FEATURE SELECTION BASED ON FISHER RATIO AND MUTUAL INFORMATION ANALYSES FOR ROBUST BRAIN COMPUTER INTERFACE

Tran Huy Dat, Cuntai Guan

Institute for Infocomm Research
21 Heng Mui Keng Terrace, Singapore 119613

ABSTRACT

This paper proposes a novel feature selection method based on two-stage analysis of Fisher Ratio and Mutual Information for robust Brain Computer Interface. This method decomposes multichannel brain signals into subbands. The spatial filtering and feature extraction is then processed in each subband. The two-stage analysis of Fisher Ratio and Mutual Information is carried out in the feature domain to reject the noisy feature indexes and select the most informative combination from the remaining. In the approach, we develop two practical solutions, avoiding the difficulties of using high dimensional Mutual Information in the application, that are the feature indexes clustering using cross Mutual Information and the latter estimation based on conditional empirical PDF. We test the proposed feature selection method on two BCI data sets and the results are at least comparable to the best results in the literature. The main advantage of proposed method is that the method is free from any time-consuming parameter tweaking and therefore suitable for the BCI system design.

Index Terms— Brain Computer Interface, Filterbank, Feature Selection, Fisher Ratio, Mutual Information

1. INTRODUCTION

Brain Computer Interface (BCI) is a fast-growing emergent technology, in which researchers aim to build a direct channel between the human brain and the computer [1]. This technology provides a new alternative of augmentative communication and control for the physically disabled. The typical BCI includes temporal-temporal filtering, spatial filter, feature extraction and classifier.

One important problem with BCIs is that its performance is very sensitive to spatial and temporal filtering but the optimal filter is strongly subject-dependent and therefore the conventional method to find the filter parameters through an exhaustive search based on cross-validations is time-consuming and inconvenience.

Recently, some automatic learning methods of the temporal filter have been proposed in the literature for some particular cases of spatial filters: the Common Spatial Pattern (CSP) [2] and the Laplacian filter [3]. These methods optimize the FIR filter using some objective function and could achieve in some cases good results close to the best result from exhaustive search. However, some limitations are remained as follows: 1) gradient-based algorithm is slow and might get stuck in local optimal; 2) final result is sensitive to the initial parameters.

In this paper we propose an alternative method which simultaneously optimize the temporal and the spatial filters. Furthermore, the method is applicable for any type of spatial filter, whose most popular in BCI are the Independent Component Analysis (ICA) and the Common Spatial Pattern (CSP). The main points of our system

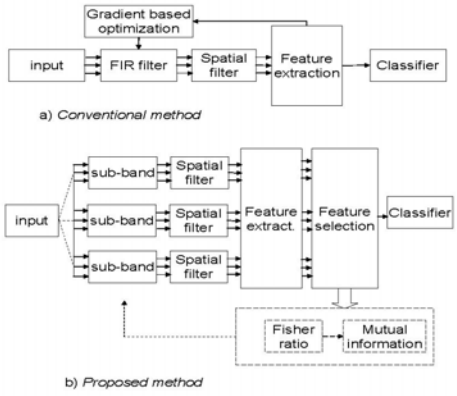

Fig. 1. Processing diagram of a) Conventional method, and b) Proposed method

can be summarised as follows: 1) a filterbank system is adopted instead of a single temporal filter as used in all existing systems; 2) the spatial filter (i.e. ICA or CSP) and feature extraction is applied in each subband after temporal filtering; 3) the feature selection, i.e. the channel and subband selection, is carried out in feature domain using a two-stage procedure combining Fisher Ratio and Mutual Information analyses on the time-series obtained from a training database for each feature index.

The feature selection, an important area of machine learning, has also been studied in BCI but the applications are limited for the channel selection only [4]-[5]. The Recursive Feature Elimination, originally proposed for gen classification, was employed for BCI [4]. However, this method includes the classifier inside the procedure and this makes the system design inconvenience. Recently, in [5], the authors proposed a method for channel selection based on searching the maximum mutual information of all possible channels' combination. The limitation of this method is that an additional ICA need to be applied for each channels' combination in order to estimate Mutual Information and this makes the processing very time-consuming. Moreover, both methods in [4] and [5] did not discuss the frequency selection as fixed band pass filters were applied.

