IOP PUBLISHING JOURNAL OF NEURAL ENGINEERING
J. Neural Eng. **10** (2013) 000000 (10pp) UNCORRECTED PROOF

Dynamically weighted ensemble classification for non-stationary EEG processing

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

[1] Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[2] Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), 1 Fusionopolis Way, 21-01 Connexis, Singapore 138632, Singapore

E-mail: sidath@nus.edu.sg, ctguan@i2r.a-star.edu.sg, hhzhang@i2r.a-star.edu.sg, kkang@i2r.a-star.edu.sg, elexujx@nus.edu.sg and eleleeth@nus.edu.sg

Received 7 December 2012
Accepted for publication 15 March 2013
Published DD MM 2013
Online at stacks.iop.org/JNE/10/000000

Abstract
Objective. The non-stationary nature of EEG poses a major challenge to robust operation of brain–computer interfaces (BCIs). The objective of this paper is to propose and investigate a computational method to address non-stationarity in EEG classification.*Approach.* We developed a novel dynamically weighted ensemble classification (DWEC) framework whereby an ensemble of multiple classifiers are trained on clustered features. The decisions from these multiple classifiers are dynamically combined based on the distances of the cluster centres to each test data sample being classified. *Main Results.* The clusters of the feature space from the second session spanned a different space compared to the clusters of the feature space from the first session which highlights the processes of session-to-session non-stationarity. The session-to-session performance of the proposed DWEC method was evaluated on two datasets. The results on publicly available BCI Competition IV dataset 2A yielded a significantly higher mean accuracy of 81.48% compared to 75.9% from the baseline SVM classifier without dynamic weighting. Results on the data collected from our twelve in-house subjects yielded a significantly higher mean accuracy of 73% compared to 69.4% from the baseline SVM classifier without dynamic weighting. *Significance.* The cluster based analysis provides insight into session-to-session non-stationarity in EEG data. The results demonstrate the effectiveness of the proposed method in addressing non-stationarity in EEG data for the operation of a BCI.

(Some figures may appear in colour only in the online journal)

1. Introduction

A brain–computer interface (BCI) is a communication system that does not require any peripheral muscular activity, with the goal of providing a direct means of communicating internal brain states to the external world [1].

A major challenge for BCI research is the non-stationarity of brain activity occurring continuously in association with diverse behavioural and mental states [2]. Non-stationarity refers to a change in class definitions over time and therefore causes a change in the distributions from which the data are drawn [3]. Consider the Bayesian posterior probability of a class ω to which instance x belongs, $P(\omega|x) = P(x|\omega) \cdot P(\omega)/P(x)$, non-stationarity is defined as any scenario where the posterior probability changes over time, i.e., $P_{t+1}(\omega|x)\, P_t(\omega|x)$, where ω is the class to which data instance x belongs.

The non-stationarity of EEG signals is caused by factors such as changes in the physical properties of the sensors, variabilities in neurophysiological conditions, psychological parameters, ambient noise and motion artefacts. Two main factors contributing to non-stationarity as reported in [4, 5] are:

the differences between the samples extracted from a training session and the samples extracted during an online session, and the changes in the users brain activity during online operation. As a result, the general hypothesis that the signals sampled in the training set follow a similar probability distribution to the signals sampled in the test set from a different session is violated [6].

Kaplan has studied the fast dynamics of quasi-stationary episodes in EEG signals and has identified different operating modes in the EEG time series [7]. Several machine learning techniques have been attempted recently to address the non-stationarity issue in BCI [8, 10, 9]. Robust principal component analysis has been proposed for visualizing spatial patterns with the most prominent variability in the data in order to automatically identify and reject outlying non-informative signals [8]. Stationary linear discriminant analysis attempts to find a direction in the feature space which is both discriminative and stationary [9]. Stationary sub-space analysis is an unsupervised learning method that finds sub-spaces in which data distributions stay invariant over time [10]. Methods such as Bayesian transduction, transfer learning, active learning and distribution matching have also been proposed to address the non-stationarity issue [11]. Even though it would be interesting to study the application of these methods, it exceeds the scope of the current study.

Density estimation for determining class conditional distributions has been attempted by Hastie *et al* [12] for discriminant analysis of Gaussian mixtures. The use of probability forecasting has been extensively studied by Dawid *et al* in [13] for probabilistic expert systems, while the Bayesian combination of classifiers has been extensively studied by Ghahramani *et al* in [14]. Recent advances include a unifying framework for determining linear combiners for classifier ensembles [15] and the Bayesian combination of multiple imperfect classifiers proposed by Simpson *et al* in [16].

In this study we propose a dynamically weighted ensemble classification (DWEC) framework to cluster features extracted using common spatial patterns (CSP) and build an ensemble of multiple classifiers on the clustered features in order to address the session-to-session non-stationarity in the EEG data for the operation of a BCI. Clustering the features extracted after CSP filtering facilitates the identification of different modes in the EEG. Classifiers trained on the clustered features offer complimentary decisions. Improved accuracies can be achieved by appropriately combining the decisions from an ensemble of multiple classifiers. An ensemble framework for constructing subject independent BCI classification was also attempted by Fazli *et al* in [17].

