Proceedings of the 29th Annual International
Conference of the IEEE EMBS
Cité Internationale, Lyon, France
August 23-26, 2007.

A Tree-Structure Mutual Information-Based Feature Extraction and Its Application to EEG-Based Brain-Computer Interfacing

Farid Oveisi and Abbas Erfanian

Abstract—**This paper presents a novel algorithm for efficient feature extraction using mutual information (MI). In terms of mutual information, the optimal feature extraction is creating a new feature set from the data which jointly have 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 using two-dimensional MI estimates. A new feature is created such that the MI between the new feature and the target class is maximized and the redundancy is minimized. The effectiveness of the proposed algorithm is evaluated by using the classification of EEG signals. The tasks to be discriminated are the imaginative hand movement and the resting state. The results demonstrate that the proposed mutual information-based feature extraction (MIFX) algorithm performed well in several experiments on different subjects and can improve the classification accuracy of the EEG patterns. The results show that the classification accuracy obtained by MIFX is higher than that achieved by full feature set.**

I. 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). 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. 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) [1]. 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 [2], LDA computes a linear transformation by maximizing the ratio of between-class distance to within-class distance, thereby achieving maximal discrimination.

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 [3]. 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 [4], 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 [5]-[6] and demonstrated the improvement of the performance.

One of the most effective approaches for optimal feature selection and extraction is based on mutual information (MI). In [7], a method was proposed 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. 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.

In this paper, we propose a novel method for efficient feature extraction which uses only two-dimensional MI estimates and describe its application to BCI.

Manuscript received April 6, 2007. This work was supported by Iran Telecommunication Research Center, Tehran, IRAN.

A. Erfanian and F. Oveisi are with the Department of Biomedical Engineering, Faculty of Electrical Engineering, Iran University of Science and Technology, Tehran, IRAN (phone: 98-21-73913553; fax: 98-21-77240490; e-mail: erfanian@iust.ac.ir).

II. METHODS

A. 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 **x** representing continuous-valued random feature vector, and a discrete-valued random variable *C* representing the class labels. In accordance with Shannon's information theory, the mutual information can be expressed as

$$I(X,C) = \sum_{c\in C} \int_{\mathbf{x}} p(\mathbf{x},c) \log \frac{p(\mathbf{x},c)}{p(c)\,p(\mathbf{x})} d\mathbf{x}.$$

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.

B. Mutual Information-Based 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. Let us denote by **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 W such that the transformed features

$$\mathbf{y} = \mathbf{wx}$$

maximizes the mutual information between the transformed features *Y* and the class labels *C*, $I(Y,C)$.

However, it is not always easy to get an accurate estimation for high-dimensional mutual information. 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 an efficient tree-structured feature extraction which is based on two-dimensional MI estimates. First the mutual information between each feature x and the class label *c* is computed and selected a feature x_i with minimal dependency. Then the mutual information between selected feature x_i and each feature in the feature set is computed and selected a feature x_j with the maximal dependency. Next, we extract a new feature *y* from the pair of features x_i and x_j such that the mutual information between the projected feature and the class label *c* is maximum. We use a genetic algorithm [8] for mutual information optimization and learning the linear mapping W. The extracted feature *y* contains the information found commonly in two features x_i and x_j about the class label *c*. The new feature is substituted into the original set for the features x_i and x_j. This process is repeated until a desired number of features is extracted.

