EEG Channel Selection Using Decision Tree in Brain-Computer Interface

Mahnaz Arvaneh[*†] , Cuntai Guan[†], Kai Keng Ang[†] and Hiok Chai Quek [*]
[*] School of Computer Engineering, Nanyang Technological University, Singapore
[†]Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore
E-mail: stuma@i2r.a-star.edu.sg

***Abstract*—Selecting appropriate channels in Brain-Computer Interface (BCI) applications helps to improve the usability and the performance of the BCI as some channels are contaminated by noise or contain irrelevant information. This paper proposes a method of using decision trees to select appropriate channels in EEG-based BCI applications. The proposed method selects the best subset of appropriate channels by considering the correlation information between them using Decision Tree. The performance of the proposed method is compared with several other methods of channel selection, such as Fisher Criterion, Mutual Information, Support Vector Machine and Common Spatial Pattern coefficients. The performances of these methods are evaluated in terms of using publicly available BCI Competition IV dataset IIa. Experimental results show that the proposed method outperforms the existing channel selection methods specifically in the case where the number of selected channels is relatively small.**

I. INTRODUCTION

A brain-computer interface (BCI) measures, analyzes and decodes brain signals directly to provide a non-muscular means of controlling a device. Thus BCIs enable users with severe motor disabilities to use their brain signals for communication and control [1], [2]. There are mainly two types of BCIs, invasive and noninvasive. Electroencephalogram (EEG) is commonly used in noninvasive BCIs because it is the least expensive compared with other methods of brain signal acquisition equipments. However, EEG signal processing is a challenging problem due to the poor resolution of EEG and its multi-channel nature in the acquisition of brain signals [3]. The use of too few channels may result in insufficient information whereas too many channels may include noisy and redundant channels that degrade BCI performance.

One method to improve the performance of EEG-based BCI is to use appropriate channels on the scalp. This is because if noisy and redundant channels are excluded, computational complexity is decreased while accuracy of the BCI may be increased [4]. Moreover, the use of a large number of channels is not practical because it involves a longer EEG setup time. Since appropriate channels may differ from subject to subject, a method of finding subject-specific optimal number of appropriate channels plays an important role in the performance of BCI applications.

The problem of EEG channel selection can be considered as a feature selection problem. Channel selection methods in the literature are mainly characterized as wrapper or filter approaches. In wrapper approaches, feature selection is coupled with classification algorithm such as the Support Vector Machine (SVM) classifier [4]. In filter approaches, feature selection is independent of induction algorithms. One example is to select features based on certain criteria such as the Mutual Information (MI) between channels and class labels [5]. The performances of wrapper-based methods depend on the accuracy of the applied classifier and properties of the features coming from channels. Although some methods have been proposed to avoid retraining classifiers [6], wrapper-based feature selection methods generally involved intensive computations. In contrast, filter-based feature selection methods are computationally less intensive than wrapper approaches, but may not select an optimal subset of features [6].

Another EEG-specific approach uses the Common Spatial Pattern (CSP) coefficients [7]. The CSP algorithm is shown to be effective in discriminating two classes of EEG measurements in BCI applications [8]. However, CSP is sensitive to outliers and EEG signals are generally noisy from various artifacts. Thus channel selection using CSP coefficients may not select an optimal subset of appropriate channels.

This paper proposes a method of using decision trees [9] to select appropriate subset of channels for EEG-based BCI applications. In the proposed algorithm, initially a multi band signal decomposition filter is presented to reduce noise by identifying the subject-specific frequency range, and then the most discriminative subset of features is selected by the defined decision tree classifier. Finally the selected features are ranked according to a tree pruning method. Since the decision tree selects a feature according to the results of previous chosen features, selected features would be more relevant and less correlated to each other. The remainder of this paper is organized as follows: Section II reviews the existing EEG channel selection methods based on Fisher Criteria (FC), MI, SVM and CSP coefficients. Section III explains the proposed method. Section IV describes the experiment performed on the publicly available BCI Competition IV dataset IIa. Section V presents the experimental results by comparing the proposed method with existing methods. Finally section VI concludes this paper.

Proceedings of the Second APSIPA Annual Summit and Conference, pages 225–230,
Biopolis, Singapore, 14-17 December 2010.

