Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2008, Article ID 673040, 8 pages
doi:10.1155/2008/673040

Research Article

A Minimax Mutual Information Scheme for Supervised Feature Extraction and Its Application to EEG-Based Brain-Computer Interfacing

Farid Oveisi and Abbas Erfanian

Department of Biomedical Engineering, Faculty of Electrical Engineering, Iran University of Science and Technology, Narmak, Tehran 16844, Iran

Correspondence should be addressed to Abbas Erfanian, erfanian@iust.ac.ir

Received 5 December 2007; Revised 29 May 2008; Accepted 3 July 2008

Recommended by Chein-I Chang

This paper presents a novel approach for efficient feature extraction using mutual information (MI). In terms of mutual information, the optimal feature extraction is creating a feature set from the data which jointly have the largest dependency on the target class. However, it is not always easy to get an accurate estimation for high-dimensional MI. In this paper, we propose an efficient method for feature extraction which is based on two-dimensional MI estimates. At each step, a new feature is created that attempts to maximize the MI between the new feature and the target class and to minimize the redundancy. We will refer to this algorithm as Minimax-MIFX. The effectiveness of the method is evaluated by using the classification of electroencephalogram (EEG) signals during hand movement imagination. 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.

1. INTRODUCTION

Classification of the EEG signals associated with mental tasks plays an important role in the performance of the most EEG-based brain-computer interface (BCI) and reducing the dimensionality of the raw input variable space is an essential preprocessing step in the classification process. There are two main reasons to keep the dimensionality of the input features as small as possible: computational cost and classification accuracy. It has been observed that added irrelevant features may actually degrade the performance of classifiers if the number of training samples is small relative to the number of features [1]. These problems can be avoided by selecting relevant features (i.e., feature selection) or extracting new features containing maximal information about the class label from the original ones (i.e., feature extraction).

A variety of linear feature extraction methods have been proposed. One well-known feature extraction methods may be principal component analysis (PCA) [2]. 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 optimally 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 dimensional reduction algorithm for supervised feature extraction [3]. 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 [4, 5]. Recently, a method based on discriminant analysis (DA) was proposed, known as subclass discriminant analysis (SDA), for describing a large number of data distributions [6]. 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 [7]. 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. In [8], a method was proposed 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 [9, 10] 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. However, it is almost impossible to get an accurate estimation for high-dimensional mutual information. In [11, 12], 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 nonparametric estimate of the mutual information. 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 [12]. To overcome the difficulties of MI estimation for feature extraction, Parzen window modeling was also employed to estimate the probability density function [13]. However, Parzen model may suffer from the "curse of dimensionality," which refers to the overfitting of the training data when their dimension is high [14]. Due to this difficulty, some recent works on information-theoretic learning have proposed the use of alternative measures for MI [14], 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 two-dimensional MI estimates. The proposed method has been applied to the problem of the classification of EEG signals during hand movement imagination. Moreover, the results of proposed method was compared to the results obtained using PCA, ICA, MRMI, and SDA.

2. METHODS

2.1. Definition of mutual information

Mutual information is a nonparametric measure of relevance between two variables. Shannon's information theory provides a suitable formalism for quantifying these concepts. Assume a random variable $\mathbf{X}$ representing continuous-valued random feature vector, and a discrete-valued random variable $\mathbf{C}$ 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 $\mathbf{C}$. The uncertainty about C given a feature vector $\mathbf{X}$ is measured by the conditional entropy as

$$H(C \mid X) = -\int_x p(\mathbf{x})\left(\sum_{c\in C} p(c \mid \mathbf{x})\log p(c \mid \mathbf{x})\right)d\mathbf{x}, \tag{2}$$

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

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 $\mathbf{C}$ and $\mathbf{X}$. The amount by which the class uncertainty is decreased is, by definition, the mutual information, $I(X;C) = H(C) - H(C \mid X)$, and after applying the identities $p(c,\mathbf{x}) = p(c \mid \mathbf{x})p(\mathbf{x})$ 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

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

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

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

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

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 [15–17]. It was proved theoretically that minimax criteria is equivalent to maximal dependency (6) if one feature is added at one time [17]. 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., $\max_i I(x_i; c)$) and minimal dependency among the new feature and already selected features (i.e., $\min_i \sum_{x_s\in S} I(x_i; x_s)$). 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 modified this criterion for purpose of feature extraction, namely minimax feature extraction, as follows:

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

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) is maximized. Formally, the problem can be stated as

