We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,900
Open access books available

185,000
International authors and editors

200M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

3

Feature Extraction by Mutual Information Based on Minimal-Redundancy-Maximal-Relevance Criterion and Its Application to Classifying EEG Signal for Brain-Computer Interfaces

Abbas Erfanian, Farid Oveisi and Ali Shadvar
Iran University of Science and Technology
Iran

1. Introduction

Dimensionality reduction of the raw input variable space is an essential preprocessing step in the classification process. In general, it is desirable to keep the dimensionality of the input features as small as possible to reduce the computational cost of training a classifier as well as its complexity (Torkkola, 2003; Murillo & Rodriguez, 2007). Moreover, using large number of features, when the number of data is low, can cause degradation of the classification performance (Chow & Huang, 2005). Reduction of the number of input features can be done by selecting useful features and discarding others (i.e., feature selection) (Battiti, 1994; Kwak & Choi, 2002; Peng et al., 2005; Estèvez et al., 2009; Sindhwani et al., 2004) or extracting new features containing maximal information about the class label from the original ones (i.e., feature extraction) (Torkkola, 2003; Hild II et al., 2006; Kwak, 2007; Murillo & Rodriguez, 2007).

In this paper, we focus on feature selection methods. A variety of linear feature extraction methods have been proposed. One well-known feature extraction methods may be principal component analysis (PCA) (Li et al., 2006). The purpose of PCA is to find an orthogonal set of projection vectors or principal components for feature extraction from given training data through maximizing the variance of the projected data with aim of optimal representing the data in terms of minimal reconstruction error. However, in its feature extraction for classification tasks, PCA does not sufficiently use class information associated with patterns and its maximization to the variance of the projected patterns might not necessarily be in favor of discrimination among classes, thus naturally it likely loses some useful discriminating information for classification.

Linear discrimination analysis (LDA) is another popular linear dimensionally reduction algorithm for supervised feature extraction (Duda et al., 2001). LDA computes a linear transformation by maximizing the ratio of between-class distance to within-class distance, thereby achieving maximal discrimination. In LDA, a transformation matrix from an n-dimensional feature space to a d-dimensional space is determined such that the Fisher criterion of between-class scatter over within-class scatter is maximized. LDA algorithm assumes the sample vectors of each class are generated from underlying multivariate

Normal distributions of common covariance matrix but different means (i.e., homoscedastic data). Over the years, several extensions to the basic formulation of LDA have been proposed (Yu & Yang, 2001; Loog & Duin, 2004). Recently, a method based on Discriminant Analysis (DA) was proposed, known as Subclass Discriminant Analysis (SDA), for describing a large number of data distributions (Zhu & Martinez, 2006). In this approach, the underlying distribution of each class was approximated by a mixture of Gaussians. Then a generalized eigenvalue decomposition was used to find the discriminant vectors that best (linearly) classify the data,

Independent component analysis (ICA) has been also used for feature extraction. ICA is a signal processing technique in which observed random data are linearly transformed into components that are statistically independent from each other (Hyvarinen, Karhunen & Oja, 2001). However, like PCA, the method is completely unsupervised with regard to the class information of the data. A key question is which independent components (ICs) carry more information about the class label. Kwak & Choi (2003) proposed a method for standard ICA to select a number of ICs (i.e., features) that carry information about the class label and a number of ICs that do not. It was shown that the proposed algorithm reduces the dimension of feature space while improving classification performance. We have already used ICA-based feature extraction for classifying the EEG patterns associated with the resting state and the imagined hand movements (Erfanian & Erfani, 2004) and demonstrated the improvement of the performance.

One of the most effective approaches for optimal feature extraction is based on mutual information (MI). MI measures the mutual dependence of two or more variables. In this context, the feature extraction process is creating a feature set from the data which jointly have largest dependency on the target class and minimal redundancy among themselves. In computing the mutual information, one needs to know the multivariate probability density function which is almost impossible to estimate.

To overcome this problem, in (Torkkola, 2003; Hild II, Erdogmnus, Torkkola & Principe, 2006), a method was proposed, known as MRMI, for learning linear discriminative feature transform using an approximation of the mutual information between transformed features and class labels as a criterion. The approximation is inspired by the quadratic Renyi entropy which provides a non-parametric estimate of the mutual information. No simplifying assumptions, such as Gaussian, need to be made about the densities of the classes. However, there is no general guarantee that maximizing the approximation of mutual information using Renyi's definition is equivalent to maximizing mutual information defined by Shannon. Moreover, MRMI algorithm is subject to the curse of dimensionality (Hild II, Erdogmnus, Torkkola & Principe, 2006). To overcome the difficulties of MI estimation for feature extraction, Parzen window modeling was also employed to estimate the probability density function (Kwak, 2007). However, Parzen model may suffer from the “curse of dimensionality,” which refers to the overfitting of the training data when their dimension is high (Murillo & Rodriguez, 2007). Due to this difficulty, some recent works on information-theoretic learning have proposed the use of alternative measures for MI (Murillo & Rodriguez, 2007) by means of an entropy estimation method that has succeeded in independent component analysis (ICA). The features are extracted one by one with maximal dependency to the target class. Although, the mutual information between the features and the classes is maximized, but the proposed scheme does not produce minimal information redundancy between the extracted features.

All the above mentioned methods are based on the idea that a linear projection on the data is applied that maximizes the mutual information between the transformed features and the class labels. Finding the linear mapping was performed using standard gradient descent-ascent procedure which suffers from becoming stuck in local minima.

The purpose of this paper is to introduce an efficient method to extract feature with maximal dependency to the target class and minimal redundancy among themselves using only one-dimensional MI estimates. The proposed method has been applied to the problem of the classification of electroencephalogram (EEG) signals for EEG-based brain-computer interface (BCI). Moreover, the results of proposed method was compared to the results obtained using PCA, ICA, MRMI, and SDA. The results confirm that the classification accuracy obtained by Minimax-MIFX is higher than that achieved by existing feature extraction methods and by full feature set.