10-0102250230©2010 APSIPA. All rights reserved.

II. REVIEW ON CHANNEL SELECTION METHODS

The goal of channel selection is to remove irrelevant or correlated channels in order to improve the performance of the BCI system. The following reviews existing channel selection methods used in BCI applications in the literature:

A. Fisher Criteria (FC)

The FC determines how strongly a feature is correlated with the labels [4] whereby the score R_j of feature j is defined as

$$R_j(X) = \frac{(\mu_j(X^+) - \mu_j(X^-))^2}{V_j(X^+) + V_j(X^-)}, \quad (1)$$

where X^+ and X^- denote the set of trials in two different classes; μ_j and V_j respectively denote the mean and variance of feature j. The rank of a channel is simply set to the mean score of the corresponding features.

B. Mutual Information (MI)

In this method, the features that have maximum MI with the class labels are ranked as the best features. The MI between input features X and the class $Y=\{1,...,N_y\}$ is defined as

$$I(X;Y) = H(Y) - H(Y|X), \quad (2)$$

where N_y is the number of classes, and H denotes the entropy function [5]. Entropy is a measure of uncertainty associated with a random variable. Given a data $T=\{T_1,T_2,...,T_d\}$, the entropy of the random variable T is defined as

$$H(T) = -\sum p(T)\log_2 p(T), \quad (3)$$

where $p(\cdot)$ is the probability function. The conditional entropy of two random variables X and Y is defined as

$$H(Y|X) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2 p(y|x). \quad (4)$$

In this study, the Parzen Window [10] is used to estimate $p(y|x)$. The Mutual Information based channel selection algorithm is described as follows:

- Step 1: Initialize a set of d features $F = \{f_1, f_2, ..., f_d\}$
- Step 2: Compute the MI of features with the output class $I(f_i;Y) \quad \forall i = 1...d, f_i \in F.$
- Step 3: Select the best features that maximize $I(f_i;Y)$

C. Support Vector Machine (SVM)

The SVM is a classification technique [4] which performs efficiently in a number of real-world problems. The SVM separates the training data $X \subseteq R^d$ with two classes $y=\{-1,1\}$ by finding a hyperplane with a weight vector $w \in R^d$ and an offset $b \in R$, as shown in (5). A good separation is achieved by a hyperplane that has the largest distance to the nearest training data points of any class. This distance is called the functional margin.

$$H: R^d \to \{-1,1\}, \quad x \to sign(w.x + b) \quad (5)$$

In SVM-based channel selection, the channels are selected using a Recursive Feature Elimination (RFE) method. The RFE method was proposed by Guyon et al. [6] based on the concept of margin maximization. The RFE algorithm begins with the subset comprising all the features and eliminates one feature at a time from the subset. In each iteration, the learning machine f, (in our case SVM classifier), is trained on the current subset of features by optimizing a cost function J to maximize the marginal function. For each remaining feature X_i, the change in J resulting from the removal of X_i is estimated without retraining the f. Thereafter, the feature $X_{v(K)}$ that results in improving or least degrading J is removed. This algorithm is iterated till only the specified number of features remains.

Guyon et al. have presented in [6] that under some conditions, the removal of one feature will induce a change in the generalization error proportional to gradient of f with respect to i^{th} feature at point x_k given by

$$\sum_{k=1}^{m} \left(\frac{\partial f}{\partial x^i}\right)^2 (\alpha, x_k). \quad (6)$$

In SVM classifier, $\frac{\partial f}{\partial x^i} \propto w_i$. So the SVM Recursive Feature Elimination method is described as follows:

- Step 1: Get w^* as the solution of SVM on the data set restricted to features.
- Step 2: Select top features as ranked by $(w_i^*)^2$. Since $(w_i^*)^2$ is proportional to the i^{th} feature, the best features are those that have greater $(w_i^*)^2$.