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

We use a genetic algorithm (GA) [18] 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. [19] have compared a number of MI estimation algorithms including standard histogram method, adaptive partitioning histogram method [20], and MI estimation based on the Gram-Charlier polynomial

expansion [19]. 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 proposed MI-based feature extraction can be summarized by the following procedure:

(i) initialization:

  (a) set $\mathbf{x}$ to the initial feature set;
  (b) set $\mathbf{s}$ to the empty set;

  feature extraction (repeat until desired number of features are extracted):

(ii) (a) set $J = \{I(\mathbf{w}_i^T\mathbf{x}, c) - \beta\sum_{y_s \in S} I(\mathbf{w}_i^T\mathbf{x}, y_s)\}$ as the fitness function;
  (b) initialize the GA;
    (1) specify type, size, and initial values of population;
    (2) specify the selection function (i.e., how the GA chooses parents for the next generation);
    (3) specify the reproduction operators (i.e., how the genetic algorithm creates the next generation)
  (c) find the weighting vector that maximizes the fitness function and denote it as $\mathbf{w}_{\text{opt}}$;
  (d) extract the feature, $y = \mathbf{w}_{\text{opt}}^T\mathbf{x}$;
  (e) put y into $\mathbf{s}$;

(iii) 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 seconds 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 seconds, 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 EEG-based BCI is the eye blink artifact suppression. The traditional method of the eye blink suppression is the removal of the segment of EEG data in which eye blinks occur. This scheme is rigid and does not lend itself to adaptation. Moreover, a great number of data is lost. To overcome these problems and to shorten the experimental session, we have already developed an adaptive noise canceller (ANC) filter using artificial neural network for real-time removing the eye blinks interference from the EEG signals [21]. In this work, we use this method for real-time ocular artifact suppression without any visual inspection.

3.2. BCI competition 2003-data set III

To validate the proposed MI-based feature extraction and classification methods for brain-computer Interfaces, the algorithms were also applied to the data set III of "BCI Competition 2003" which is obtained by Graz group [22]. This data set was recorded from a healthy subject during a feedback session. Three bipolar EEG channels were measured over C3, Cz, and C4. EEG signals were sampled with 128 Hz and was filtered between 0.5 and 30 Hz. The task was to control a feedback bar in one dimension by imagination of left- or right-hand movements. The experiment included seven runs with 40 trials each. All runs were conducted on the same day with breaks of several minutes in between. The data set consists of 280 trials of 9 seconds length. The first 2 seconds were quiet. At $t = 2$ seconds, an acoustic stimulus indicated the beginning of the trial, and a cross ("+") was displayed for 1 seconds. Then, at $t = 3$ seconds, an arrow (left or right) was displayed as a cue stimulus. The subject was asked to use imagination as described above to move the feedback bar into the direction of the cue.

3.3. Multiple classifiers

Multiple classifiers are employed for classification of extracted feature vectors. The *Multiple Classifier* s are 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 is -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 Figure 1. The diagonal linear discrimination analysis (DLDA) [23] 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 1second interval of EEG data of each channel, in the time period 2.3–3.3 seconds, during each trial of experiment. The window starting 0.3 seconds after cue presentation is used for classification. The number of local extrema within interval, zero crossing, 5 AR

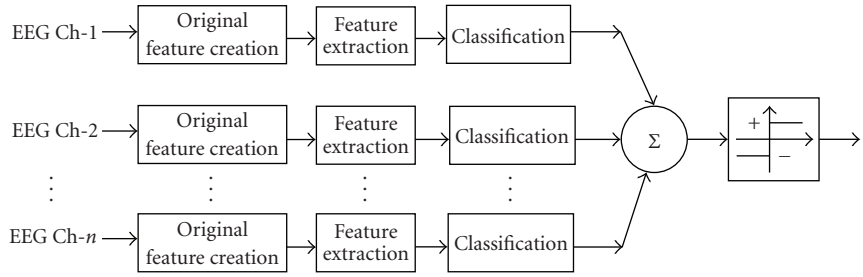

Figure 1: The block diagram of classification process.

