of a high quality, however, the second and third stage are usually considered as most crucial [10][3].

The necessity of choosing or designing right method for feature extraction is broadly discuss in scientific literature but there is surprisingly a very small number of papers addressing the issue of feature selection [5][9]. And the problem of feature selection in case of BCI is not a trivial one because the EEG feature space is often very large (at least several hundred features).

The paper presents a comparison of two methods of feature selection – PCA (Principal Components Analysis), which is a BSS (Blind Source Separation) method, and method using features from the original feature space. The methods were compared in terms of classification accuracy obtained with neural classifiers using selected features. The comparison was carried out with a data set submitted to the second BCI Competition (data set III – motor imaginary) by Department of Medical Informatics, Institute for Biomedical Engineering, University of Technology Graz [3]. The data set was recorded from a normal subject (female, 25 years old) which task was to control a feedback bar by means of imagery of left or right hand movements. The data set contains 140 EEG signals, measured over three canals: C3, Cz and C4, sampled with 128Hz and preliminary filtered between 0.5 and 30Hz. In order to perform analysis, the signals were refiltered in 12 different frequency bands and signal power was calculated, separately per each frequency band and each second of the recording.

2. Feature selection

Considering the recording of EEG signals from only 32 electrodes and calculating the signal power separately in 5 frequency bands and individually per each of 5 seconds, the

number of features in BCI feature space would be equal to 800. Obviously such large number of features could not be introduced to the classifier due to enormous number of trials which had to be performed in order to collect enough training data. Since, it is recommended to use at least 5 to 10 times more training data per class than the number of features [12], at least 160000 data records would have been needed in case of only two classes and only 20 (from original 800) input features. Even such amount of data is impossible to obtain in BCI domain because the shape and the power of EEG signal in different frequency bands is subject specific [11] and so all signals needed to train the classifier have to be recorded from the same subject. Due to this fact, in real applications of BCI, the number of trials is not greater than one or two hundreds. Such small amount of training data put significant limits on the number of features which can be used in the interface and so imply the necessity of applying methods of feature selection.

Feature selection methods can be divided into two categories: methods based on interpretable features extracted from the original feature space and BSS (Blind Source Separation) methods, it is methods based on artificially created, non-interpretable features. In case of methods from the second group, the selection process is straightforward because new features are arranged according to a given criterion, e.g. decreasing variance. Hence, in order to reduce the feature space, it is enough to use chosen BSS method to calculate the set of artificial features and then to choose assumed number of features from the top places of created ranking. Some of the methods from this group are: ICA (Independent Component Analysis), PCA (Principal Component Analysis) and MNF (Maximum Noise Fraction).

The selection process in case of methods based on features extracted from the original feature space is much more demanding because mostly there are no cues indicating which of extracted features are more and which are less significant for a given subject. That is why before the final selection of several from hundreds of features, features significant for a given subject has to be determined. Since, the features are to be used as inputs for a classifier, the search for subset of most significant features is driven by the criterion of maximizing classification accuracy. That means that for each examined subset of features, a chosen classifier has to be trained and tested. Since, there is no possibility of performing the exhaustive search in space of several hundreds of features, some heuristics are used to shorten the searching process. The most widely used are GA (Genetic Algorithms) which are search algorithms applying mechanisms of natural evolution. The alternative method, giving promising results in respect to length of the search process and classification accuracy is a forward step-selection method, applied in this paper.

3. Applied methods

3.1. Principal Components Analysis

Principal components analysis (PCA) is an example of BSS methods, it is methods which in order to reduce the original feature space transform data from this space to another, artificially created one [13]. To be more specific, PCA is defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest

variance lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on.

