Dynamically Weighted Classification with Clustering to Tackle Non-stationarity in Brain Computer Interfacing

Sidath Ravindra Liyanage[1], Cuntai Guan[2], Haihong Zhang[2], Kai Keng Ang[2], Jian-Xin Xu[1] and Tong Heng Lee[1]

1 National University of Singapore, Singapore
2 Institute for Infocomm Research, A*STAR, Singapore
sidath@nus.edu.sg, {ctguan, hhzhang, kkang}@i2r.a-star.edu.sg, {elexujx, eleleeth}@nus.edu.sg

Abstract—**This paper addresses an important problem known as EEG non-stationarity in Brain-computer Interfacing. We propose a novel technique called Dynamically Weighted Classification with Clustering (DWCC), which explores hidden states in non-stationary EEG using a modified k-means clustering method by combining cosine distance measure and mutual information criterion. DWCC builds a set of classifiers, one for each pair of clusters from different classes. A dynamically-weighted classifier ensemble network is trained to combine the outputs of the classifiers, where we propose to dynamically assign the weight of a classifier for each test sample based on its distances to the cluster centres associated with the classifier. Experimental results on publicly available BCI Competition IV Dataset 2a yielded a mean accuracy of 81.5% which is statistically significant (t-test p<0.05) compared to the baseline result of 75.9% using a single classifier.**

Keywords-Brain–computer interface (BCI); motor imagery; clustering; classification.

I. Introduction

A brain–computer interface (BCI) is a communication system that does not require any peripheral muscular activity [1]. BCI research aims at the automatic translation of neural commands into control signals to control applications such as text input programs, electrical wheelchairs or neuroprostheses. The field of BCI has been developed with the goal of providing a direct means of communicating internal brain states to the external world [1].

A typical BCI system includes the acquisition of brain signals, the processing and classification of the acquired signals, the feedback of the interpreted brain state, and the use of the classified signals to perform a task. A major challenge in classification of EEG signals is their non-stationarity. The brain signals substantially vary after the initial calibration, such that a classifier trained on one session can rarely be reused in the next experimental session. Long periods of low performance have often been observed even when the classifier is trained with data obtained on the same day. Typically, the models developed using machine learning techniques are based on the assumption that underlying distributions of features are more or less static. However, EEG data are apparently non-stationary due to various factors underlying the distributions change between training and testing sessions. This non-stationarity impedes the continuous use of BCI, particularly for the disabled. Therefore BCI is a difficult and inspiring application area with respect to nonstationarity.

Non-stationary of EEG signals has been identified to be caused by factors such as, changes in the physical properties of the sensors, variabilities in neurophysiological conditions, psychological parameters, ambient noise and motion artifacts. Hence, a few methods such as, Bayesian transduction, transfer learning, active learning and distribution matching have been suggested to address this problem [2].

Studies have shown that classifiers trained using clustered Common Spatial Patterns (CSP) features are able to capture generic and invariant discriminative features of the BCI task and address the non-stationarity inherent in EEG due to subject fatigue, attention, or diverse strategies for motor imageries across different sessions [3]. The effectiveness of clustering and partitioning methods in general pattern classification has also been empirically examined in several studies [4-10]. In addition, theoretical analyses of some of these methods are available in [7-11]. In this study we apply clustering on features selected after CSP transformation and build multiple classifiers on the clustered feature sets in order to address the non-stationarity in the EEG signals.

Partitioning the training data can be achieved by clustering the features based on geometric distribution. Given multivariate data, clustering finds a partition of the points into clusters such that the points within a cluster are more similar to each other than to points in different clusters [12]. It has been suggested that clustering can also be considered as a method to encode background knowledge, allowing the transfer of inferential steps, and schemes that facilitate to identify common features between two situations [2].

Ensemble classifiers are known to combine different classifiers in a complementary manner resulting in improved performance [12]. The base classifiers and the aggregation technique used in ensemble methods are vital factors that affect the classification accuracy [13]. The decision of the ensemble classifier is obtained by weighting the classification decisions of individual classifier by the normalized reciprocal of the distance of each test sample to the cluster centres associated with each classifier in the ensemble. In this study, Support Vector Machine (SVM) classifiers were employed as the base classifiers.

Our main contribution is proposing a novel Dynamically Weighted Classification with Clustering (DWCC) to partition the training data and to combine the classification decisions from multiple classifiers in order to address the non-stationarity inherent in EEG data. In the proposed DWCC method, the EEG data is partitioned into clusters after obtaining the features using Common Spatial Patterns (CSP) algorithm. The clusters are class specific, and the clustered features are combined to form training data sets for multiple classifiers. The multiple classifiers are trained independently on the clustered features to form an ensemble classifier.