2. Methods

2.1 Definition of mutual information

Mutual information is a non-parametric measure of relevance between two variables. Shannon's information theory provides a suitable formalism for quantifying this concepts. Assume a random variable $\boldsymbol{X}$ representing continuous-valued random feature vector, and a discrete-valued random variable $\boldsymbol{C}$ representing the class labels. In accordance with Shannon's information theory, the uncertainty of the class label C can be measured by entropy $H(C)$ as

$$H(C) = -\sum_{c \in C} p(c) \log p(c) \tag{1}$$

where $p(c)$ represents the probability of the discrete random variable $\boldsymbol{C}$. The uncertainty about C given a feature vector $\boldsymbol{X}$ is measured by the conditional entropy as

$$H(C|X) = -\int_{\boldsymbol{x}} p(\mathbf{x}) \left(\sum_{c \in C} p(c|\mathbf{x}) \log\ p(c|\mathbf{x}) \right) d\mathbf{x} \tag{2}$$

where $p(c|\mathbf{x})$ is the conditional probability for the variable $\boldsymbol{C}$ given $\boldsymbol{X}$.

In general, the conditional entropy is less than or equal to the initial entropy. It is equal if and only if one has independence between two variables $\boldsymbol{C}$ and $\boldsymbol{X}$**.** The amount by which the class uncertainty is decreased is, by definition, the mutual information, $I(X;C) = H(C) - H(C|X)$, and after applying the identities $p(c,\mathbf{x}) = p(c|\mathbf{x})p(\mathbf{x})$ and $p(c) = \int_x p(c,\mathbf{x}) d\mathbf{x}$ can be expressed as

$$I(X;C) = \sum_{c \in C} \int_x p(c,\mathbf{x})\ \log \frac{p(c,\mathbf{x})}{p(c)p(\mathbf{x})} d\mathbf{x} \tag{3}$$

If the mutual information between two random variables is large, it means two variables are closely related. Indeed, MI is zero if and only if the two random variables are strictly independent.

2.2 Minimax mutual information approach to feature extraction

The optimal feature extraction requires creating a new feature set from the original features which jointly have largest dependency on the target class (i.e., maximal dependency). Let us denote by $\mathbf{x}$ the original feature set as the sample of continuous-valued random vector, and by discrete-valued random variable C the class labels. The problem is to find a linear mapping $\mathbf{W}$ such that the transformed features

$$\mathbf{y} = \mathbf{W}^T\mathbf{x} \tag{4}$$

maximizes the mutual information between the transformed features Y and the class labels C, $I(Y,C)$. That is, we seek

$$\mathbf{W}_{opt} = \arg \max_{\mathrm{W}} I(Y;C) \tag{5}$$

$$I(\boldsymbol{Y},C) = \sum_{c\in C} \int\!\cdots\!\int p(y_1 \ldots y_m)\log \frac{p(y_1 \cdots y_m, c)}{p(y_1 \cdots y_m)p(c)} dy_1 \cdots dy_m \tag{6}$$

However, it is not always easy to get an accurate estimation for high-dimensional mutual information. It requires the knowledge on the underlying probability density functions (pdfs) of the data and the integration on these pdfs. Moreover, due to the enormous computational requirements of the method, the practical applicability of the above solution to complex classification problems requiring a large number of features is limited.

To overcome the abovementioned practical obstacle, we propose a heuristic method for feature extraction which is based on minimal-redundancy-maximal-relevance (minimax) framework. The max-relevance and min-redundancy criterion has been already used for feature selection (Battiti, 1994; Kwak & Choi, 2002; Peng et al., 2005). It was proved theoretically that minimax criteria is equivalent to maximal dependency (6) if one feature is added at one time (Peng et al., 2005). This criterion is given by

$$J = \left\{ I(x_i;c) - \beta \sum_{x_s \in S} I(x_i;x_s) \right\} \tag{7}$$

According to this criteria, at each time, a new feature x_i is selected with maximal dependency to the target class (i.e., $\max_{x_i} I(x_i;c)$) and minimal dependency among the new feature and already selected features (i.e., $\min_{x_i} \sum_{x_s \in S} I(x_i;x_s)$). The parameter β is the redundancy parameter which is used in considering the redundancy among input features and regulates the relative importance of the MI between the new extracted feature and the already extracted features with respect to the MI with the output class.

In this paper, we modify these criteria for purpose of feature extraction, namely minimax feature extraction, as follows:

$$J = \left\{ I(y_i;c) - \beta \sum_{y_s \in S} I(y_i;y_s) \right\}; \qquad y_i = \mathbf{w}_i^T\mathbf{x} \tag{8}$$

where y_i and y_s are the new and already extracted features, respectively. The parameter β was assigned the value $1/m$, where m is the number of already extracted features. The proposed feature extraction method is an iterative process which begins with an empty feature set and additional features are created and included one by one such that the criteria (8) maximized. Formally, the problem can be stated as

$$\mathbf{w}_{opt} = \arg \max_{\mathbf{w}} \left\{ I(y_i; c) - \beta \sum_{y_s \in S} I(y_i; y_s) \right\}; \quad y_i = \mathbf{w}_i^T \mathbf{x} \tag{9}$$

We use a genetic algorithm (GA) (Goldberg, 1989) for mutual information optimization and learning the linear mapping $\mathbf{w}$. Unlike many classical optimization techniques, GA does not rely on computing local first- or second-order derivatives to guide the search process; GA is a more general and flexible method that is capable of searching wide solution spaces and avoiding local minima (i.e., it provides more possibilities of finding an optimal or near-optimal solution). To implement the GA, we use Genetic Algorithm and Direct Search Toolbox for use in Matlab (The Mathworks, R2007b). The algorithm starts by generating an initial population of random candidate solutions. Each individual (chromosomes) in the population is then awarded a score based on its performance. The value of the fitness function (i.e., the function to be optimize) for an individual is its score. The individuals with the best scores are chosen to be parents, which are cut and spliced together to make children. The genetic algorithm creates three types of children for the next generation: Elite children, Crossover children, and Mutation children. Elite children are the individuals in the current generation with the best fitness values. These individuals automatically survive to the next generation. Crossover children are created by combining the genes of two chromosomes of a pair of parents in the current population. Mutation, on the other hand, arbitrarily alters one or more genes of a selected chromosome, by a random change with a probability equal to the mutation rate. These children are scored, with the best performers likely to be parents in the next generation. After some number of generations, it is hoped that the system converges with a near-optimal solution.

In this application, the genetic algorithm is run for 70 generations with population size of 20, crossover probability 0.8, and uniform mutation probability of 0.01. The number of individuals that automatically survive to the next generation (i.e., elite individuals) is selected to be 2. The scattered function is used to create the crossover children by creating a random binary vector and selects the genes where the vector is a 1 from the first parent, and the genes where the vector is a 0 from the second parent.