The formal definition of PCA method is as follows. Let's define a data matrix X (*M*x*N*), where *M* – number of observations, *N* – number of features. C (*N*x*N*) is the covariance matrix of matrix X:

$$C = \begin{bmatrix} Var(x_1) & Cov(x_1, x_2) & \dots & Cov(x_1, x_N) \\ Cov(x_2, x_1) & Var(x_2) & \dots & Cov(x_2, x_N) \\ \dots & \ddots & \dots & \dots \\ Cov(x_N, x_1) & Cov(x_N, x_2) & \dots & Var(x_N) \end{bmatrix}, \tag{1}$$

where: $\mathrm{Cov}(x_i, x_j) = \frac{1}{M}\sum_{k=1}^{M}(x_i(k) - \bar{x_i})(x_j(k) - \bar{x_j})$, $\mathrm{Var}(x_i) = \sigma_{x_i}^2$. The matrix C can be presented as a product of matrix of eigenvectors and matrix of eigenvalues:

$$C\Gamma = \Lambda\Gamma \implies C = \Gamma\Lambda\Gamma^T, \tag{2}$$

where Λ(*NxN*) – matrix of sorted eigenvalues ($\lambda_1 \geq \lambda_2 \geq \cdots \lambda_N$):

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \dots & \dots & \ddots & \dots \\ 0 & 0 & \dots & \lambda_N \end{bmatrix} \tag{3}$$

and Γ(*NxN*) - matrix of eigenvectors (each column corresponds to one eigenvalue from matrix Λ):

$$\Gamma = \begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1N} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2N} \\ \dots & \dots & \ddots & \dots \\ \gamma_{N1} & \gamma_{N2} & \cdots & \gamma_{NN} \end{bmatrix}. \tag{4}$$

The matrix of Principal Components is defined as:

$$Y = X\Gamma = [y_1, y_2, \dots, y_N] = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1N} \\ y_{21} & y_{22} & \cdots & y_{2N} \\ \dots & \dots & \ddots & \dots \\ y_{M1} & y_{M2} & \cdots & y_{MN} \end{bmatrix}. \tag{5}$$

Each of *N* features $x_1, x_2, \dots, x_n$ can be described with parameters from matrix Y as linear combination of K first principal components: $x_{mn} = \sum_{k=1}^{K} \gamma_{nk} y_{nk}$.

3.2. Forward step-selection

Step-selection methods are heuristic methods defining the overall strategy of the search process. In case of non linear data all of them act according to the so-called "wrapper" approach to feature selection [4]. That means that the process of searching for the suitable feature subset is driven by the results of the model created with this subset. So, when the classification accuracy is the criterion of the model quality, the search process is also aimed at maximizing the classification accuracy.

There are three main types of step-selection methods: forward selection, backward selection and bidirectional selection. The only difference between them is the search direction of the feature space. In case of the first method, the search process starts with an empty set of features which is then extended by adding one feature at each step of the procedure. At each step this feature is added to the set which causes the greatest increase in classifier accuracy. The whole process ends when none of the remaining features is able to improve the classifier performance. In case of the second method, the selection process starts with the set containing all possible features which are one by one discarded from the model in succeeding steps of the search process. Of course this time this feature is eliminated from the feature set which causes the smallest (or none) improvement in the classifier accuracy. The last method is a simple mix of forward and backward selection – both elementary strategies are used alternately.

The decision whether to use a forward or backward selection is always determined by the characteristics of a given data set. In case of EEG feature space the forward step-selection is the only choice because of very limited number of observations and hence very limited number of features possible to introduce to a classifier.

3.3. Neural classifier

The neural classifier [15] is a mapping function, $F: \mathcal{R}^d \rightarrow \mathcal{R}^M$, where d-dimensional input vector x is submitted to the network and M-dimensional output y is obtained. The network is typically built such that an overall error measure such as mean squared error (MSE) is minimized. The mapping function which minimizes the expected squared error $E[y - F(x)]^2$ is the conditional expectation of y given x: $F(\mathrm{x}) = E[y|x]$. In the classification problem, where the desired network output y is a vector of binary values, the least squares estimation for the mapping function is the posterior probability of membership of x to class j: $F(\mathrm{x}) = \mathrm{P}[\omega_j|x]$, where ω is the membership variable that takes a value of ω_j if an object x belongs to class j.