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.*, [9] have compared a number of MI estimation algorithms including standard histogram method, adaptive partitioning histogram method [10], and MI estimation based on the Gram-Charlier polynomial expansion [9]. 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 mutual information based feature extraction (MIFX) can be summarized by the following procedure:

1) Initialization:
   - Set $\boldsymbol{F} \leftarrow$ "initial set of *n* features."
2) Estimation of the MI between each feature and output class:
   - Compute $I(f_i, C) \;\; \forall \; f_i \in \boldsymbol{F}$.
4) Selection of a feature with minimal dependency on the target class:
   - Find a feature f_i that minimizes $I(f_i, C)$;
   - Set $\boldsymbol{F} \leftarrow \boldsymbol{F} \backslash \{f_i\}$
5) Computation of the MI between the couples of features:
   - Compute $I(f_i, f_j) \;\; \forall \; f_j \in \boldsymbol{F}$.
6) Selection a feature with maximal dependency on f_i:
   - Find a feature f_s that maximizes $I(f_s, f_i)$;
   - Set $\boldsymbol{F} \leftarrow \boldsymbol{F} \backslash \{f_s\}$
7) Extraction of the feature:
   - Extract a feature f from the pair of features (x_i, x_j) such that maximizes $I(f, C)$
   - Set $\boldsymbol{F} \leftarrow \{f\}$;
8) Repeat steps 2-7 until desired number of features are created.
9) Output the set $\boldsymbol{F}$ containing the created features.

C. Multiple Classifiers

A multiple classifiers is employed for classification of extracted feature vectors. The *Multiple Clssifier*s are used if different information sources (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. The decision system can by implemented in many different ways; depending on the problem a simple logical majority vote function, or a rule based expert system may be employed. In this work, for each EEG channel, separate classifier was trained and the final decision was

implemented by a simple logical majority vote function. The multilayer perceptron (MLP) with back-propagation learning rule is used to implement each classifier. The MLP network considered in this study consists of two hidden layers each containing hyperbolic tangent units and one output node. The classifier is trained to distinguish between rest state and imaginative hand movement.

III. Experimental Setup

The EEG data of 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 $5\,\mathrm{s}$ during each trial experiment and low-pass filtered with a cutoff 45 Hz. There were 100 trails acquired from each subject during each experiment day. At $t = 2\,\mathrm{s}$, a cross ("+") was displayed on the monitor of computer as a cue visual stimulus. The subjects were asked to imagine the hand grasping in synchronization with the cue and to not perform a specific mental task before displaying the cue. In the present study, the tasks to be discriminated are the imaginative hand movement and the idle state.

Eye blink artifact was suppressed by using independent component analysis. The artifactual independent components were visually identified and set to zero.

IV. Results

Original features are formed from 0.8-s interval of single-channel EEG data, in the time period 2.2-3.0 s, during each trial of experiment. The window starting 0.2 s after cue presentation is used for classification because it takes time for the subjects to start imagination. The mean absolute value (MAV), variance, zero crossing and number of extrema of each interval, 5 AR parameters, energy of 8 wavelet subbands, 1-Hz frequency components between 1 and 45 Hz constitute the full set of features. 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 network training, while the rest is kept aside for validation purposes. Training and validating procedure is repeated 10 times and the results are averaged.

Table I summarizes the results of classification accuracy of the EEG signals using full set of features consisting of 62 features. Using the full set of features for classification, the average success rate for five subjects and different experiment days, is 76%.

Table II summarizes the results of classification accuracy for different subjects using mutual information based feature selection (MIFS-U) proposed by Kwak, and Choi [11]. Among 62 features, 30 feature was selected. The average classification accuracy over all subjects is 78% which 2% better than that obtained by full feature set. An average classification rate of 80% is achieved by using MIFX with 30 extracted features. It is observed a classification accuracy rate as high as 94% is achieved in subject AE using MIFX. The results show that the MIFX performed better than full feature set by 4% in classification rate. The results of BCI performance using full feature set, MIFS-U, and MIFX for different subjects are shown in Fig. 1. It is observed that the MIFX algorithm provides better performance than MIFS-U and full set. Fig. 2 shows the classification accuracy rate for two experiment trials for different sizes of feature set obtained by MIFS-U and MIFX methods.