One is to implement MI-based feature extraction scheme, estimation of MI always poses a great difficulties as it requires the knowledge on the underlying probability density functions (pdfs) of the data and the integration on these pdfs. One of the most popular ways to estimate mutual information for low-dimensional data space is to use histograms as a pdf estimator. Histogram estimators can deliver satisfactory results under low-dimensional data spaces. Trappenberg et al. (2006) have compared a number of MI estimation algorithms including standard histogram method, adaptive partitioning histogram method (Darbellay & Vajda, 1999), and MI estimation based on the Gram-Charlier polynomial expansion (Trappenberg et al., 2006). They have demonstrated that the adaptive partitioning histogram method showed superior performance in their examples. In this work, we used a two-dimensional mutual information estimation using adaptive partitioning histogram method.

The minimax MI-based feature extraction can be summarized by the following procedure:

1. Initialization:
   - Set $\mathbf{x}$ to the initial feature set;
   - Set $\mathbf{s}$ to the empty set;
2. Feature extraction (repeat until desired number of features are extracted).
   - Set $J = \left\{ I(\mathbf{w}_i^T \mathbf{x}, c) - \beta \sum_{y_s \in S} I(\mathbf{w}_i^T \mathbf{x}; y_s) \right\}$ as the fitness function;
   - Initialize the GA;
     - Specify type, size, and initial values of population;
     - Specify the selection function (i.e., how the GA chooses parents for the next generation);
     - Specify the reproduction operators (i.e., how the genetic algorithm creates the next generation)
   - Find the weighting vector that maximize the fitness function and denote it as $\mathbf{w}_{opt}$;
   - Extract the feature, $y = \mathbf{w}_{opt}^T \mathbf{x}$;
   - Put y into $\mathbf{s}$;
3. Output the set $\mathbf{s}$ containing the extracted features.

3. Experimental setup and data set

3.1 Our experiments

The EEG data of five healthy right-handed volunteer subjects were recorded at a sampling rate of 256 from positions Cz, T5, Pz, F3, F4, Fz, and C3 by Ag/AgCl scalp electrodes placed according to the International 10-20 system. The eye blinks were recorded by placing an electrode on the forehead above the left brow line. The signals were referenced to the right earlobe. Data were recorded for 5 *s* during each trial experiment and low-pass filtered with a cutoff 45 Hz. Depending on the cue visual stimuli which was appeared on the monitor of computer at 2 s, the subject imagines either right-hand grasping or right-hand opening. If the visual stimuli was not appeared, the subject did not perform a specific task. In the present study, the tasks to be discriminated were the imagination of hand grasping and the idle state. The imaginative hand movement can be hand closing or hand opening. There were 200 trails acquired from each subject during each experiment day.

One of the major problems in developing an online EEG-based BCI is the ocular artifact suppression. In this work, eye blink artifacts are suppressed automatically by using a neural adaptive noise canceller (NANC) proposed in (Erfanian & Mahmoudi, 2005). The structure of adaptive noise canceller is shown in Fig. 1. The primary signal is the measured EEG data from one of the EEG channels. The reference signal is the data recorded from the forehead electrode. Here the adaptive filter is implemented by means of a multi layer perceptron neural network.

3.2 BCI competition III-data set IIIb

To validate the proposed MI-based feature extraction and classification methods for brain-computer Interfaces, the algorithms were also applied to the data set IIIb of "BCI Competition III (Blankertz et al., 2006).". This data set contained 2-class EEG data from 3

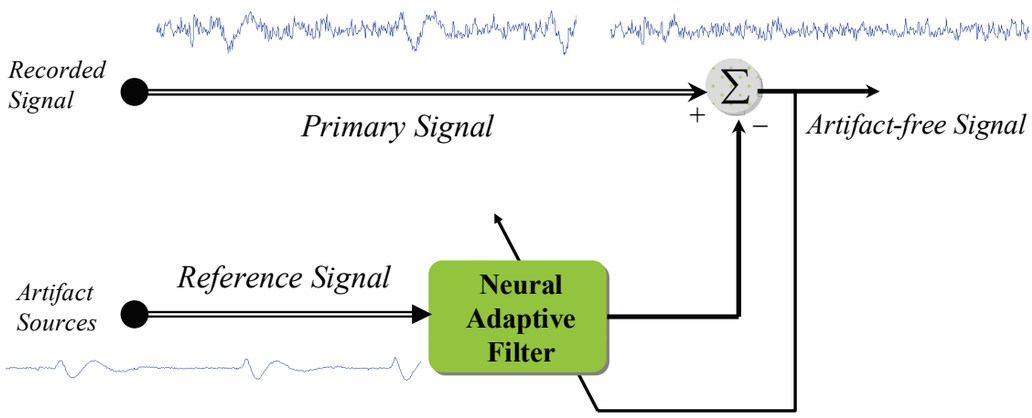


Fig. 1. The structure of the neural adaptive noise canceller used for online ocular artifact suppression.

Subjects. Each data set contained recordings from consecutive sessions during a BCI experiment.Thee experiment consists of 3 sessions for each subject. Each session consisted of 9 runs and each run consisted of 40 feedback trials. For each subject the total number of trials is 1080. The recordings were made with a bipolar EEG amplifier from g.tec (Guger Technologies OEG Austria) . The EEG was sampled with 125 Hz, it was filtered between 0.5 and 30 Hz with Notchfilter on. The experiment was based on the basket paradigm (Vidaurre et al., 2006). In each trial, the subject saw a black screen for a fixed length pause (3 s). Then, two different colored baskets (green and red) appeared at the bottom of the screen. At this moment, also a little green ball appeared at the top of the screen. After 1 s more, the ball began to fall downward with constant speed. The horizontal position of the ball was directly controlled by the output of the classifier. The subject's task was to control the green ball by the imagination of left- or right-hand movements, and try to keep it as long as possible in the side where the green basket appeared. The duration of each trial was 7 s,

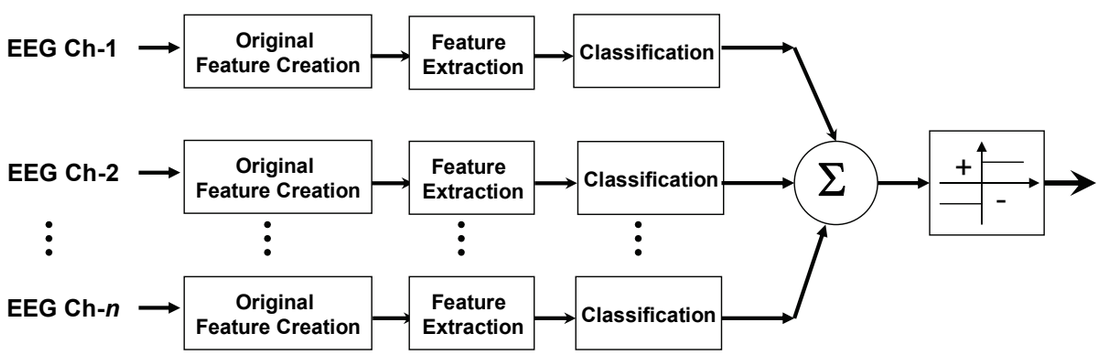


Fig. 2. The block diagram of a multiple classifier for EEG classification.