In our proposed method (Fig.1.b), the feature selection, by mean of subband and channel selection, is independent from the classifier and the type of feature to be applied. The combination of Fisher Ratio and Mutual Information is the principal merit of the proposed method. From a theoretical point of view, the Fisher Ratio analysis should be able to select the most discriminate (i.e. less noisy) feature

components but can not to provide the "best" combination between selected components because there might be cross correlations. On another hand, the Mutual Information without Fisher Ratio analysis would be able to provide the maximum independence between feature components but this might mistakenly select the noisy components which even degrade the system. In this work we show that the combination of Fisher Ratio and Mutual Information greatly improve the performance of the BCI systems.

In the approach, we developed two flexible and practical solutions for the application of Mutual Information. First, to avoid the typical difficulty in estimating high dimensional mutual information, we use two-dimensional Mutual Information as a distance measurement to cluster the pre-selected components into groups. Their best subset is then chosen picking up the component with the best Fisher Ratio score in each group.

In the approach, we develop two practical solutions which avoid the difficulties in the application of high dimensional Mutual Information: the feature indexes clustering using cross Mutual Information and the former estimation based on conditional empirical PDF. The proposed feature selection is applied for both Independent Component Analysis (ICA) and Common Spatial Pattern (CSP) spatial filters and the evaluation is tested on two datasets of ECoG and EEG signals. In next section we describe more details of the proposed method.

2. FEATURE SELECTION BASED ON FISHER RATIO AND MUTUAL INFORMATION

In this section we discuss following issues of the proposed method: 1)the pre-selection of feature components based on their Fisher Ratio (FR) scores and the optimization of the number of filters; 2)the final selection by Mutual Information.

2.1. Pre-selection by Fisher Ratio Analysis

2.1.1. Fisher Ratio Analysis

Suppose that two investigating classes on a feature component domain (i.e. a subband-channel index) have mean vectors μ_1, μ_2 and covariances $\mathbf{\Sigma}_1, \mathbf{\Sigma}_2$, respectively. The Fisher Ratio is defined as the ratio of the variance of the between classes to the variance of within classes noted by

$$d = \frac{\sigma^2_{between}}{\sigma^2_{within}} = \frac{(\mathbf{w}\cdot\mu_1 - \mathbf{w}\cdot\mu_2)^2}{(\mathbf{w}^T\mathbf{\Sigma}_1\mathbf{w} + \mathbf{w}^T\mathbf{\Sigma}_2\mathbf{w})} = \frac{\left[\mathbf{w}\cdot(\mu_1-\mu_2)^2\right]}{\mathbf{w}^T(\mathbf{\Sigma}_1+\mathbf{\Sigma}_2)\mathbf{w}}. \tag{1}$$

The maximum of class separation (discriminative level) is obtained when

$$\mathbf{w} = (\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2)^{-1}(\mu_1 - \mu_2) \tag{2}$$

As the Fisher Ratio can be considered as a "Signal-to-Noise" Ratio measurement, this step also means to reject the noisy components in the feature domain. In the processing, the Fisher Ratios are estimated from time series obtained from a training database on each feature index. The index components with top scores are then selected.

2.1.2. Optimize the number of filters

A arising question is how to set the optimal number of filters? In this work we propose an iterative method to define this number by comparing the maximum Fisher Ratio value from each filterbank set. The reason of the choice of maximum measure is come from the fact what we have learned from the experiences that is the best subband makes the largest contribution in the final performance [6]. To provide a common framework in the processing, we start with a fixed number of filters, says 4 subbands, and this number will be iteratively increased until the maximum of Fisher Ratios stops increasing. Now given the pre-selected feature components we discuss how to select their best subset for the classification task.

2.2. Feature selection by mutual information maximization

The Mutual Information maximization is a natural idea for subset selection since this can provide a maximum information combination of pre-selected components. However, a big problem is that the estimation of high-dimensional mutual information requires very large number of observations to be accurate but this is often not provided. To solve this problem we develop a flexible solution using only two-dimensional cross Mutual Information. This measurement is used as a distance to cluster the feature components into groups and then select from each group the best component by mean of the Fisher Ratio.