Different architectures of neural networks (NN) can be used as classifiers but in BCI research mostly a classic feedforward network, it is a network of the following parameters is applied: flow of signals: one-way, architecture of connections between layers: all to all, hidden layers: mostly 1 hidden layer with an appropriate number of non-linear neurons, output layer: 1 neuron per class (or only one neuron in two classes problems). The appropriate number of hidden neurons has to be chosen in relation to the training set - to be more specific, in relation to: number of input features, number of observations and number of output classes.

In scientific papers the number of observations sufficient to built a good (it means not overfitted) classifier is correlated with the number of input features and output classes. Mostly it is recommended to use at least 5 to 10 times more training data per class than the number of input features [12]. This rule is valid in linear classification problems where the degree of classifier's freedom grows linearly with number of input features. However, in case of non-linear classifiers this estimation is too weak because it has nothing to do with real dimensionality of the training process, which depends only on the number of training parameters. Therefore, in order to choose the right number of hidden neurons for a non-linear neural classifier, the rule given above should be strengthen and number of training data should be correlated not with the number of input features but with the number of training parameters (and of course with the number of output classes). For example, in case of the classification problem of: 400 observations, 2 classes and 5 features, the number of hidden neurons should be equal or smaller than 5 (5 hidden neurons gives 36 training parameters – assuming 2 parameters per one neuron in hidden and output network layer).

4. Experiments settings and results

The experiments described later in this section were aimed at comparing two methods of dimensionality reduction – PCA method, returning a set of artificial, non-interpretable features and forward step-selection method, returning a reduced set of original, interpretable features. The methods were compared in terms of classification accuracy obtained with neural classifiers trained over both set of features.

The comparison was carried out with a data set submitted to the second BCI Competition (data set III – motor imaginary) by Department of Medical Informatics, Institute for Biomedical Engineering, University of Technology Graz [3]. The data set was recorded from a normal subject (female, 25y) which task was to control movements of a feedback bar by means of imagery movements of left and right hand. Cues informing about direction in which the feedback bar should be moved were displayed on a screen in form of left and right arrows. The order of left and right cues was random. The experiment consisted of 280 trials, each trial lasted 9 seconds. The first 2s was quite, at t=2s an acoustic stimulus sounded and a cross "+" was displayed for 1s; then at t=3s, an arrow (left or right) was displayed as cue. The EEG signals were measured over three bipolar EEG channels (C3, Cz and C4), sampled with 128Hz and preliminary filtered between 0.5 and 30Hz. The whole data set, containing data from 280 trials, was then divided into two equal subsets – first intended for classifier training and second intended for external classifier test. Since only data from the first subset was published with target values, only this subset was used in the experiments described in this paper.

4.1. Data preprocessing

Each of 140 trials from the original data set was described by 3456 features (128 data per second, 9 seconds, 3 channels). Since, it is well known that hand motor area representation is located on the mantle of the cortex and is lateralized [1], data from canal Cz (placed between hemispheres) were discarded from the survey. Also data from the first and second

seconds of the recordings of each trial were discarded (they could be removed because the applied feature extraction method does not need any reference period). After the both manipulations the data set was preliminary reduced to 1792 features (128 data per second, 7 seconds, 2 channels).

To further reduce the number of features for classifier training, the original data set was transformed to set of frequency band power features. The signal power was calculated separately for:

- 12 frequency bands: alfa band (8-13Hz) and five sub-bands of alfa band (8-9Hz; 9-10Hz; 10-11Hz; 11-12Hz; 12-13Hz); beta band (13-30Hz) and also five sub-bands of beta band (13-17Hz; 17-20Hz; 20-23Hz; 23-26Hz; 26-30Hz),
- each of 7 seconds of the trial,
- each of 2 canals (C3, C4).

In this way 168 band power features were obtained. Obviously, taking into account very small number of trials equal to 140, the number of features had to be still significantly reduced before the classifier training. In fact with 140 trials, no more than several features could be used without the threat of overfitting. Hence, in order to perform the final reduction of the data set, forward step-selection and PCA methods were used.