3.3 Multiple classifier

A multiple classifier is employed for classification of extracted feature vectors. The *Multiple Classifier is* used if different sensors are available to give information on one object. Each of

the classifiers works independently on its own domain. The single classifiers are built and trained for their specific task. The final decision is made on the results of the individual classifiers. In this work, for each EEG channel, separate classifier is trained and the final decision is implemented by a simple logical majority vote function. The desired output of each classifier are -1 or +1. The output of classifiers is added and the *signum function* is used for computing the actual response of the classifier. The block diagram of classification process is shown in Fig. 1. The diagonal linear discrimination analysis (DLDA) (Krzanowski, 2000) is here considered as the classifier. The classifier is trained to distinguish between rest state and imaginative hand movement.

4. Results

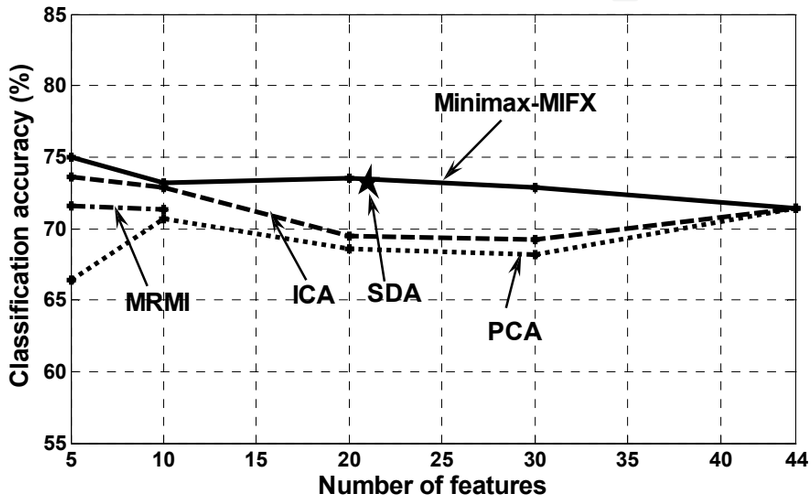
4.1 Our experiments

Original features are formed from 1-s interval of EEG data of each channel, in the time period 2.3-3.3 s, during each trial of experiment. The window starting 0.3 s after cue presentation is used for classification. The number of local extrema within interval, zero crossing, 5 AR parameters, variance, the mean absolute value (MAV), and 1-Hz frequency components between 1 and 35 Hz constitute the full set of features with size 44. In this application, the genetic algorithm was run for 70 generations with population size of 20, crossover probability 0.8, and mutation probability of 0.01. For each channel, one classifier is designed. The classifier is trained to distinguish between rest state and imaginative hand movement. The imaginative hand movement can be hand closing or hand opening. From 200 data sets, 100 sets are randomly selected for training, while the rest is kept aside for validation purposes. Training and validating procedure is repeated 10 times and the results are averaged.

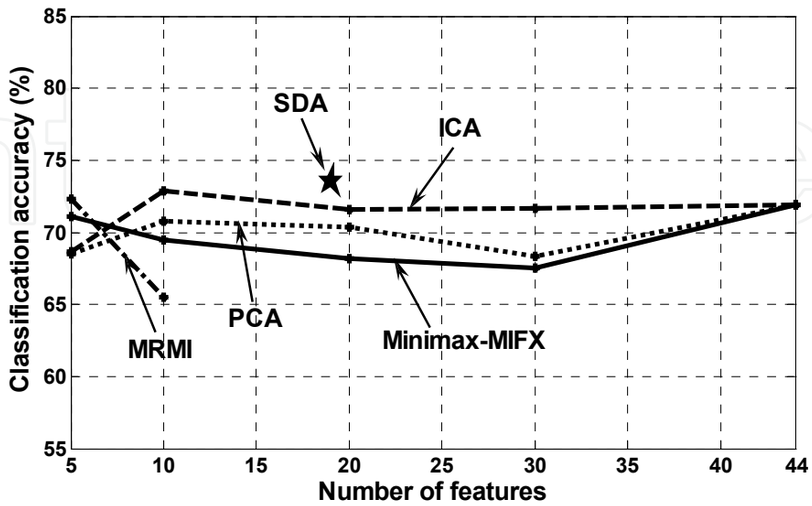
Fig. 3 shows the classification accuracy for subject ST during different experiment days for different sizes of feature set obtained by minimax-MIFX, PCA, MRMI, and ICA methods. During the first day, the best classification accuracy as high as 75.0% was obtained using minimax-MIFX with 5 features. During the second day, the best results obtained are 72.9% with 10 features using ICA, 72.3% using MRMI and 71.1% using Minimax-MIFX with 5 features, and 71.9% using full feature set. During the third experiment day, the best classification accuracy obtained is 83.4% by using Minimax-MIFX with 5 features, while the rate is 74.0% with full feature set. Fig. 2 (d) shows the average classification accuracies over three experiment days for the subject ST. It is observed that the Minimax-MIFX method provides a better performance compared to the other feature extraction methods. On average, the best rate for the subject ST is 76.5% which is obtained by Minimax-MIFX method with 5 extracted features. The average classification performance of SDA for the subject ST is 73.96% which is poorer than that obtained by the Minimax-MIFX. The performance for full feature set is 72.43%. It is observed that the best performance of MRMI method takes place when the number of extracted to be small. It should be noted that the MRMI method is subject to the curse of dimensionality as the number of extracted feature increases (Hild II et al., 2006). Due to this fact and low computation speed of MRMI, this method are performed for extraction of 5 and 10 features.