2.2.1. Cluster pre-selected components for the selection

The Mutual Information based clustering is an iterative procedure likes vector quantization and therefore might be sensitive to the initial setting. However, taking into account the fact that the largest contribution is expected to come from the best pre-selected component, the initialization is set as follows.

1. Fixed the feature component with the best Fisher Ratios; calculate the cross Mutual Informations from each component to this; sort the estimated sequence;
2. Set the initial cluster centers uniformly from the sorted sequence;
3. Classify the components according to the lowest cross mutual information to the centers.
4. Recompute the center of each group;
5. Repeat (3) and (4) until the centers do not change;
6. Select in each group the component with highest FR score.

2.2.2. Two-dimension mutual information estimation

Now we pay our attention to the estimation of cross Mutual Information. Conventional method estimates the two-dimensional Mutual Information through the marginal and joint histograms

$$I(X,Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y)\log_2\frac{p(x,y)}{p(x)\,p(y)}, \tag{3}$$

where x, y denote the observation of random variables X and Y. In our case, X and Y are a pair of feature components. The typical problems of the estimation (3) are the complexity and possible presence of null bins in the conventional joint histogram.

In this work, we develop an estimation method by using the empirical conditional distributions We first rewrite (3) as

$$I(X,Y) = \sum_y p(y)\sum_x p(x|y)\log_2\frac{p(x|y)}{p(x)}. \tag{4}$$

equation (4) can be simplified by clustering the y into clusters and estimate the conditional densities in each cluster

$$\begin{aligned} I(X,Y) &= \sum_y\sum_i p(y|i)\sum_x p(x|i)\log_2\tfrac{p(x|i)}{p(x)} \\ &= \sum_y\sum_i p(y|i)\sum_x\sum_{k_i} p(x|i,k_i)\log_2\tfrac{p(x|i,k_i)}{p(x)}, \end{aligned} \tag{5}$$

where i is the cluster index and k_i is the number of observations in each cluster. For clustering y we apply a fast method based on order statistics. We briefly describe this idea as follows. Given an observed sequence $Y\{y_1, y_2, ..., y_N\}$ and a number M, a set of M order statistics are defined as

$$c_i = \{y'_{q_i}\} \tag{6}$$

where,

$$q_i = \left[(i-1)\frac{N-1}{M-1}\right] + 1, \tag{7}$$

and

$$y'_1 \leq y'_2 \leq .. \leq y'_N \tag{8}$$

is the sorted sequence of Y, $i = 1, 2, ..., M$. The clustering of sequence Y into group number i is given by matching the sequence to an inequality as

$$c_{q_{i-1}} \leq y < c_{q_i}. \tag{9}$$

The conditional density for each group $p(y|i)$ is estimated by a "piecewise" density function using the order statistics, calculated from the clustered samples in each group [13]. For example, the "piecewise" pdf of sequence Y with the order statistics in (5) is noted by

$$p(z) = \begin{cases} 0 & z < z'_{q_1} \\ \frac{1}{(z_{q_{i+1}} - z_{q_i})} & z'_{q_i} \leq z < z'_{q_{i+1}} \\ 0 & z > z'_{q_M} \end{cases} \tag{10}$$

The selection of parameter M is determined by a bias variance trade-off and typically, $M = \sqrt{N}$.

2.3. Algorithm of two-step sub-band selection with ICA and CSP spatial filtering

In this paragraph, we describe in more details the final algorithms for each case of using ICA or CSP.

2.3.1. Feature selection for ICA

The ICA is an unsupervised method which unmix the source components by maximizing the independence between the output components. A question is which ICA components are useful if there is no prior knowledge on the sources. To overcome this problem, conventionally, a follow-up supervised component selection should be applied. In our algorithm, the ICA is processed in each subband and the feature selection combines both the channel (i.e. ICA component) and the subband selection. The final algorithm is as follows.

1. Initially set a filterbank set of N-bands (N=4)
2. Apply filtering on training database.
3. Perform ICA and on each sub-band. Extract the features.
4. Calculate the Fisher Ratio for each feature component (i.e. subband-channel index).
5. Decrease the subband bandwidth twice and repeat items 2-4.
6. Repeat item 5 until the best Fisher Ratio does not increase.
7. Pre-select components yielding more than 70% of the best score.
8. Apply the subset selection based on Mutual Information described in section 2.2.