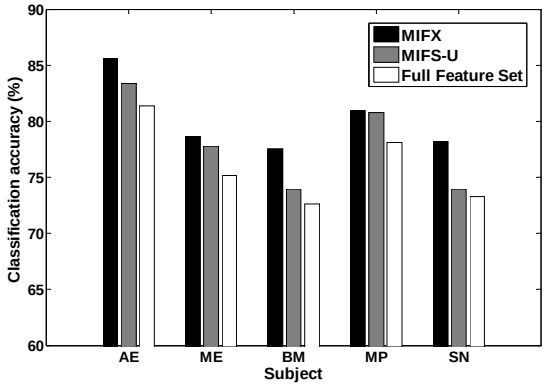

Fig. 1. Mean classification accuracy of EEG patterns for different subjects using full feature set and features sets obtained by MIFS-U and MIFX methods.

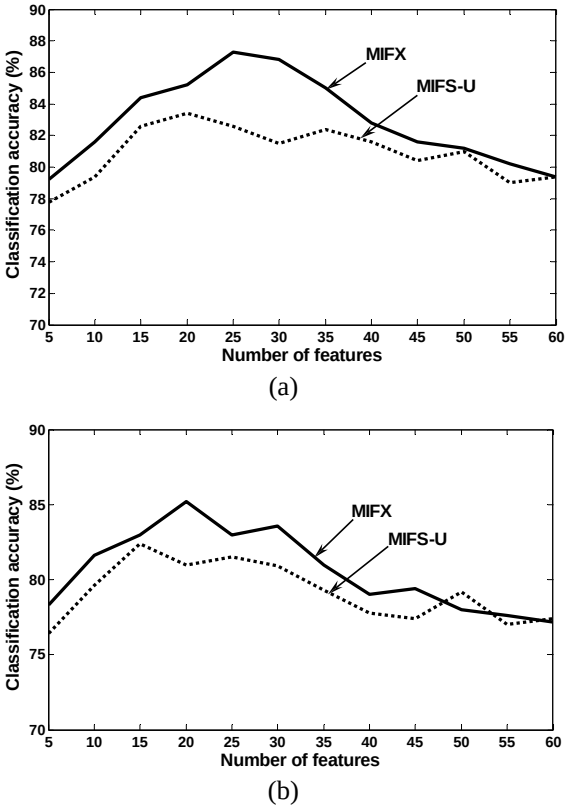

Fig. 2. Classification performance for two experiment trials for different sizes of feature set obtained by MIFS-U and MIFX methods.

It is observed that MIFX method provides better performance than MIFS-U for different sizes of feature set. The best performance is obtained using feature set with the size of 26 for one trial and of 20 for the second trial experiment.

V. Conclusion

Feature extraction plays an important role in classification systems. In this paper a novel method for feature extraction based on mutual information was proposed. 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. In contrast, feature selection is an approach to selecting a relevant subset from the original feature set [11], [12]. The proposed MIFX method iteratively creates new feature (*i.e.*, new generation) from a pair of features (*i.e.*, parent) based on two-dimensional MI estimates. The new generation is substituted for the parents into the original set. This process is repeated until a desired number of features are created. The method was applied to the classification of EEG signals. The tasks to be discriminated are the imaginative hand movement and the resting state. The results demonstrated that the method can improve the performance of BCI. The analysis of variance (ANOVA) shows that the mean classification accuracies achieved by using MIFX and full feature set are significantly different ($p < 0.0005$).

References

[1] H. Li, T. Jiang, and K. Zhang, "Efficient and Robust Feature Extraction by Maximum Margin Criterion," *IEEE Trans. Neural Networks*, vol. 17, no. 1, pp. 157–1165, Jan. 2006.

[2] R.O. Duda, P.E. Hart, and D. Stork, *Pattern Classification*. Wiley, 2000.

[3] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, 2001.

[4] N. Kwak and C.-H. Choi, "Feature extraction based on ICA for binary classification problems," *IEEE Trans. Knowledge and Data Engineering*, vol. 15, no. 6, Nov/Dec 2003.