Fig. 4 shows the average of classification accuracies over three days for all other subjects. The best classification accuracy is obtained by the Minimax-MIFX in all subject and is 78.4% with 5 features in AE, 80.0% with 10 features in ME, 78.37% with 20 features in BM, and 78.3% with 10 features in MM. Fig. 3(e) shows the average of classification accuracy over all

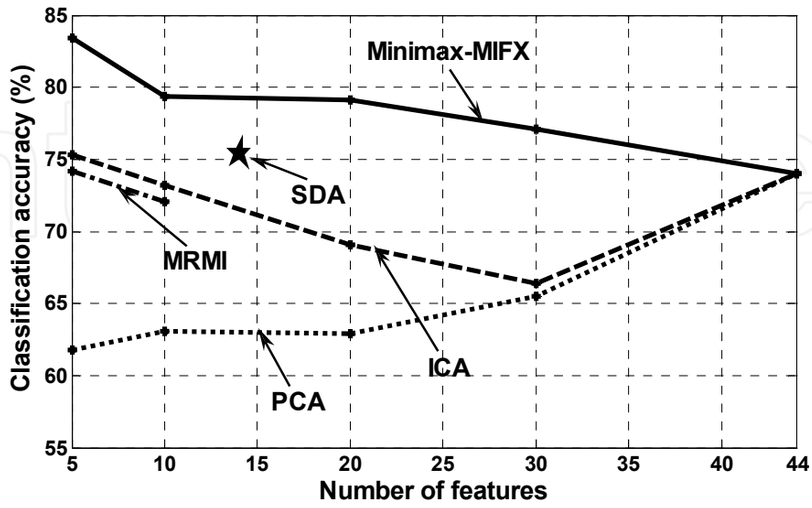
subjects. The classification performance obtained using ICA method is almost the same as that obtained using PCA. The best performance of MRMI method is achieved when five extracted features are used for classification. However, the performance of MRMI degrades as the number of extracted features increases. The results indicate that classification accuracy obtained by the Minimax-MIFX method is generally better than that obtained by other methods. The best classification accuracy as high as 78.0% is obtained by minimax-MIFX method only with 5 extracted features. The average performance of SDA is 77.85% which is identical to that obtained using minimax-MIFX.



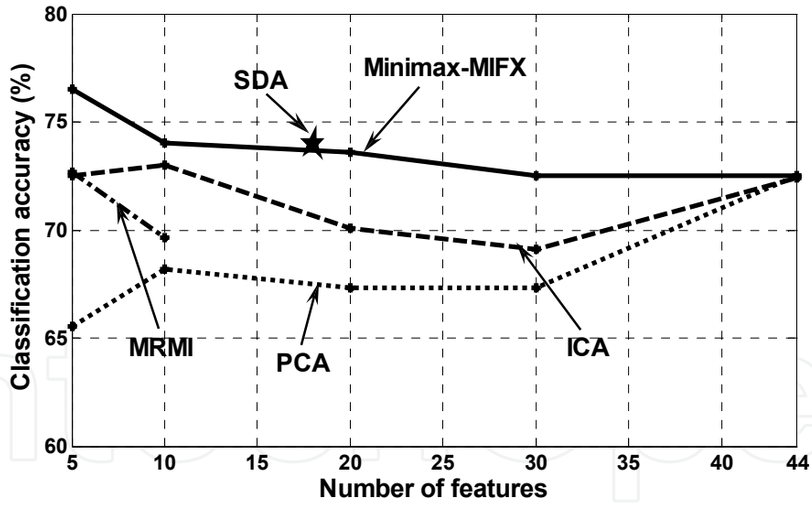
(a)



(b)

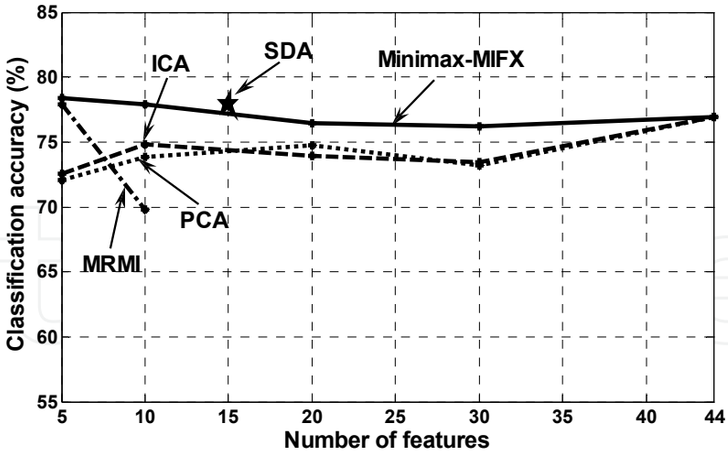


(c)

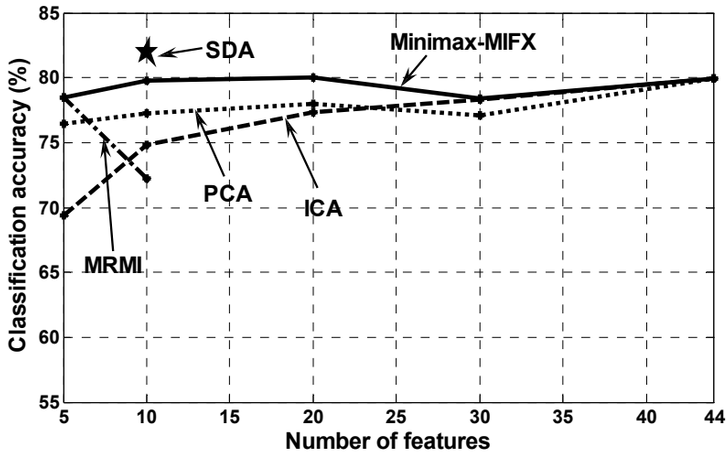


(d)

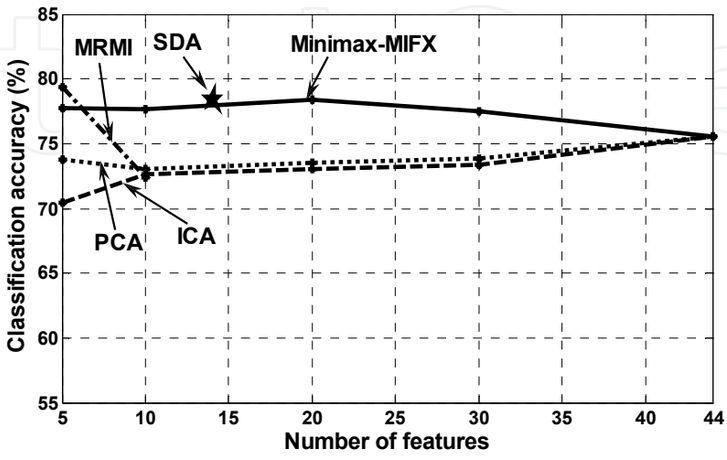
Fig. 3. Classification accuracy for subject ST with different sizes of feature set obtained by different feature extraction methods: (a-c) Different experiment days. (d) Average classification accuracy over different days.