D. Common Spatial Pattern (CSP)

The CSP algorithm [8] is an effective technique to discriminate between two classes of EEG data. The CSP algorithm projects the raw signal to a spatially filtered signal Z as given in (7) that maximizes the variance of one class while minimizes the variance of the other class.

$$Z = WX \quad (7)$$

Let $X \in R^{N\times T}$ denotes a matrix that represents the EEG of a single-trial; N and T denotes the number of features and the number of measurement samples per feature respectively. The rows of projection matrix W are the stationary spatial filters and the columns of W^{-1} are the common spatial patterns. The CSP algorithm performs simultaneous diagonalization of the covariance matrices of both classes. For each centered and scaled X, the estimated covariance matrix in class C, $\sum^{(C)} \in R^{T\times T}$, is given by

$$\sum^{(C)} = \frac{1}{|I_C|}\sum_{i\in I_c} X_i X_i^T \qquad (C \in \{+,-\}), \quad (8)$$

where $|I_C|$ denotes the number of trials belonging to class C. The CSP projection matrix W is computed by simultaneous diagonalization of the two covariance matrices given by

$$W^T \sum^{(+)} W = \Lambda^{(+)}, \quad W^T \sum^{(-)} W = \Lambda^{(-)}, \tag{9}$$

where $\Lambda^{(C)}$ is the diagonal eigenvalues and the scaling of W is commonly determined such that $\Lambda^{(+)} + \Lambda^{(-)} = I$ [8].

The proposed CSP channel selection by Wang et al. [7] is defined as follows: Optimal channels for every motor imagery task are determined through the maximums of the absolute value of the concerned spatial pattern. Let SP_{Ri} and SP_{Hi} denote i^{th} optimal channels of spatial pattern for right and left hand motor imagery respectively, therefore equation (10) is calculated to obtain overall ranking, where i varies from 1 to the total number of channels. Finally since every channel has been iterated twice in CH, the lower rank is discarded. As shown in equation (10), in this method channels are pair-wisely selected from both left and right motor imagery areas.

$$CH_{2i-1} = Find(Max(SP_{Ri}, SP_{Li})) \quad CH_{2i} = Find(Min(SP_{Ri}, SP_{Li})) \tag{10}$$

III. METHODOLOGY

The general structure of the proposed EEG channel selection method is shown in Fig.1. In this method, a multiband signal decomposition filter is applied to all channels of the multichannel EEG. Subsequently, the EEG signals are subject-specifically filtered into the most relevant frequency range. Finally the relevant channels are selected using the proposed decision tree based algorithm.

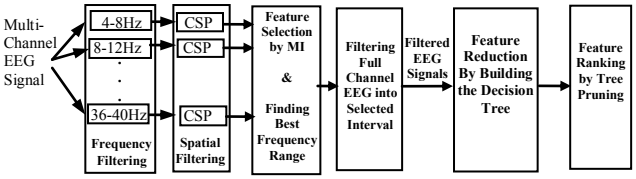

Fig.1. Proposed EEG channel selection method

A. Filtering by Multi Band Signal Decomposition

Fig. 1 shows that the subject-specific filter used in our work comprises four progressive steps described as follows:

The first step employs a Chebyshev filter bank to perform bandpass filtering of EEG in multiple frequency bands. The second step performs spatial filtering on each of these bands using the CSP algorithm. Thus, each pair of bandpass and spatial filters yields CSP features which are corresponding to the frequency range of the bandpass filter. The third step employs a Mutual Information-based feature selection algorithm to select the best discriminative CSP features. Subsequently, the best discriminative CSP features indicate the most relevant sub-bands. The fourth step employs an Elliptic bandpass filter to filter the original unfiltered EEG into the smallest frequency range including all the most relevant sub-bands.

B. Decision Tree-based Channel Selection

Decision Trees are classifiers that provide interpretable solutions. Since training the induction algorithm and selecting the features are performed simultaneously, the decision tree (DT)-based feature selection method is characterized as an embedded approach [6]. The DT comprises a root, internal (non-terminal) decision nodes and a set of terminal nodes or leaves, each representing a class.

The decision tree based feature selection method consists of two phases:

1- Building the tree for feature reduction: A training data set including all the features is used to build the tree. Features that do not appear in the tree are discarded.
2- Tree pruning for feature ranking: The remaining features are ranked backward by pruning the tree.

B.1 Building the tree

In this work, Classification And Regression Tree (CART) [11], one of the widely used DT algorithms, is used. CART is a technique that uses binary tree structure (with only two branches at each internal node). An example of CART decision tree is illustrated in Fig 2.