[5] A. Erfanian and A. Erfani, "EEG-based brain-computer interface for hand grasp control: feature extraction by using ICA," in *Proc.* 9th *Annual Conf. Int. Functional Electrical Stimulation Society*, UK, 2004.

[6] A. Erfanian and A. Erfani, "ICA-based classification scheme for EEG-based brain-computer interface: the role of mental practice and concentration skills," in *Proc.* 26th *Annual Conf. Int. Conf. IEEE/EMBS*, San Francisco, USA, 2004.

[7] K. Torkkola, "Feature extraction by non-parametric mutual information maximization," *Journal of Machine Learning Research 3*, pp. 1415–1438, 2003.

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

[9] T. Trappenberg, J. Ouyang and A. Back "Input variable selection: mutual information and linear mixing measures," *IEEE Trans. Knowledge and Data Engineering*, vol. 15, no. 1, Jan 2006.

[10] G. Darbellay and I. Vajda, "Estimation of the information by an adaptive partitioning of the observation space," *IEEE Trans. Information Theory*," vol. 45, no. 4, pp. 1315-1321, May 1999.

[11] N. Kwak and C.-H. Choi, "Input feature selection for classification Problems," *IEEE Trans. Neural Networks*, vol. 13, no. 1, pp. 143–159, Jan. 2002.

[12] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Trans. Neural Networks*, vol. 5, no. 4, pp. 537-550, July 1994.

TABLE I
Classification Accuracy Rate of EEG Signals During Hand Movement Imagination using Full Feature Set.

Subject	Day1			Day2			Day3			Day4			Day5			mean
	min	mean	max	min	mean	max	min	mean	max	min	mean	max	min	mean	max	
AE	75	79.4	82	80	81.4	84	80	83	**88**	78	81.6	86	-	-	-	81.35
ME	71	75.6	82	68	72.9	77	75	78.9	86	69	73.2	78	-	-	-	75.15
BM	66	71.3	77	64	67.5	72	68	74	82	73	77.3	85	-	-	-	72.59
MP	69	75.9	81	74	76.6	79	75	78.8	84	76	79	84	76	80.6	83	78.1
SN	72	74.7	80	71	73.9	78	74	79.7	82	62	67.1	72	66	71.1	76	73.3

TABLE II
Classification Accuracy Rate of EEG Signals During Hand Movement Imagination using Feature Set Obtained by MIFS-U Method.

Subject	Day1			Day2			Day3			Day4			Day5			mean
	min	mean	max	min	mean	max	Min	mean	max	min	mean	max	min	mean	max	
AE	77	81.5	84	80	82.8	85	83	85.7	**89**	81	83.6	87	-	-	-	83.4
ME	73	75.7	78	69	74.3	79	77	80.1	83	75	80.9	86	-	-	-	77.75
BM	71	74.9	81	66	67.9	72	67	72.2	77	77	80.9	84	-	-	-	73.97
MP	71	74.6	78	79	81.8	85	79	81.3	83	78	81.1	84	81	85	**89**	80.76
SN	69	74.6	79	73	76.6	80	71	75	79	62	68.6	72	65	74.6	80	73.92

TABLE III
Classification Accuracy Rate of EEG Signals During Hand Movement Imagination using Feature Set Obtained by MIFX Method.

Subject	Day1			Day2			Day3			Day4			Day5			mean
	min	mean	max	min	mean	max	min	mean	max	min	mean	max	min	mean	max	
AE	81	86.8	**94**	80	85	90	85	87	90	81	83.9	87	-	-	-	85.67
ME	73	80.9	85	73	76	80	77	80.1	83	73	77.6	82	-	-	-	78.65
BM	73	75.6	81	68	72.2	77	74	78.8	82	81	83.6	86	-	-	-	77.55
MP	75	79.1	86	75	80	84	75	79	82	79	81.7	86	82	85.2	89	81
SN	74	81	87	76	78.8	82	77	82.1	86	70	73.4	76	71	75.9	78	78.24