4.2. Forward step-selection results

At the first step of the forward step-selection method 168 neural classifiers of one input feature were trained. The classifiers parameters were as follows [8]: flow of signals: one-way, architecture of connections between layers: all to all, hidden layers: 1 hidden layer with 4 sigmoid neurons, output layer: 1 linear neuron, training method: backpropagation algorithm with momentum and changing learning rates, training time: 1000 epoch. The data set was divided into three subsets: training set (70%), testing set (20%) and validation set (10%). The classification threshold was set to 0.5 and hence all network results greater than 0.5 were classified as class "2" (right hand) and results smaller or equal 0.5 were classified as class "1" (left hand).

Four accuracy measures were calculated per each classifier: training accuracy (Acc_{tr}), testing accuracy (Acc_t), validation accuracy (Acc_v) and general accuracy (Acc_g). All measures were calculated according to the same equation:

$$Acc = \frac{R}{U} * 100\%, \tag{6}$$

where: Acc – accuracy measure, R - number of properly classified cases (from testing set for Acc_t, from training set for Acc_{tr}, from validation set for Acc_v, or from all sets together for Acc_g), U – number of cases in an appropriate set (or the total number of cases in case of Acc_g). The choice of the best classifier was made on the basis of Acc_g, but only these classifiers were compared for which the weighted average of Acc_t and Acc_v was not greater than Acc_{tr}:

$$Acc_{av} = \frac{0.2 * Acc_t + 0.1 * Acc_v}{0.3} * 100\%. \quad (7)$$

The best classifier, it is the classifier of the highest value of Acc_g (equal to 74.29) occurred to be the classifier using feature number 52 (signal power in frequency band: 10-11Hz, second: 3, canal: C4). All accuracy measures of this and next classifiers are presented in Table 1.

In succeeding step of the experiment 167 two-input classifiers were trained. First input of each classifier was feature number 52, second input was one of the remaining features. The classifiers parameters were the same as in the first step, except of the number of hidden neurons which was equal to three. This time the highest accuracy (Acc_g equal to 82.86) was obtained by the classifier with input features number 52 and 45 (signal power in band: 10-11, second: 3, canal: C3).

Three more steps of forward step-selection procedure were performed. At each step one input of the highest accuracy was added to the classifier:

- step 3 – feature number 137 – the band power in band: 20-23, second: 4, canal: C4; number of hidden neurons: 2; Acc_g=85.00,
- step 4 – feature number 129 – the band power in band: 10-11, second: 4, canal: C4; number of hidden neurons: 2; Acc_g =86.43,
- step 5 – feature number 53 – the band power in band: 20-23, second: 3, canal: C3; number of hidden neurons: 1; Acc_g =87.86.

Table 1. Classifiers accuracy measures – forward step-selection method (description of features in text)

No. of features	Acc_g	Acc_t	Acc_v	Acc_{tr}	Acc_{av}
52	**74.29**	75.00	78.57	73.47	76.20
52-45	**82.86**	85.71	85.71	81.63	85.71
52-45-137	**85.00**	89.29	85.71	83.67	88.10
52-45-137-129	**86.43**	92.86	78.57	85.71	88.10
52-45-137-129-53	**87.86**	89.29	85.71	87.76	88.10

4.3. PCA results

The whole set of 168 features was transformed with PCA algorithm. As a result a set of 168 components was obtained. Five first components were then used to build five neural classifiers corresponding in number of inputs and number of hidden neurons to classifiers obtained with forward step-selection method. The classifiers parameters were the same as in Section 4.2. Accuracy measures for trained classifiers are presented in Table 2.

Table 2. Classifiers accuracy measures – PCA method

No. of component	No. of neurons	Acc_g	Acc_t	Acc_v	Acc_{tr}	Acc_{av}
1	4	**55.71**	64.29	50.00	54.08	59.50
2	3	**70.71**	82.14	64.29	68.37	76.20
3	2	**72.86**	78.57	78.57	70.41	78.57
4	2	**79.29**	85.71	78.57	77.55	81.40
5	1	**80.00**	75.00	92.86	79.59	81.00

4.4. Discussion of results

Results obtained in the first experiment show that the highest discrimination abilities had signal power calculated for sub-band of alfa band in third second after presenting the visual cue to the subject (features number 52 and 45). These results are in line with previously reported experiments considering ERS/ERD (Event Related Synchronization/Desynchronization), where it is often emphasized that during imagination of hand movements, ERD in alfa band in relevant canals takes place [7]. Also high discrimination capabilities of features number 129 and 137, reflecting the signal power calculated for sub-band of beta band in third and fourth second after presenting the visual cue to the subject, coincide with BCI literature. For example Pfurtscheller and Lopes da Silva in [11] reports that during hand movement ERD in beta band over the hemisphere contralateral to the hand and ERS in beta band over the parallel hemisphere are observed.