Classification accuracy (%)
Number of features
ICA
SDA
Minimax-MIFX
PCA
MRMI

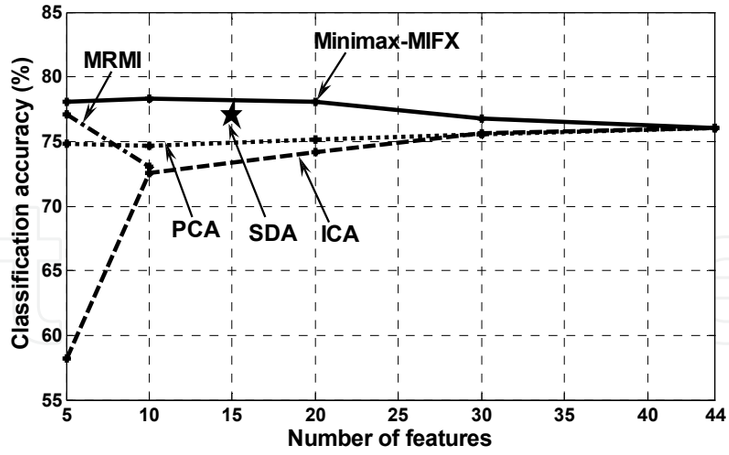
(a)


Classification accuracy (%)
Number of features
SDA
Minimax-MIFX
PCA
ICA
MRMI

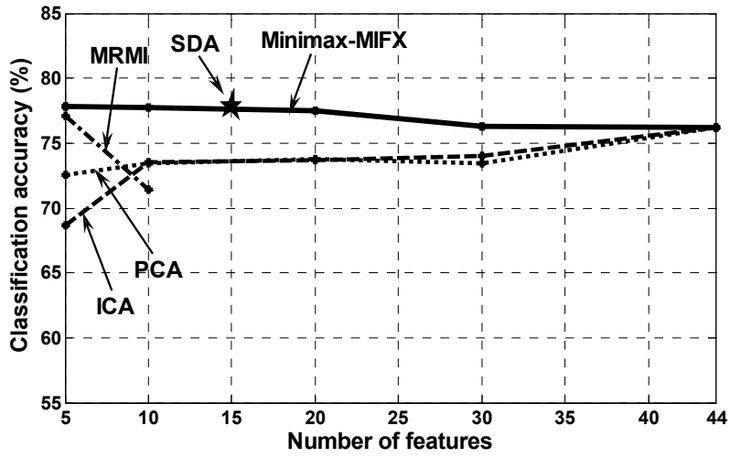
(b)


Classification accuracy (%)
Number of features
MRMI
SDA
Minimax-MIFX
PCA
ICA

(c)



(d)

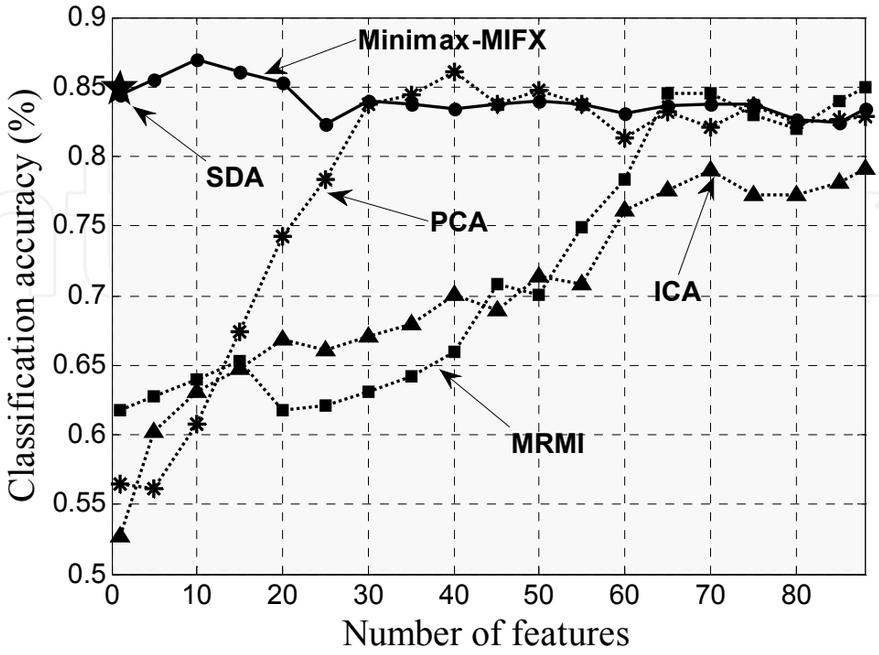


(e)

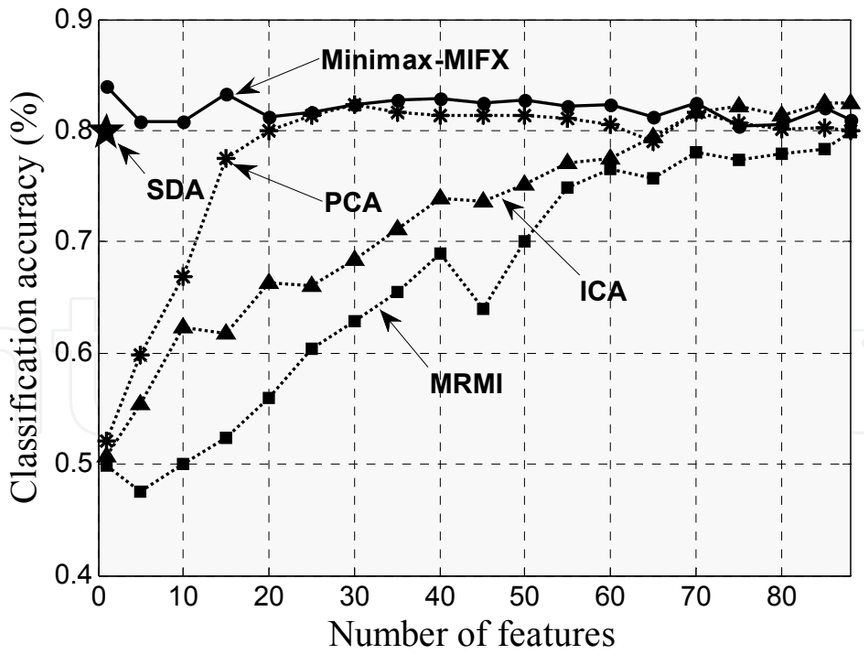
Fig. 4. The average of classification accuracy over the three days for the subjects AE (a), ME (b), BM (c), and MM (d). Average classification accuracy over all days and all subjects (e).