II. Methods

The proposed DWCC framework consists of two steps: a training step and a test step. In the training step the EEG data used for training the classifiers are subject to bandpass filter 8-30Hz and is subject to CSP algorithm. The selected features are subject to supervised clustering. The clustered features are used to train multiple classifiers considering all possible combinations of clusters from the two classes. The figure 1 illustrates the framework of the DWCC method.

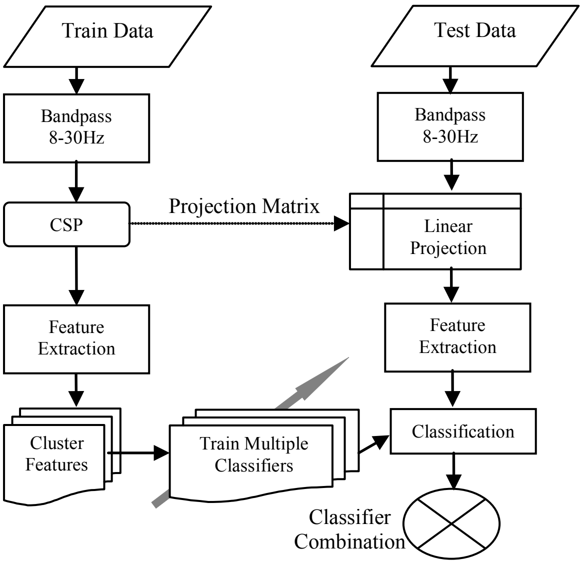

Figure 1. Schematic diagram of the DWCC method.

The CSP filters that are found in training step are used for projection of test data. Features for the test data are selected after the CSP transformation and the multiple classifiers will give their classification decision independently. The classifier decisions are combined using a weighted majority voting method. The weights given to each individual classifier in the ensemble is based on a similarity measure calculated by taking the normalized reciprocals of the distances from the test data sample to the cluster centres that contain the training data for a given classifier.

A. Common Spatial Patterns

Proper preprocessing of EEG data is vital for the ultimate success of the overall BCI system. Non-informative dimensions of the data can be discarded and the features of interest for classification can be selected by Common Spatial Patterns [14].

The CSP algorithm was first presented in [15] as a method to extract the abnormal components from EEG, using a set of patterns that are common to both the normal and the abnormal recordings and have a maximally different proportion of the combined variances. CSP was later extended to classification of movement related EEG signals [14]. The first and last few CSP components (the spatial filters that maximize the difference in variance) are used to classify the trials with high accuracy.

The main concept of CSP is to use a linear transformation to project multichannel EEG data into low–dimensional spatial subspace by a projection matrix. Each row of this projection matrix consists of weights for each channel. If the random variable $\vec{x} \in R^N$ represents the EEG data, recorded through N electrodes, from which the intention of the BCI user $c \in C = \{c_1, \ldots c_M\}$ is to be inferred. Denote the class probability by $P(C_i), i = 1,\ldots,M$ and assume that the EEG data conditioned on any class follows a Gaussian distribution with zero mean, i.e., $P(\vec{x} \mid C_i) = N(0, R_{\vec{x}|C_i}), i = 1,\ldots,M$. Then a linear transformation $w \in R_{N \times L}$ can be found where $L < N$, such that for finite training data using the reduced dimension $\hat{x} = W^T x$. This transformation maximizes the variance of two class signal matrices and leads to an increased classification accuracy in comparison to using $\vec{x}$. These features $\hat{x}$ are clustered in the subsequent step.

B. Clustering of EEG

The EEG data is clustered based on cosine distances among the feature vectors. K-Means algorithm with different distance measures such as Euclidean, Mahalabonis, cityblocks, and cosine were attempted for the clustering of the features. However, Euclidean and Mahalabonis distances were not able to cluster the features satisfactorily. Cityblocks and cosine distance measures were successful in clustering the features with the k-Means algorithm. However, clusters generated using cosine distance measures were found to contain maximum diversity. Therefore, k-Means clustering with cosine distance is employed in generating the initial clusters, which are optimized using the information theoretic criterion.