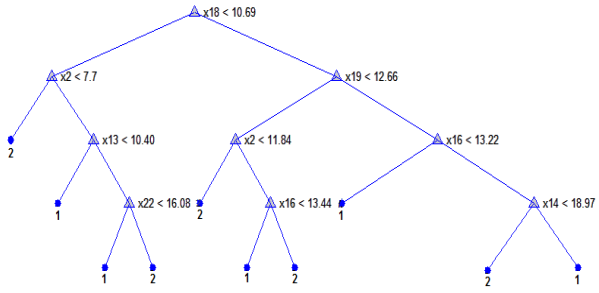

Fig.2. Example of CART decision tree

The process of tree building starts at the root (first internal node) with the entire training dataset being split into two subsets. Similarly, the partitioning of every subset into two 'subsubsets' is continued at each internal node based on a predefined criterion. According to the predefined criterion, a test is conducted at each step to find out the most suitable feature that gives the best separation of the training samples. This work uses the "gini index" criterion [11] to build the tree. The gini index provides a measure of the impurity degree in a dataset. For data set S, the gini index is computed using

$$gini(S) = 1 - \sum_{j=1}^{m} p_j , \tag{11}$$

where p_j and m are the probability of class j and the total number of classes respectively. The minimum value of the gini index occurs when the set consists of only one class; and the maximum value occurs when all the classes in the set have equal probability.

The goodness of a split point is specified by the gini value. If the decision splits the database S, into sets S1 and S2, the gini value of the divided data is

$$gini_{split}(S) = \frac{n_1}{n} gini(S_1) + \frac{n_2}{n} gini(S_2), \quad (12)$$

where n_i is the number of instances in S_i. The gini value is evaluated for every possible split of an attribute. This implies that not only a certain feature is chosen but also a split point that is used in a node. The procedure is repeated for each feature with its own split points and finally the feature with the lowest gini value is selected for the next split.

The tree building continues until all the remaining instances belong to a same class; or there is no new splitting to improve the overall accuracy of the tree.

B.2 Pruning the tree

In this work, a pruning process based on the overall accuracy of the tree is applied to rank the features. While exploring over the internal nodes from the bottom to the top, the procedure checks the overall accuracy of the tree after replacing each internal node with a leaf labeled by the highest represented class. Therefore the nodes with small decreases in performance are known as less important nodes. Consequently features are ranked according to the importance of the corresponding nodes.

IV. EXPERIMENTS

A. Data description

The EEG data from publicly available BCI Competition IV datasets 2a [13] are used in this study, which comprises two classes: right and left hand motor imageries. The EEG was recorded from nine subjects using 22 electrodes per subject. During each experiment, the subject was given visual cues that indicated four motor imageries should be performed: left hand, right hand, foot and tongue. 140 single-trials of EEG from each class of right and left hand motor imageries with the time segment of 0.5 to 2.5 seconds after cue were applied in this study. The EEG data from foot and tongue motor imageries were not used.

B. Data processing

In this work, the raw EEG is filtered using the best subject-specific frequency range extracted by the multi band signal decomposition algorithm. Subsequently, the covariance of each channel is computed as the feature corresponding to the channel. These features are used in all the channel selection methods except CSP-based method, because in CSP channel selection method channels are directly selected from the spatial patterns (refer to section II.C).

The performances of the different channel selection methods are evaluated by calculating the accuracy of the classification using different number of optimal channels. For this purpose, the common spatial filters are employed to spatially filter the EEG. Then the variances of three first and three last rows of the filtered signals [12] are used as inputs of the SVM classifier. It is noted that the classification accuracies of different methods are evaluated by averaged 10×10-fold cross-validation.

V. EXPERIMENTAL RESULTS

For evaluation purpose, the classification accuracies after the channel reduction (the first phase of our method) were compared with the results obtained from: (1) all the channels, (2) three typically used channels for left and right motor imageries, (C3, C4, Cz).

The experimental results for 9 subjects are shown in Table 1. The first row presents the averaged 10×10-folds classification accuracies of full channel EEG. Averaged number of selected channels by the decision tree, and achieved accuracies after selecting those channels are indicated on the second and third rows respectively. Finally obtained accuracies by using only C3, C4 and Cz are presented in the last row.