Note that here we adopt the fast-ICA algorithm proposed in [8].

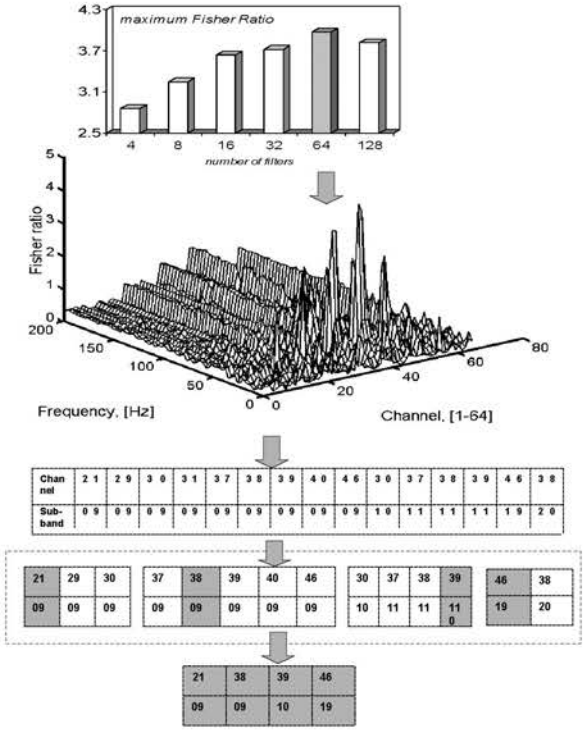

Fig. 2. Example of feature selection on ECoG database

2.3.2. Feature selection for CSP

This supervised method [2] projects the multi-dimensional signal into a direction that maximize the discrimination of class variances. As the result, the common patterns (i.e. background noises) in two class should be formed in the middle rows and can be removed. In our method, the CSP is processed in each subband. The feature selection algorithm here is almost the same as for ICA case but since the dimension reduction is done inside the CSP, the initial number of feature components is much more smaller than in ICA case.

3. EXPERIMENT AND DISCUSSIONS

In this section we evaluate the proposed sub-band selection method and compare it to the conventional ones.

3.1. Databases

The evaluation is for the offline motor imagination task and was carried out on two datasets of the BCI III competition [7]: ECoG dataset-1 and EEG dataset-4a. The ECoG database contains recordings from two different sessions about one week apart, for cued motor imagery (left finger or tongue) from one subject. The electrical brain activity signals were recorded using a 64 channel ECoG platinum electrode grid which was placed on the contra lateral (right) motor cortex. Every trial consisted of either an imagined tongue or

Table 1. Classification accuracies [%] of the proposed and reference methods on EEG and ECoG datasets

Data	ICA-crv	ICA-subb-Fis.	ICA-subb	CSP-crv	CSS-SP	CSP-subb-Fis.	CSP-subb	BCI-III
EEG	78.64	80.72	82.21	86.21	88.96	86.84	94.08	94.17
ECoG	87.00	88.00	93.00	82.00	81.00	84.00	85.00	91.00

an imagined finger movement. The labeled training database contains 278 training trials and 100 test trials. The EEG 4a dataset contains EEG signals (118 channels, sampled at 100 Hz) for five healthy subjects. Two classes of motor imagination: right hand or right foot movements. Thus there are 2x140 trials in total for each subject.

3.2. Method to evaluate

The following methods are implemented and evaluated in this paper:

1. Reference methods: The ICA and CSP using single bandpass filter with cross-validation (ICA-crv, CSP-crv). The CSSSP method [3]. The subband ICA and CSP with only 1-step Fisher Ratio Analysis (ICA-subb-Fis., CSP-subb-Fis.)
2. Proposed methods: The proposed 2-step selection combining Fisher Ratio and Mutual Information (ICA-subb, CSP-subb).

We also compare to the best results from BCI III competition (BCI-III). For the ECG database the conventional log-energy feature is adopted and for the ECoG database, we use the inverse cumulative distribution feature developed in our previous work [6]. The Linear Discriminant Analysis Classifier is applied for the classification.