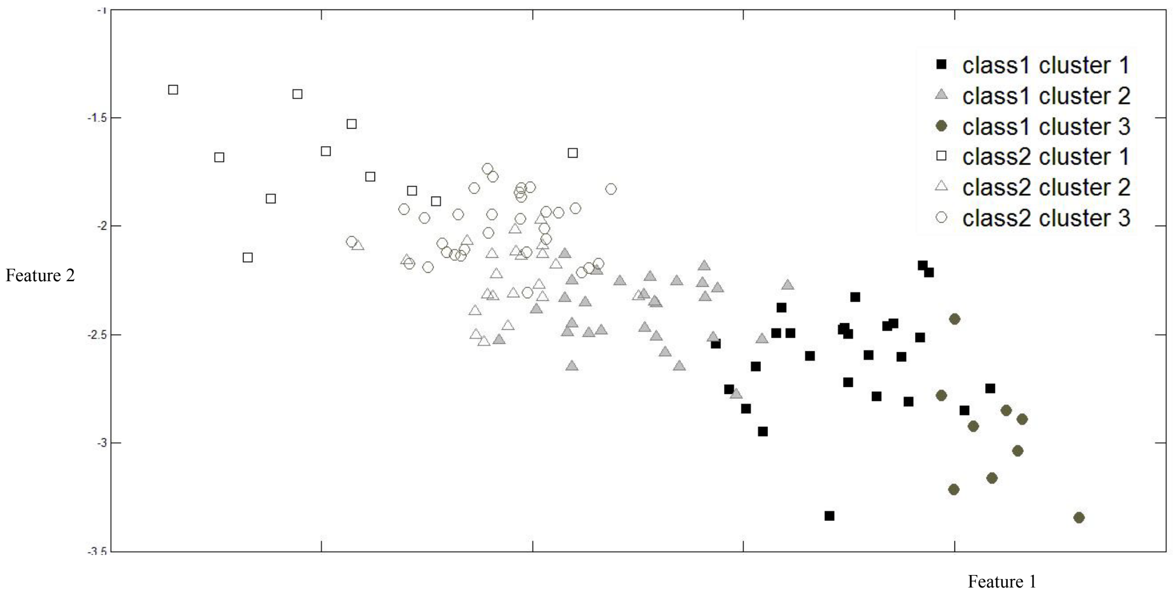

Figure 2. Clustered Features.

Different numbers of clusters from two clusters up to seven clusters were considered. The ensemble classifiers are combined in a novel method in order to maximize the overall classification accuracy. Ensembles with four to forty nine classifiers were trained on the corresponding clusters. It was observed that as the number of classifiers increased the overall increase of performance is marginal while the computational complexity increases quadratically. Best performance was observed when the feature space was partitioned to 3 clusters resulting in 9 classifiers in the ensemble. The figure 2 illustrates the case of three clusters in a two dimensional feature space.

C. Information Theoretic Criterion for Clustering

The initial clusters generated by k-Means algorithm are optimized using mutual information maximization procedure [16]. In this information theoretic criterion, Given a data set $X = \{x_1,...,x_N\}$ of N data items in R^d , a partitional clustering $C = \{c_1,...,c_k\}$ is a way to divide X into K non-overlapped subsets. If C is the space of all possible K-cluster partitions of X, the optimal clustering C^* in C would have maximum mutual information between the data and the clustering: $C^* = \underset{c \in C}{\arg\max}\{I(C;X)\}$.

This is also equivalent to, $C^* = \underset{c \in C}{\arg\min}\{H(X \mid C)\}$.

This criterion is based on the argument that the optimal clustering would maximize the information shared between the clustering and data. It has been shown that, by using Havrda-Charvat's structural entropy measure the conditional entropy can be estimated without any assumptions about the distribution of the data.

Havrda-Charvat's structural entropy is defined as:

$$H_\alpha = \left(2^{1-\alpha} - 1\right)^{-1}\left[\sum_{k=1}^{K} p_k^\alpha - 1\right], \alpha > 0, \alpha \neq 1 .$$

With $\alpha = 2$ the following quadratic Havrda-Charvat's entropy (with the constant coefficient discarded for simplicity) gives:

$$H_2 = 1 - \sum_{k=1}^{K} p_k^2 .$$

The conditional quadratic Havrda-Charvat's entropy of X given C is defined as:

$$H_2(X \mid C) = \sum_{k=1}^{K} p(c_k) H_2(X \mid C = c_k).$$

With this measure of entropy the objective function can be expressed as

$$C^* = \underset{c \in C}{\arg\min}\left\{\sum p(c_k) H_2(X \mid C = c_k)\right\}.$$

The Gaussian kernel in d-dimensional space:

$$G\left(x - a, \sigma^2\right) = \frac{1}{(2\pi\sigma)^{d/2}} \exp\left\{\frac{-\|x - a\|^2}{2\sigma^2}\right\},$$

where σ is the kernel width parameter and a is the center of the Gaussian window. The density estimation of X can be expressed as, $p(x) = \frac{1}{N}\sum_{i=1}^{N} G\left(x - x_i, \sigma^2\right)$.