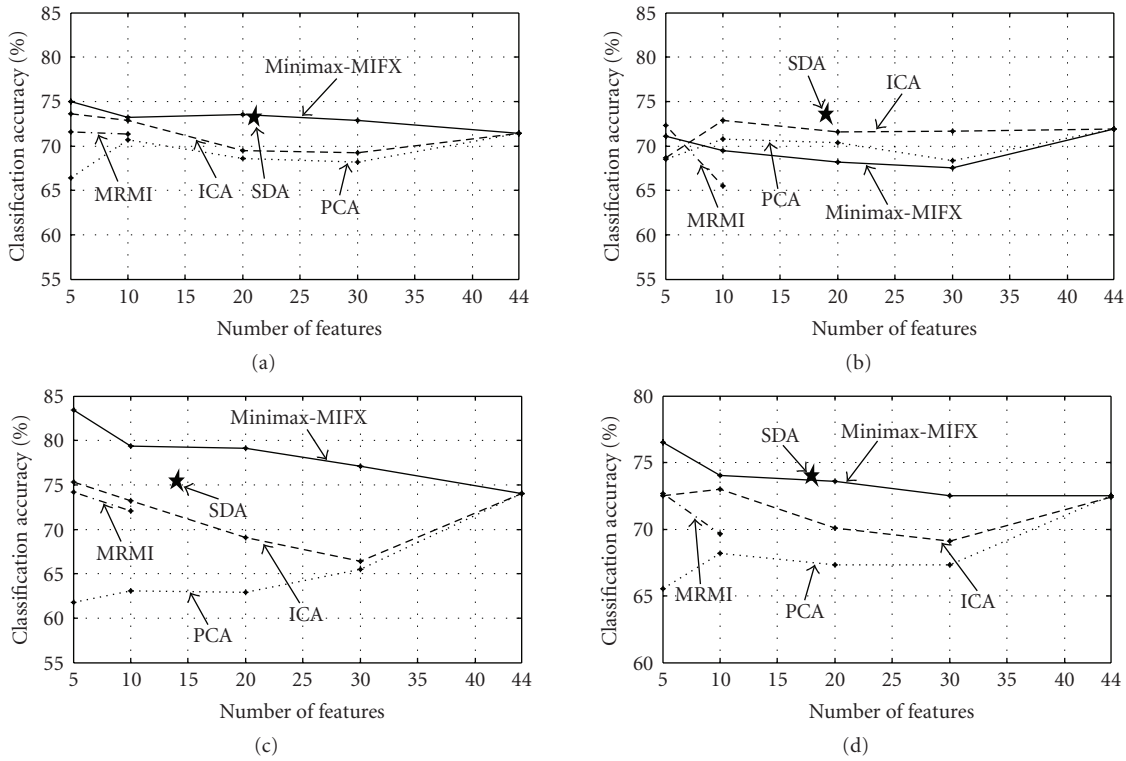

Figure 2: Classification accuracy for subject ST with different sizes of feature set obtained by different feature extraction methods: (a)–(c) different experiment days. (d) Average classification accuracy over different days.

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.01. 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.

Figure 2 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. Figure 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