4.2 BCI Competition III-Data Set IIIb

For classification, the features are extracted from 3-s epoch of EEG data recorded from channels C3 and C4 in the interval 4-7 s. A classifier is trained to differentiate between EEG patterns associated with left- and right-hand movement imagery. The entire feature set are formed from each data window, separately and consist of 23 features including the number of local extrema within interval, zero crossing, energy of 8 wavelet packet nodes of a three level decomposition, 5 AR parameters, variance, the mean absolute value (MAV), the first three eigenvalues of correlation matrix, and the relative power in three common frequency bands of EEG spectral density—theta (4–8 Hz), alpha (9–14 Hz), and beta (15–30 Hz). From 1080 feature sets, 540 sets are assigned for training of each classifier, while the rest is kept aside for validation purposes.



(a)



(b)

Fig. 5. Classification accuracy obtained by using different feature extraction methods for BCI Competition III-Data Set IIIb for two subjects S4 (a) and X11 (b).

Fig. 5 shows the classification accuracies obtained by different feature extraction methods for different number of extracted features. In subject S4, the best classification accuracies obtained are 87.0% using minimax-MIFX with 10 extracted features, 86.1% using PCA with 36 features, 79.1% using ICA with 68 features, 85.1% using MRMI with 70 extracted features, and 77.2% using full feature set. In subject X11, it is observed that the best classification accuracy was achieved by using the proposed feature extraction method which is 84.1% with only one extracted feature. The accuracy rate with full feature set in subject X11 is 82.9%.
The results show that that minimax-MIFX provides a robust performance against changes in the number of features extracted, while the performance of other feature extraction methods is sensitive with respect to the number of features.

5. Conclusion

In this paper, we have proposed a novel approach for feature extraction which is based on mutual information. The goal of mutual information-based feature extraction (MIFX) is to create new features from transforming the original features such that the dependency between the transferred features and the target class is maximized. However, the estimation of MI poses great difficulties as it requires the estimating the multivariate probability density functions (pdfs) of the data space and the integration on these pdfs. The proposed MIFX method iteratively creates a new feature with maximal dependency to the target class and minimal redundancy among the new feature and previously extracted features. Our minimax-MIFX scheme avoids the difficult multivariate density estimation in maximizing dependency and minimizing redundancy. Only two-dimensional (2-D) MIs are directly estimated, whereas the higher dimensional MIs are analyzed using the 2-D MI estimates. The effectiveness of the proposed method was evaluated by using the classification of EEG signals during hand movement imagination and the results compared to the performance of some existing feature extraction methods including PCA, ICA, SDA, and MRMI. Moreover, the MIFX algorithms were also applied to the data set IIIb of BCI Competition III. The results demonstrate that the classification accuracy can be improved by using the proposed feature extraction scheme compared to the existing feature extraction methods.

6. References

Battiti, R. (1994). Using mutual information for selecting features in supervised neural net learning. *IEEE Transaction Neural Networks*, Vol. 5, No. 4, (July 1994) pp. (537-550), ISSN 1045-9227.

Blankertz B. & Muller, K. (2004). The BCI competition 2003: progress and perspectives in detection and discrimination of EEG single trials, *IEEE Transaction Biomedical Engineering*, Vol. 51, No. 6, (June 2004) pp. (1044-1051), ISSN 0018-9294.

Blankertz, B.; Müller, K. R.; Krusienski, D.; Schalk, G.; Wolpaw, J. R.; Schlögl, A.; Pfurtscheller, G.; Millán, J. R.; Schröder, M. & Birbaumer, N. (2006). The BCI competition III: validating alternative approaches to actual BCI problems. *IEEE Transaction on Neural Systems and Rehabilitation Engineering*, Vol. 14, No. 2, (June 2006) pp. (153-159), ISSN 1534-4320.

Chow, T. & Huang, D. (2005). Estimating optimal feature subsets using efficient estimation of high-dimensional mutual information. *IEEE Transaction Neural Network*, Vol. 16, No. 1, (January 2005) pp. (213-224), ISSN 1045-9227.

Darbellay, G. & Vajda, I. (1999). Estimation of the information by an adaptive partitioning of the observation space. *IEEE Transaction Information Theory*, Vol. 45, No. 4, (May 1999) pp. (1315-1321), ISSN 0018-9448.

Duda, R.O.; Hart, P.E. & Stork, D. (2001). *Pattern Classification*. Wiley, second edition ISBN 0-471-05669-3.

Erfanian, A. & Erfani, A. (2004). ICA-based classification scheme for EEG-based brain-computer interface: the role of mental practice and concentration skills. *26th Annual International Conference of the IEEE/EMBS*, pp. 235 – 238, ISBN 0-7803-8439-3, USA, September 2004, San Francisco.

Erfanian, A. & Mahmoudi, B. (2005). Real-time ocular artifacts suppression using recurrent neural network for EEG-based brain computer interface, *Medical & Biological Engineering & Computation*, Vol. 43, No. 2, (March 2005) pp. (296-305), ISSN 0140-0118.

Estèvez, P. A.; Tesmer, M.; Perez, C. A. & Zurada, J. M. (2009). Normalized mutual information feature selection. *IEEE Transaction Neural Networks*, Vol. 20, No. 2, (February 2009) pp. (189 - 201), ISSN 1045-9227.

Goldberg, D. E. (1989). Genetic algorithms in search, *Optimization and Machine Learning*. Addison-wesley, ISBN 0201157675, Boston, MA, USA.

Hild II, K. E.; Erdogmnus, D.; Torkkola, K. & Principe, J. C. (2006). Feature extraction using information-theoretic learning. *IEEE Transaction Pattern Analysis and Machine Intelligence,* Vol. 28, No. 9, (September 2006) pp. (1385-1393), ISSN 0162-8828.