The quadratic entropy of $p(x)$ can be estimated as,

$$H_2(X) = 1 - \int_x p^2(x) = 1 - \frac{1}{N^2}\int_x \left(\sum_{i=1}^{N} G(x - x_i, \sigma^2) \right)^2 dx$$

Because convolution of two Gaussians remain a Gaussian the above can be expressed as

$$H_2(X) = 1 - \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G(x_i - x_j, 2\sigma^2).$$

In a similar fashion,

$$H_2(X \mid C = c_k) = 1 - \frac{1}{n_k^2}\sum_{x_i \in c_k}\sum_{x_j \in c_k} G(x_i - x_j, 2\sigma^2)$$

where n_k is the number of the data items in cluster c_k. Given this estimate the objective function can be written as

$$C^* = \underset{c \in C}{\arg\max}\left\{ \sum_{k=1}^{K} p(c_k)\frac{1}{n_k^2}\sum_{x_i \in c_k}\sum_{x_j \in c_k} G(x_i - x_j, 2\sigma^2) \right\}.$$

Here the probability of encountering the cluster c_k in C is n_k/N, therefore the conditional entropy based objective function becomes,

$$C^* = \underset{c \in C}{\arg\max}\, CE(C),$$

where,

$$CE(C) = \sum_{k=1}^{K}\frac{1}{n_k}\sum_{x_i, x_j \in c_k} \exp\left\{ \frac{-\|x_i - x_j\|^2}{4\sigma^2} \right\}.$$

Therefore, by maximizing $CE(C)$ the conditional entropy criterion is minimized.

D. Classifier Ensemble

The clustered features are combined to form training feature sets that are used to train individual classifiers in the ensemble of classifiers. The dynamic fusion method to combine classifier decisions in the ensemble to achieve higher classification accuracies exploit the geometrical relationships among clustered training samples to quantify the similarity between test data and training data.

If the classifiers in an ensemble are not of identical accuracy, then it is reasonable to give the more competent classifiers more power in making the final decision. Label output for a test sample x by a classifier *i* in an ensemble of *L* classifiers can be represented in terms of degree of support for each class *j* as,

$$d_{i,j} = \begin{cases} 1, \textit{if } D_i \textit{ labels } x \textit{ in } j \\ 0, \textit{otherwise} \end{cases},$$

where D_i represents i^{th} classifier in the ensemble. The discriminant function for class *j* obtained through weighted voting by *L* classifiers in the ensemble can be expressed as,

$$g_j(x) = \sum_{i=1}^{L} b_i d_{i,j},$$

where b_i is a coefficient for decision given by classifier D_i regarding class *j*. Thus the value of the discriminant function is the sum of the coefficients for these members of the ensemble whose output for x is class *j*.

$$\sum_{i=1}^{L} b_i d_{i,k} = \max_{j=1}^{c} \sum_{i=1}^{L} b_i d_{i,j} \cdots\cdots\cdots (1)$$

For a *c* class classification scenario, classifier decisions from constituent classifiers are combined through weights b_i as shown in equation (1). The class *k* which gets the maximum support through the weighted majority voting scheme, given in equation (1), is chosen as the ensemble's final decision.

In the proposed DWCC method, the weights b_i are actively calculated for each test sample based on the distances from a test sample to the centres of the clusters consisting of training data. The reciprocals of the distances from the test sample to the cluster centres are classwise normalized and summed together. For the two class case, let the clusters that make up the training features for a classifier *i* be denoted as C_1 and C_2. The reciprocal distance from a given test sample to the two cluster center can be denoted as, r_{c1} and r_{c2}. The values normalize for each class where there are *m* clusters for each class,

$$Nr_{c1} = \frac{r_{c1}}{\sum_{c1=1}^{m} r_{c1}} \text{ and } Nr_{c2} = \frac{r_{c2}}{\sum_{c1=2}^{m} r_{c2}}.$$

The final weight b_i assigned to each classifier *i* is calculated as sum of the normalized reciprocal distances associated with the two particular clusters that make up the specific classifier,

$$b_i = Nr_{c1} + Nr_{c2}.$$

Support Vector Machine (SVM) classifiers were employed as the base classifiers in all the ensembles.