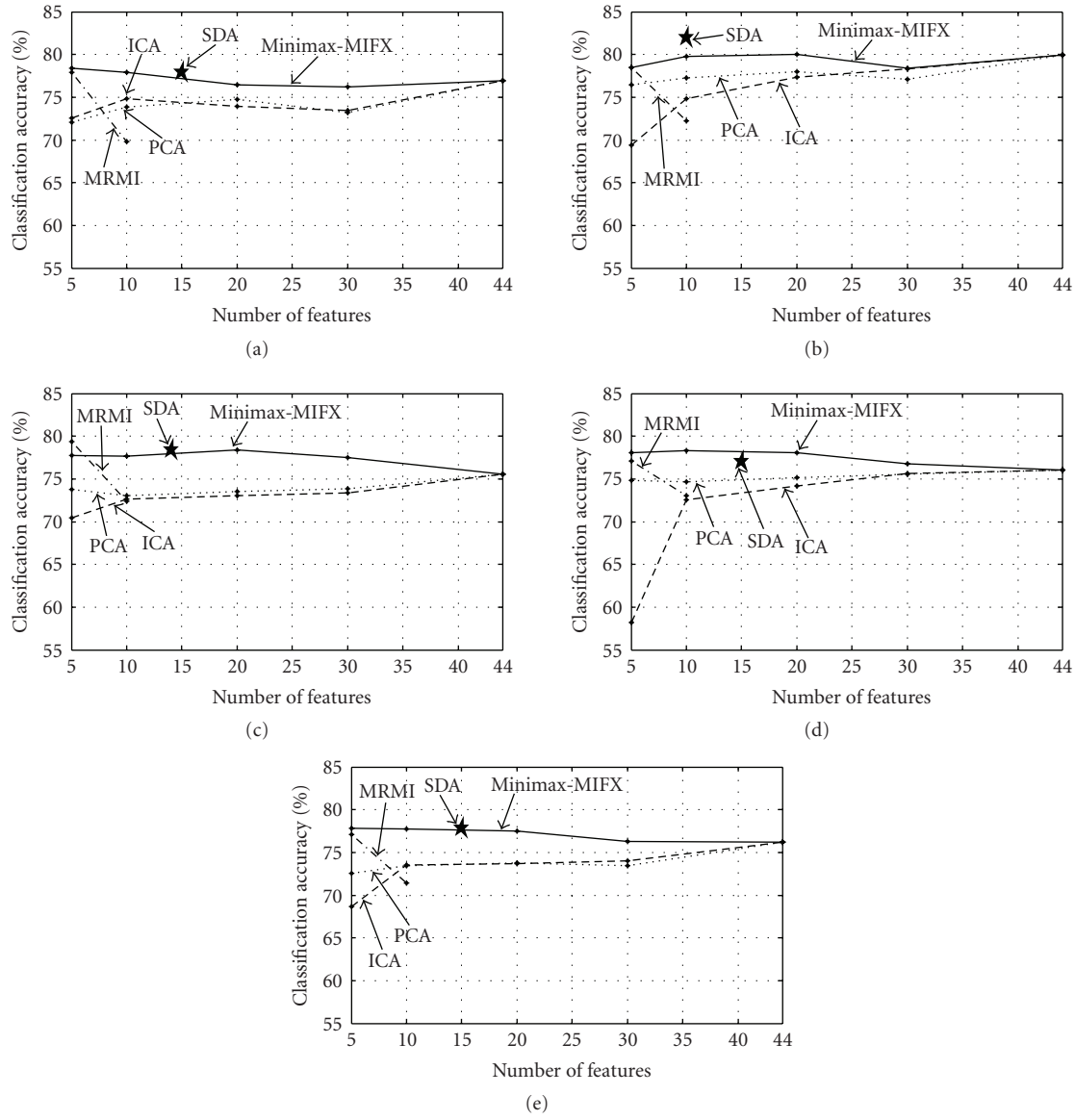

Figure 3: The average of classification accuracy over the three days for the subjects AE (a), ME (b), BM (c), and MM (d). Average classification accuracy over all days and all subjects (e).

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 [12]. Due to this fact and low computation speed of MRMI, this method is performed for extraction of 5 and 10 features.

Figure 3 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 subjects 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. Figure 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.

4.2. BCI competition 2003-data set III

Six 0.7 second intervals of EEG data of each channel (i.e., C3 and C4) are considered during each trial of experiment. The first window starts 0.5 seconds after cue stimulus and all 0.7 seconds windows overlap by 0.2 seconds. For each data window of each channel, one classifier is designed.

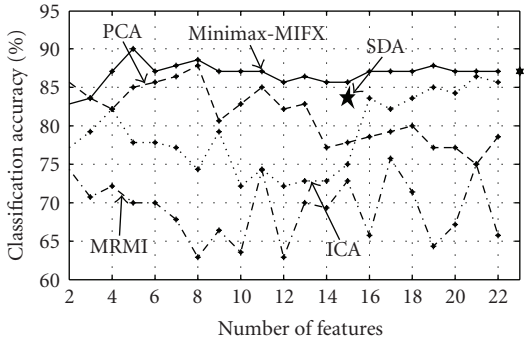

Figure 4: Classification accuracy obtained by using different feature extraction methods for BCI competition 2003-data set III.

The final decision is made on the results of the individual classifiers. The classifiers are trained to differentiate between EEG patterns associated with left- and right-hand movement imagery. The entire feature sets are formed from each data window, separately and consisted 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), 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). Each classifier is trained to differentiate between EEG patterns associated with left- and right-hand movement imagery. For each data window of each channel, one classifier is designed. The final decision is made on the results of the individual classifiers. From 280 data sets, 140 sets are assigned for training of each classifier, while the rest is kept aside for validation purposes. The same data set of "BCI Competition 2003" provided for training and testing are also used here for training and testing, respectively.

Figure 4 shows the classification accuracies obtained by different feature extraction methods for different number of extracted features. It is observed that the best classification accuracy obtained is 90.0% using Minimax-MIFX with 7 extracted features, 87.85% using PCA with 8 features, 86.42% using ICA with 21 features, 75.71% using MRMI with 17 extracted features, and 87.14% using full feature set. It is observed that minimax-FX 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. The performance of SDA for BCI competition data set is 83.57% with 15 extracted features. It is worthy to note that the best rate reported in the BCI competition 2003 for this data set is 89.3% [24].