TABLE 1
PERFORMANCE COMPARISON OF DECISION TREE BASED EEG CHANNEL REDUCTION (CH: CHANNELS, ACC: ACCURACY, #: NUMBER)

Subject	1	2	3	4	5	6	7	8	9	Mean±Std
Full Ch Acc (%)	87.3	56.8	93.1	63.6	87.6	62.6	77.1	94.2	93.8	79.5±14.9
#Selected Ch	9	11	8	9	11	6	10	7	5	8.4± 2.1
Remained Ch Acc (%)	79.6	53.0	89.8	64.4	80.6	64.4	69.8	92.2	89.5	75.9±13.7
C3, C4, Cz Acc (%)	73.3	52.0	86.6	63.9	61.2	64.2	55.0	85.9	87.3	69.9±13.8

As can be seen in Table 1, the proposed method decreased the number of electrodes on average to 38% (of 22 electrodes), with sustaining only a drop of 3.63% in accuracy. Interestingly enough, the accuracies of selected channels in subjects 4 and 6 are even more than the full channel accuracies. It can happen due to removing redundant and noisy channels which degraded the performance.

According to the results, the proposed method performs significantly (mean=6% and p=0.029) better than three typical channels (C3, C4 and Cz). Furthermore, compared with the full channel, the use of only three C3, C4 and Cz channels leaded to a significant drop (mean=9.4% and p=0.016) in accuracy. As a result, although the use of only three C3, C4 and Cz channels certainly alleviates the inconvenience of preparation, but it inevitably causes performance drop which was in our study around 10%. On contrary, the proposed method can bring the benefit of reducing the number of channels with a small drop in accuracy.

After performing channel selection, the remaining channels are ranked using the proposed pruning method. To consider the performance of the proposed ranking method, the accuracy of best ranked channels (from 2 to all the remaining channels) were calculated, and compared with four other

channel selection methods based on Fisher Criterion (FC), Mutual Information (MI), Support Vector Machine (SVM) and Common Spatial Pattern coefficients (CSP). Fig. 3 depicts the averaged accuracy versus different number of channels selected by 5 different channel selection methods. This figure shows that, the classification accuracy of subject 2 was close to random, and the results of channel selection were scattered. According to the applied algorithm, this subject is identified as a "BCI illiterate" meaning he cannot use a BCI. Hence we ignored the results obtained by this subject and compared the rest.