III. Experiments

The data set 2A of the fourth BCI Competition [17] was considered in this study. This data set is composed of EEG data collected from 9 subjects that had been recorded during two sessions on different days for each subject. The synchronous BCI data had been collected for four different motor imagery tasks. The imagination of movements of the left hand (class 1), right hand (class 2), both feet (class 3), and tongue (class 4) had been considered as the four motor imagery tasks. Each session had been made up of 6 runs separated by short breaks. One run had included 48 trials (12 for each of the four possible classes), amounting to a total of 288 trials per session.

The subjects had been seated on an armchair in front of a computer screen and at the beginning of a trial (t = 0 s), a fixation cross had appeared on the black screen. Short acoustic warning tones had also been presented at the start of the trial. After two seconds (t = 2 s), a cue had been presented. This cue could have been in the form of an arrow pointing either to the left, right, down or up (corresponding to one of the four classes left hand, right hand, foot or tongue). The cue had appeared and stayed on the screen for 1.25 s and this was expected to induce the subjects to perform the desired motor imagery task. The subjects had been instructed to carry out the motor imagery tasks until the fixation cross disappeared from the screen at t = 6 s without any feedback on their performance. A short break had been given before the next trial and this procedure had been repeated for each of the 6 runs in a session.

EEG signals had been recorded from 22 scalp positions, mainly covering the primary motor cortices bilaterally. The signals had been sampled at 250 Hz and had been subjected to a bandpass filter between 0.5 Hz and 100 Hz. The sensitivity of the amplifier had been set to 100 μV. An additional 50 Hz notch filter had been utilized to suppress line noise.

IV. RESULTS AND DISCUSSIONS

The performance of the DWCC framework was evaluated under different number of clusters. The number of component classifiers in the ensemble depends on the number of clusters. Too many clusters result in smaller partitions of training data that lead to over fitting and result in lower classification accuracies. Therefore only six different cluster numbers, from two to seven clusters, resulting in four to forty nine individual classifiers respectively were considered.

TABLE I. CLASSIFICATION ACCURACY RATES (%) ON DATA SET 2A OF BCI COMPETITION IV.

Subject	Baseline	DWCC: Number of Clusters					
		2	3	4	5	6	7
A1	87.3	95.2	**95.4**	94.8	94.4	94.8	94.6
A2	56.8	63.8	**64.2**	64.0	62.5	63.8	63.4
A3	93.1	**96.9**	96.8	96.1	96.5	95.2	95.9
A4	63.6	66.7	**67.2**	66.7	66.8	66.4	65.4
A5	54.86	**75.9**	75.9	75.5	75.4	75.7	75.5
A6	62.6	64.9	65.2	63.6	**65.8**	63.7	64.5
A7	77.1	78.1	78.1	77.9	78.1	78.4	**78.7**
A8	94.2	96.1	96.1	**96.3**	95.2	95.7	95.5
A9	93.8	92.6	**93.1**	92.7	93.2	92.8	**93.2**
Mean	75.92	81.26	**81.48**	81.01	80.86	80.77	80.90
S.D.	16.65	14.31	14.21	14.409	14.058	14.084	14.4
P	-	0.039	0.032	0.047	0.047	0.059	0.048

P-values denote results of pairwise t-tests against the baseline.

The results obtained for BCIC IV data set 2a are shown in table 1. The nine subjects are denoted as A1 to A9. All classifiers were trained on the training data obtained on the previous date and were evaluated on test data which had been collected on a subsequent date. The highest classification accuracies for each subject are shown in boldface. The mean accuracies and standard deviations calculated for all the subjects are denoted as mean and S.D. in the table 1. Highest mean accuracy of 81.48% is achieved by the DWCC with three clusters.

The baseline classification utilizes a single SVM classifier. The features used for training the baseline SVM classifier were subject to the same preprocessing steps as the other cases except clustering. The observed mean baseline accuracy is 75.92%. The baseline result was compared against the results obtained using DWCC method. A series of pairwise t-tests were carried out between the baseline results and each of the DWCC approaches. The P-value denotes the probability, under the null hypothesis (difference of means is zero), of observing a value as extreme or more extreme of the test statistic t. i.e. at a confidence level of 0.05 if the P value is less than 0.05 then the two means are significantly different.

The mean accuracies from DWCC method are found to be significantly higher than the baseline, except in the case of six clusters, at a confidence level of 0.05. It should be noted that accuracies of all subjects show improvements over the baseline in all DWCC based cases irrespective of the number of clusters. Results suggest that the proposed method can significantly improve classification accuracies in most cases.