5. CONCLUSIONS

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 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 (2D) MIs are directly estimated, whereas the higher dimensional MIs are analyzed using the 2D MI estimates. The effectiveness of the MIFX methods is evaluated by using the classification of EEG signals during hand movement imagination. Our comprehensive experiments and BCI Competition 2003-Data Set III—demonstrate that the classification accuracy can be improved by using the proposed feature extraction scheme.

REFERENCES

[1] T. W. S. Chow and D. Huang, "Estimating optimal feature subsets using efficient estimation of high-dimensional mutual information," *IEEE Transactions on Neural Networks*, vol. 16, no. 1, pp. 213–224, 2005.

[2] H. Li, T. Jiang, and K. Zhang, "Efficient and robust feature extraction by maximum margin criterion," *IEEE Transactions on Neural Networks*, vol. 17, no. 1, pp. 157–165, 2006.

[3] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley-Interscience, New York, NY, USA, 2000.

[4] H. Yu and J. Yang, "A direct LDA algorithm for high-dimensional data—with application to face recognition," *Pattern Recognition*, vol. 34, no. 10, pp. 2067–2070, 2001.

[5] R. P. W. Duin and M. Loog, "Linear dimensionality reduction via a heteroscedastic extension of LDA: the Chernoff criterion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 6, pp. 732–739, 2004.

[6] M. Zhu and A. M. Martinez, "Subclass discriminant analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 8, pp. 1274–1286, 2006.

[7] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, New York, NY, USA, 2001.

[8] N. Kwak and C.-H. Choi, "Feature extraction based on ICA for binary classification problems," *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 6, pp. 1374–1388, 2003.

[9] A. Erfanian and A. Erfani, "EEG-based brain-computer interface for hand grasp control: feature extraction by using ICA," in *Proceedings of the 9th Annual Conference of the International Functional Electrical Stimulation Society (IFESS '04)*, Bournemouth, UK, September 2004.

[10] A. Erfanian and A. Erfani, "ICA-based classification scheme for EEG-based brain-computer interface: the role of mental practice and concentration skills," in *Proceedings of the 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEMBS '04)*, vol. 26, pp. 235–238, Francisco, Calif, USA, September 2004.

[11] K. Torkkola, "Feature extraction by non-parametric mutual information maximization," *The Journal of Machine Learning Research*, vol. 3, no. 7-8, pp. 1415–1438, 2003.

[12] K. E. Hild II, D. Erdogmus, K. Torkkola, and J. C. Principe, "Feature extraction using information-theoretic learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 9, pp. 1385–1392, 2006.

[13] N. Kwak, "Feature extraction based on direct calculation of mutual information," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 21, no. 7, pp. 1213–1232, 2007.

[14] J. M. Leiva-Murillo and A. Artes-Rodriguez, "Maximization of mutual information for supervised linear feature extraction," *IEEE Transactions on Neural Networks*, vol. 18, no. 5, pp. 1433–1441, 2007.

[15] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Transactions on Neural Networks*, vol. 5, no. 4, pp. 537–550, 1994.

[16] N. Kwak and C.-H. Choi, "Input feature selection for classification problems," *IEEE Transactions on Neural Networks*, vol. 13, no. 1, pp. 143–159, 2002.

[17] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[18] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley, Reading, Mass, USA, 1989.

[19] T. Trappenberg, J. Ouyang, and A. Back, "Input variable selection: mutual information and linear mixing measures," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 1, pp. 37–46, 2006.

[20] G. A. Darbellay and I. Vajda, "Estimation of the information by an adaptive partitioning of the observation space," *IEEE Transactions on Information Theory*, vol. 45, no. 4, pp. 1315–1321, 1999.

[21] A. Erfanian and B. Mahmoudi, "Real-time ocular artifact suppression using recurrent neural network for electroencephalogram based brain-computer interface," *Medical & Biological Engineering & Computing*, vol. 43, no. 2, pp. 296–305, 2005.

[22] B. Blankertz, K.-R. Müller, G. Curio, et al., "The BCI competition 2003: progress and perspectives in detection and discrimination of EEG single trials," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 6, pp. 1044–1051, 2004.

[23] W. J. Krzanowski, *Principles of Multivariate Analysis: A User's Perspective*, Oxford University Press, Oxford, UK, 2000.

[24] S. Lemm, C. Schäfer, and G. Curio, "BCI competition 2003-data set III: probabilistic modeling of sensorimotor μ rhythms for classification of imaginary hand movements," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 6, pp. 1077–1080, 2004.