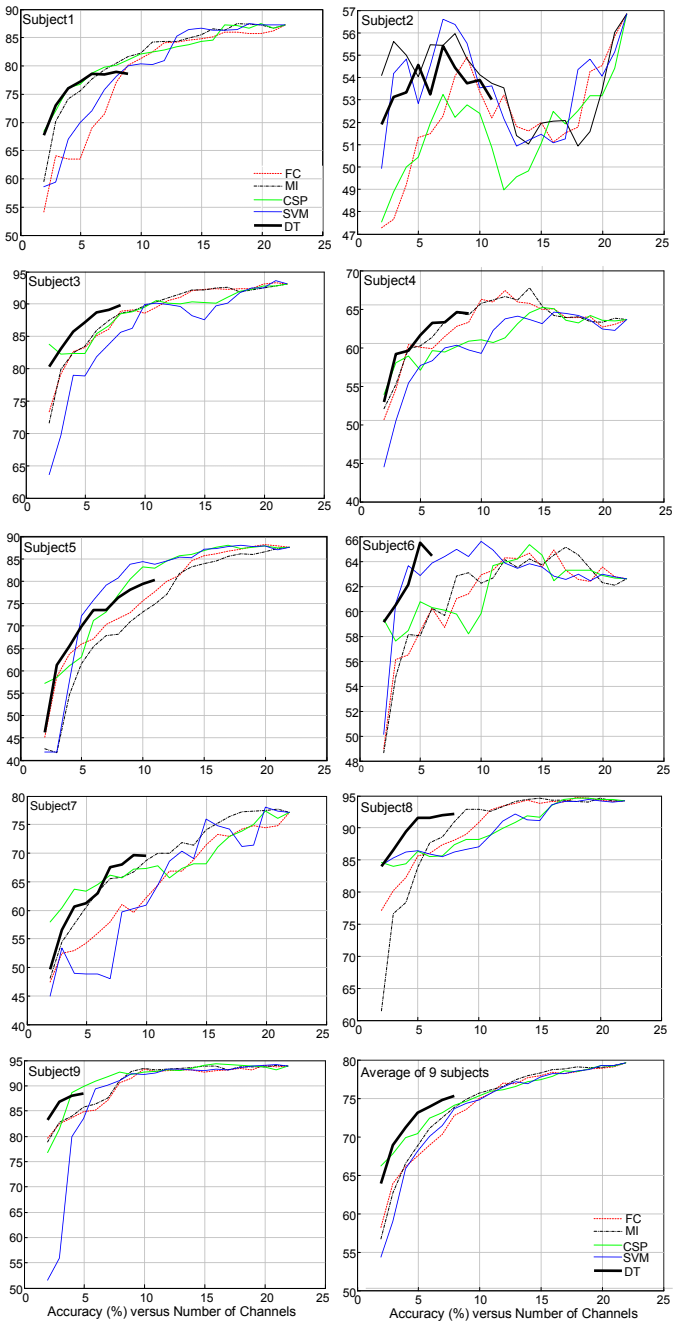

Fig.3 . Comparison of 5 EEG channel selection methods

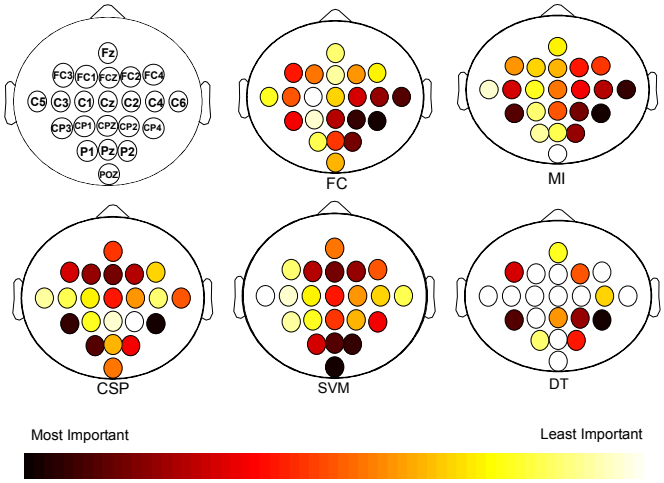

Fig. 4. Visualization of channels importance for subject 1

The results of other subjects illustrate that proposed method outperformed the other channel selection methods. As can be seen, the proposed method yielded superior classification accuracy in selecting 3 to 6 channels. It might be due to selecting channels according to the results of previous chosen channels; hence the selected channels would be more relevant and less correlated to each other.

Beside the proposed method, the CSP-based channel selection method is capable of selecting more relevant channels compared to FC, MI and SVM methods. On the contrary, FC and MI methods perform rather a channel ranking than a channel selection method. Therefore they are mostly not as good as the proposed method in selecting a few channels. As it is visible in Fig.3, a sharp decrease of accuracy around 6 to 2 channels happens for hired SVM method.

Visualization of the channel positions according to their ranks may support the analysis of our applied methods. As the experimental paradigm is well known, we investigated whether the best selected channels were those situated over or close to the motor areas. Fig. 4 visualizes the selected channels of subject 1 for five considered methods where darker colors show more important channels which are selected earlier, and lighter ones show less important channels. It should be noted that in this step cross validation was not applied.

According to Fig. 4, the best channels selected by FC method are neighbor and just in one side of the brain (down-right). Consequently, selecting a few channels results in poor performance because selected channels are full of redundant information without supporting both task activities. MI channel positions are slightly better than FC but still quiet near to each others. In subject 1, SVM recognized top and down of the brain channels as the most important ones, which both parts are not related to motor area. It might be the reason that in Fig. 3 a sharp drop in accuracy around 8 to 2 channels are seen for hired SVMs.