The figure 2 shows the clustered features for the case where features were clustered into three partitions. In line with our basic assumption that non-stationary sources form clusters in the feature space that are distinctly apart from one another, it is clear that clustering has been able to uncover encoded background knowledge. Therefore, it is suggestive that inherent non-stationarity in EEG data can be effectively addressed by the proposed clustering based DWCC method.

V. CONCLUSION

In this study, a novel framework to improve classification of non-stationary EEG data named DWCC was proposed. In DWCC approach EEG data is partitioned using clustering and multiple classifiers are trained on the partitioned features. It is assumed that regions in the feature space that are substantially apart from other features are due to non-stationarity in EEG signals. Training data for multiple classifiers are selected by supervised partitioning of the features of the two class training data. The decisions of ensemble classifiers are obtained by weighted majority voting of the classification decisions of individual classifiers. A novel combination method based on the distance of the test sample to the constituent clusters that form the specific classifier is used to give weights to the classifier decisions.

The proposed DWCC method has been evaluated on a publicly available data set from the BCI Competition IV. Empirical results suggest that the method can yield statistically significant improvement in classification accuracy for non-stationary EEG data.

Future work based on this approach might include an adaptive clustering framework where clusters will be added and removed in an online data driven manner.

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. 113 767–91, 2002.

[2] J. Quinonera-Candela, M.Sugiyama, Schwaighofer, A., Lawrence, N. D.,.Dataset Shift in Machine Learning, MIT Press, 2009.

[3] M. Krauledat, M. Schröder, Benjamin Blankertz, Klaus-RobertMüller, Reducing Calibration Time For Brain-Computer Interfaces: A Clustering Approach, In Bernhard Schölkopf, J. Platt, and T. Hoffman, editors, Advances in Neural Information Processing Systems 19, Cambridge, MA, MIT Press, pp. 753-760, 2007.

[4] N. Chawla, T. Moore, L. Hall, L. Bowyer, P. Kegelmeyer, and C. Springer, "Distributed Learning with Bagging-like Performance," Pattern Recognition Letters, vol. 24, pp. 455-471, 2003.

[5] K. Woods, W. Kegelmeyer, and K. Bowyer, "Combination of Multiple Classifiers Using Local Accuracy Estimates," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, no. 4, pp. 405-410, Apr. 1997.

[6] T. Dietterich, "An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization," Machine Learning, vol. 40, no. 2, pp. 139-158, 2000.

[7] Y. Freund and R. Schapire, "Experiments with a New Boosting Algorithm," Proc. 13th Int'l Conf. Machine Learning, pp. 148-156, 1996.

[8] Y. Freund and R.E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," J. Computer and System Sciences, vol. 55, pp. 119-139, 1997.

[9] D. Frosyniotis, A. Stafylopatis, and A. Likas, "A Divide-and- Conquer Method for Multi-Net Classifiers," Pattern Analysis and Applications, vol. 6, pp. 32-40, 2002.

[10] L. Breiman, "Bagging Predictors," Machine Learning, vol. 24, no. 2, pp. 123-140, 1996.

[11] D. Tao, X. Tang, X. Li, and X. Wu, "Asymmetric Bagging and Random Subspace for Support Vector Machines-Based Relevance Feedback in Image Retrieval," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 28, no. 7, pp. 1088-1099, July 2006.

[12] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, New York:John Wiley, 2001.

[13] F. C. Meinecke, S. Harmeling, and K.-R. Müller, Robust ICA for Super-Gaussian Sources, in: C. G. Puntonet and A. Prieto, eds., Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA), 2004.

[14] H. Ramoser, J. Mueller-Gerking, and G. Pfurtscheller, Optimal spatial filtering of single trial EEG during imagined hand movement, IEEE Trans. Rehabil. Eng., vol. 8, no. 4, pp. 441–446, Dec. 2000.

[15] Z. J. Koles, The quantitative extraction and topographic mapping of the abnormal components in the clinical EEG, Electroencephalogr. Clin. Neurophysiol., vol. 79, pp. 440–447, 1991.

[16] N. X. Vinh, J. Epps, and J. Bailey. Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance ,Journal of Machine Learning Research 11, pp. 2837-2854, 2010.

[17] B. Blankertz, BCI Competition IV, Fraunhofer FIRST.IDA, http://ida.first.fraunhofer.de/projects/bci/competition_iv/.