5. Summary

Comparing the results obtained with both methods of feature selection (PCA and forward step-selection method), it can be noticed that at each step of the experiment more precise results were obtained when the classifier was trained with a set of original features from the feature set. In case of BCI, where the accuracy is of the highest importance, this feature of the selection method is a crucial one and should be taken into account when the method is to be chosen. Moreover, when the selection method works with original features, obtained results give some insight into the analyzed system (brain), and perhaps could allow the researcher to find out new patterns of brain behavior. Of course BSS methods also have some benefits - most important is their speed. However, since the selection process in BCI is not performed online, the speed of the selection method should not be a decisive feature.

References

[1] Boord P., Craig A., Tran Y., Nguyen H.: Discrimination of left and right leg motor imagery for brain-computer interfaces, *Medical & Biological Engineering & Computing*, **48**, 4, 2010, 343–350.

[2] Hammon P.S., de Sa V.R., Preprocessing and meta-classification for brain-computer interfaces. IEEE Transactions on Biomedical Engineering, **54**, 3, 2007, 518–525.
[3] II BCI Competition; http://bbci.de/competition/ii/index.html; data set III – motor imaginary
[4] Kohavi R., John G.H., Wrappers for feature subset selection*, Artificial Intelligence*, **1-2**, 1997, 273-324.
[5] Lakany H., Conway B. A., Understanding intention of movement from electroencephalograms, *Expert Systems*, **24**, 5, 2007, 295–304.
[6] Lebedev M.A., Nicolelis M.A.L., Brain-machine interfaces: past, present and future. *Trends in Neurosciences*, **29**, 9, 2006, 536–546.
[7] Leocani L., Toro C., Zhuang P., Gerloff C., Hallet M., Event-related desynchronization in reaction time paradigms: a comparison with event-related potentials and corticospinal excitability, *Clinical Neurophysiology*, **112**, 2001, 923–930.
[8] Masters T., *Practical Neural Networks Recipes in C++,* Academic Press Inc, 1993.
[9] Peterson D. A., Knight J. N., Kirby M. J., Anderson Ch. W., Thaut M. H., Feature Selection and Blind Source Separation in an EEG-Based Brain-Computer Interface, *EURASIP Journal on Applied Signal Processing*, **19**, 2005, 3128–3140.
[10] Pfurtscheller G., Flotzinger D., Kalcher J., Brain-computer interface – a new communication device for handicapped persons, *Journal of Microcomputer Application*, **16**, 1993, 293–299.
[11] Pfurtscheller G., Lopes da Silva F. H., Event-related EEG/MEG synchronization and desynchronization: basic principles, *Clinical Neurophysiology***, 110**, 1999, 1842–1857
[12] Raudys S. J., Jain A. K., Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **13**, 3, 1991, 252–264.
[13] Reed R. D., Marks II R. J., *Neural Smithing, Supervised Learning in Feedforward Artificial Neural Networks*, MIT Press, London, England, 1998
[14] Wolpaw J. R., Birbaumer N., McFarland D. J., Pfurtscheller G., Vaughan T. M., Brain–computer interfaces for communication and control, *Neurophysiology* **113**, 2002, 767–791.
[15] Zhang G. P., Neural Networks for Classification: A Survey, *IEEE Transactions on Systems, Man and Cybernetics – Part C: Applications and Reviews*, 30, 4, 2000, 451-462

Presented at the Congress of Young IT Scientists, Międzyzdroje, Poland, 20-22.09.2012