The preference of CSP based method is selecting channels pair wisely from both sides of the brain. The best channels

selected by CSP method are CP4 and CP3 (near motor imagery areas), and after a while some channels from top are also selected. It would be the reason that CSP based method achieved quietly good results in our experiments. As it can be seen, the decision tree method selected just some of neighbor electrodes from both sides of the brain. Thus redundant information is reduced and performance is increased. In summary, visualization presents that the preference of the decision three based method to the other method is selecting channels from neurophysiological relevant areas, and removing redundant and correlated channels.

VI. CONCLUSIONS

This paper presents a decision tree-based method for EEG channel selection in BCI applications. The proposed method first employs a subject-specific multiband filter to filter the EEG, then irrelevant channels are removed using a decision tree. Subsequently, the remaining channels are ranked using a pruning method. Since the decision tree selects channels based on previously chosen features, the selected channels are more relevant and less correlated to each other. Moreover, since training the decision tree and channel selection are performed simultaneously, the proposed method is computationally efficient.

The experimental results showed that the proposed method reduces the averaged number of electrodes from 22 to 8.44, whereas the classification accuracy decreases only 3.63%. While if three typical channels (Cz, C3, C4) are used, the accuracy drops around 9.6%. A comparative study of the proposed method with other channel selection methods using Fisher Criterion, Mutual Information, Support Vector Machine and CSP on 9 subjects for two motor imagery tasks showed that our method outperformed the others in selecting around 3 to 6 channels. A visualization of the selected channels illustrated that the proposed method improves the results by removing some of the neighboring channels and selecting those from both hemispheres of the brain.

ACKNOWLEDGMENT

The authors would like to thank Mr. Hamed Ahmadi from National University of Singapore for his constructive comments on this manuscript.

REFERENCES

[1] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-computer interfaces for communication and control," Clin Neurophysiol, vol. 113, no. 6, pp. 767-791, 2002.

[2] N. Birbaumer, "Brain-computer-interface research: Coming of age," Clin Neurophysiol, vol. 117, no. 3, pp. 479- 483, 2007.

[3] A. Al-Ani and A. Al-Sukker, "Effect of Feature and Channel Selection on EEG Classification," in Proc. IEEE/ EMBS'06, 2006, pp. 2171-2174.

[4] T. N. Lal, M. Schroder, T. Hinterberger, J. Weston, M. Bogdan, N. Birbaumer, and B. Scholkopf, "Support vector channel selection in BCI," IEEE Transactions on Biomedical Engineering,, vol. 51, no. 6, pp. 1003-1010, 2004.

[5] L. Tian, D. Erdogmus, A. Adami, M. Pavel, and S. Mathan, "Salient EEG Channel Selection in Brain Computer Interfaces by Mutual Information Maximization," in Proc. IEEE/ EMBS '05,, 2005, pp. 7064-7067.

[6] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," Journal of Machine Learning Research vol. 3, no. pp. 1157-1182, 2003.

[7] Y. Wang, S. Gao, and X. Gao, "Common Spatial Pattern Method for Channel Selelction in Motor Imagery Based Brain-computer Interface," in Proc. IEEE/ EMBC'05, 2005, pp. 5392-5395.

[8] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K. R. Muller, "Optimizing Spatial filters for Robust EEG Single-Trial Analysis," IEEE Signal Processing Magazine,, vol. 25, no. 1, pp. 41-56, 2008.

[9] L. Rokach and O. Maimon, "Top-down induction of decision trees classifiers - a survey," Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, vol. 35, no. 4, pp. 476-487, 2005.

[10] K. K. Ang, Z. Y. Chin, H. Zhang, and C. Guan, "Filter Bank Common Spatial Pattern (FBCSP) in Brain-Computer Interface," in Proc. IEEE/ IJCNN '08, 2008, pp. 2390-2397.

[11] L. Breiman, Classification and Regression Trees. Boca Raton: Chapman & Hall,, 1993.

[12] H. Ramoser, J. Muller-Gerking, and G. Pfurtscheller, "Optimal spatial filtering of single trial EEG during imagined hand movement," IEEE Transactions on Rehabilitation Engineering,, vol. 8, no. 4, pp. 441-446, 2000.

[13] "Data Sets 2a for the BCI Competition IV", http://www.bbci.de/competition/iv/#datasets.