3.3. Feature selection

Fig.2 illustrates an example of the proposed sub-band selection: the sub-band ICA with proposed feature selection for the ECoG database. The uppermost diagram compares the maximum score of Fisher Ratio over different number of filters. The set of 64 filters covering the frequency range from 0-200Hz is found to be optimal. The second graphic plot the Fisher Ratio analysis for the 64-filter system. We can see that only few components concentrated near to the 21rd, 28th and 46th channel indexes are useful. The best subbands are band number 9 and 11 which are 25.39Hz-28.57Hz and 31.74Hz-34.92Hz. However, some other bands, which can also be useful, are the 19th and 20th bands. The next tables shows the step-by-step selection. Four subband-channel components were finally selected for the classification.

3.4. Evaluation of classification

The overall classification accuracies of evaluated methods on ECoG and EEG datasets are shown in Tab.1. The proposed feature selection method performs best for both datasets but with different spatial filters. The proposed ICAs-subb is the best for ECoG database. This method even outperformed the best score in the BCI-III competition. For the EEG database, the proposed CSP-subb get almost the same result as the best one in the Competition. Note that the best methods of BCI-III used a manual tuning of temporal filter plus a heuristic fusion of some classifiers which was also manually applied for each selective subject [7]. As these methods could get a high accuracy results, no unique framework for the system designing is provided. The main point of our method is that this can utilize an unique, and free from any parameter tweaking framework while achieves at least the same good results as the best available ones.

The question of which spatial filter should be applied for ECoG and EEG is an interesting question. Although this is not the scope of this paper, from the experiment results, it seems that the CSP is very effective for EEG but not good for ECoG. The reason might be the fact that the ECoG, observed with much more higher SNR, is more closed to the real "sources" and therefore has less "common spatial patterns" than the EEG one.

The combination of Fisher Ratio and Mutual Information significantly improved the performance. The absolute accuracy of the 2-step selection method overcome the 1-step of Fisher Ratio Analysis up to 6 %. The proposed method also greatly outperformed the methods with single bandpass filter where the cross validation is applied. This confirmed the fact that the use of information in well selective bands should be more effective than that of a single one. The gradient based CSSSP method seems to not always perform well as this is the third best for the EEG dataset but even worst for the ECoG one. The ICA-subb seems not optimal for EEG and this might be caused by the over separation problem of the ICA.

4. CONCLUSIONS

The main points of this paper are summarised as follows: 1) the combination of Fisher Ratio and Mutual Information analyses is shown to be very effective for the feature selection in BCI; 2) the proposed feature selection method achieves the same good results as the best available in literature but is free from time-consuming parameter tweaking. The proposed method utilizes an unique and robust framework for the design of BCI systems.

5. REFERENCES

[1] J. R. Wolpaw, Brain-Computer Interfaces (BCIs) for Communication and Control: Current Status, *in 2nd Int. BCI Workshop & Training*, Course, 2004, pp. 29-32.

[2] Dornhege, G., Blankertz, B., Krauledat, M., Losch, F., Curio, G., and Muller, K.-R. Optimizing spatiotemporal filters for improving BCI, *In Advances in Neural Info. Proc. Sys.*, 2006.

[3] Le Song, and Julien, E., Classifying EEG for brain-computer interfaces: learning optimal filters for dynamical system features, *In Proceeding of the 23rd ICML*, 2006

[4] M. Schrder, T.N. Lal, T. Hinterberger, et al., Robust EEG Channel Selection across Subjects for Brain-Computer Interfaces, *EURASIP Journal on Applied Signal Processing* , vol. 2005, no. 19, pp. 3103-3112, 2005

[5] T. Lan, D. Erdogmus, A. Adami, M. Pavel, S. Mathan, Salient EEG Channel Selection in Brain Computer. Interfaces by Mutual Information Maximization, *In Proceedings 27th IEEE EMBC, September 1-4*, 2005, pp. 7064-7067.

[6] T. H. Dat, L. Shue, C. Guan , Electrocorticographic signal classification based on time-frequency decomposition and nonparametric statistical modeling, In Proceedings 28th IEEE EMBC, 30 Aug - 3 Sept, 2006, pp. 2292-2295.

[7] BCI competition III:
http://ida.first.fraunhofer.de/projects/BCI/competition III/.

[8] The FastICA package for MATLAB
http://www.cis.hut.fi/projects/ica/fastica/