Hyvarinen, A.; Karhunen, J. & Oja, E. (2001). *Independent Component Analysis,* John Wiley & Sons, ISBN 9780471405405.

Krzanowski, WJ. (2000). Principles of multivariate analysis: *a user's perspective,* Oxford University Press, Oxford, ISBN 9780198507086.

Kwak, N. (2007). Feature extraction based on direct calculation of mutual information. *International Journal of Pattern Recognition and Artificial Intelligence,* Vol. 21, No. 7, (november 2007) pp. (1213-1231), ISSN 0218-0014.

Kwak, N. & Choi, C. H. (2002). Input feature selection for classification problems. *IEEE Transaction Neural Networks*, Vol. 13, No. 1, (January 2002) pp. (143–159), ISSN 1045-9227.

Kwak, N. & Choi, C. H. (2003). Feature extraction based on ICA for binary classification problems. *IEEE Transaction Knowledge and Data Engineering*, Vol. 15, No. 6, (November/December 2003) pp. (1374-1388), ISSN 1041-4347.

Lemm, S.; Schäfer, C. & Curio, G. (2004). BCI competition 2003—data set III: probabilistic modeling of sensorimotor μ rhythms for classification of imaginary hand movements. *IEEE Transaction Biomedical Engineering*, Vol. 51, No. 6, (June 2004) pp. (1077–1080), ISSN 0018-9294.

Li, H.; Jiang, T. & Zhang, K. (2006). Efficient and robust feature extraction by maximum margin criterion. *IEEE Transaction Neural Networks*, Vol. 17, No. 1, (January 2006) pp. (157–165), ISSN 1045-9227.

Loog, M. & Duin, R. P. W. (2004). Linear dimensionality reduction via a heteroscedastic extension of LDA: the Chernoff criterion. *IEEE Transaction Pattern Analysis and Machine Intelligence*, Vol. 26, No. 6, (June 2004) pp. (732-739), ISSN 0162-8828.

Murillo, J. & Rodriguez, A. (2007). Maximization of mutual information for supervised linear feature extraction. *IEEE Transaction Neural Network*, Vol. 18, No. 5, (September 2007) pp. (1433-1441), ISSN 1045-9227.

Peng, H.; Long, F. & Ding, C. (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transaction Pattern Analysis and Machine Intelligence*, Vol. 27, No. 8, (august 2005) pp. (1226-1238), ISSN 0162-8828.

Sindhwani, V.; Rakshit, S.; Deodhar, D.; Erdogmus, D.; Principe, J. & Niyogi, P. (2004). Feature selection in MLPs and SVMs based on maximum output information. *IEEE Transaction Neural Networks*, Vol. 15, No. 4, (July 2004.) pp. (937–948), ISSN 1045-9227.

Torkkola, K. (2003). Feature extraction by non-parametric mutual information maximization. *Journal of Machine Learning Research*, Vol. 3, No. 7-8, (March 2003) pp. (1415–1438), ISSN 1533-7928.

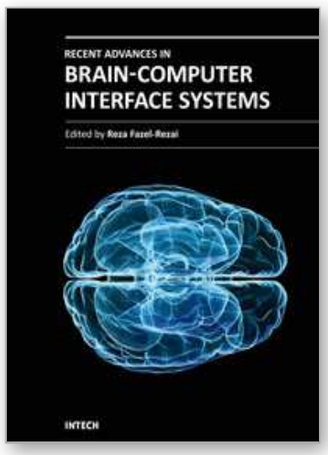
Trappenberg, T.; Ouyang J. & Back A. (2006). Input variable selection: mutual information and linear mixing measures. *IEEE Transaction Knowledge and Data Engineering*, Vol. 18, No. 1, (january 2006) pp. (37-46), ISSN 1041-4347.

Vidaurre, C.; Schlöogl, A.; Cabeza, R.; Scherer, R. & Pfurtscheller, G. (2006). A fully on-line adaptive BCI. *IEEE Transaction Biomedical Engineering*, Vol. 53, No. 6, (June 2006) pp. (1214-1219), ISSN 0018-9294.

Yu, H. & Yang, J. (2001). A direct LDA algorithm for high-dimensional data—with applications to face recognition. *Pattern Recognition*, Vol. 34, No. 11, (October 2001) pp. (2067-2070), ISSN 0031-3203.

Zhu, M. & Martinez, A. M. (2006). Subclass discriminant analysis. *IEEE Transaction Pattern Analysis and Machine Intelligence,* Vol. 28, No. 8, (august 2006) pp. (1274-1286), ISSN 0162-8828.

IntechOpen



Recent Advances in Brain-Computer Interface Systems
Edited by Prof. Reza Fazel

ISBN 978-953-307-175-6
Hard cover, 222 pages
Publisher InTech
Published online 04, February, 2011
Published in print edition February, 2011

Brain Computer Interface (BCI) technology provides a direct electronic interface and can convey messages and commands directly from the human brain to a computer. BCI technology involves monitoring conscious brain electrical activity via electroencephalogram (EEG) signals and detecting characteristics of EEG patterns via digital signal processing algorithms that the user generates to communicate. It has the potential to enable the physically disabled to perform many activities, thus improving their quality of life and productivity, allowing them more independence and reducing social costs. The challenge with BCI, however, is to extract the relevant patterns from the EEG signals produced by the brain each second. Recently, there has been a great progress in the development of novel paradigms for EEG signal recording, advanced methods for processing them, new applications for BCI systems and complete software and hardware packages used for BCI applications. In this book a few recent advances in these areas are discussed.

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following:

Abbas Erfanian, Farid Oveisi and Ali Shadvar (2011). Feature Extraction by Mutual Information Based on Minimal-Redundancy-Maximal-Relevance Criterion and Its Application to Classifying EEG Signal for Brain-Computer Interfaces, Recent Advances in Brain-Computer Interface Systems, Prof. Reza Fazel (Ed.), ISBN: 978-953-307-175-6, InTech, Available from: http://www.intechopen.com/books/recent-advances-in-brain-computer-interface-systems/feature-extraction-by-mutual-information-based-on-minimal-redundancy-maximal-relevance-criterion-and

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

© 2011 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike-3.0 License, which permits use, distribution and reproduction for non-commercial purposes, provided the original is properly cited and derivative works building on this content are distributed under the same license.

IntechOpen

IntechOpen