For stationary data, the Bayesian optimal classifier combination was proposed by Kuncheva [18]. This work extends the concept of Bayesian optimal combination for non-stationary data. Since the underlying distribution of the test data is unknown, classification accuracies for each classifier need to be re-estimated. Particularly, we consider each test sample to dynamically estimate the classification accuracy based on the relative location of samples with respect to the clusters.

The remainder of this paper is organized as follows: section 2 describes the synthesized materials followed by the methods in section 3. Section 4 presents comparative results and the discussion. Finally, section 5 concludes this paper.

2. Materials

Two datasets were evaluated using the proposed method: the publicly available BCI Competition IV dataset 2A [19] and the ARTS12 motor imagery dataset collected from 12 healthy subjects.

The BCI Competition IV dataset 2A comprises EEG data collected from nine subjects that were recorded during two sessions on separate days for each subject. The data were collected on four different motor imagery tasks: left hand (class 1), right hand (class 2), both feet (class 3) and tongue (class 4). Each session comprised six runs separated by short breaks, each run comprised 48 trials (12 for each class), amounting to a total of 288 trials per session. Only the class 2 classification between left hand and right hand motor imagery was considered in the current study. For more details on the protocol please refer to [19]. The motor imagery data from the first session were used to train the classifiers and the motor imagery data from the second session were used as the test data.

The ARTS12 motor imagery data were collected using Nuamps EEG acquisition hardware (http://www.neuroscan.com) with unipolar Ag/AgCl electrodes, digitally sampled at 250 Hz with a resolution of 22 bits for voltage ranges of ± 130 mV. EEG signals from 22 scalp positions, mainly covering the primary motor cortices bilaterally, were recorded. The sensitivity of the amplifier was set to $100\,\mu$V. Twelve healthy subjects were recruited for the study. Two subjects chose to perform left hand motor imagery while the remaining ten subjects chose to perform on the right hand. The subjects were instructed, in the form of visual cues displayed on the computer screen, to perform kinaesthetic motor imagery of the chosen hand and to rest during the background rest condition.

EEG data were collected in two sessions for this study from each subject on two different days. In the first session, two runs of EEG data were collected from a subject while performing motor imagery of the chosen hand and during the background rest condition. In the second session on another day, three runs of EEG data were collected while performing motor imagery of the chosen hand and during the background rest condition. Each run lasted approximately 16 min, which comprised 40 trials of motor imagery and 40 trials of background rest condition. The motor imagery data collected during the first session were used to train the classifiers and the motor imagery data from the subsequent sessions were used as test data.

3. Methods

The proposed framework consists of two steps: training and testing. In the training step, the EEG data used for training are subject to pre-processing and feature extraction. In this

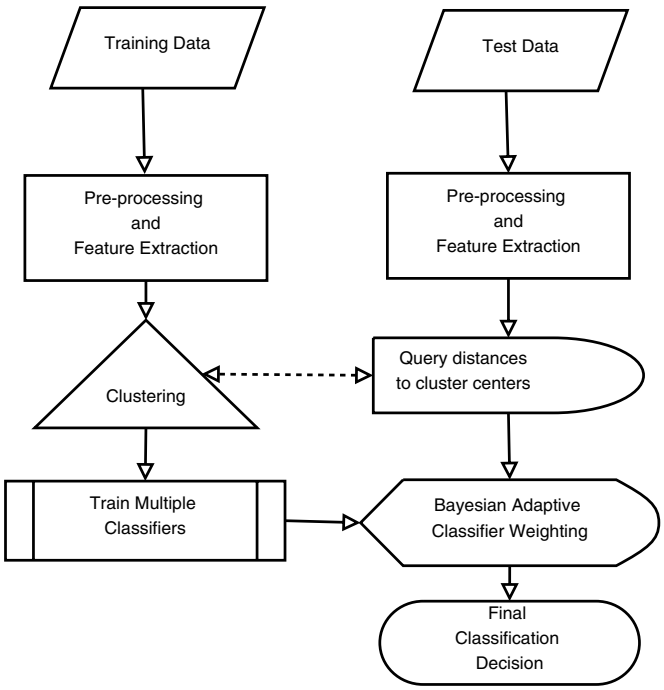

Figure 1. Schematic diagram. The training data and test data are pre-processed and features are extracted. Training data are clustered and multiple classifiers are trained on clustered features. The decisions from multiple classifiers are dynamically weighted to arrive at the final classification decision.

experiment EEG data were bandpass filtered at 8–30 Hz and spatially filtered using the CSP algorithm. The extracted features of each class were subjected to clustering separately. The clustered features were subsequently used to train an ensemble of multiple classifiers by combining all possible clusters from each class.

In the testing step, the EEG data used for testing were subjected to pre-processing and feature extraction similar to the training data. In this experiment the EEG data used for testing were bandpass filtered at 8–30 Hz and spatially filtered using the CSP filter trained during the training step. The extracted features were then evaluated by the ensemble of multiple classifiers. The decisions from the classifiers in the ensemble were dynamically combined using a weighted majority voting method based on a similarity measure computed from the distance of the test data to each cluster centre of each classifier.

The following subsections provide a more detailed description of the proposed framework. Figure 1 summarizes the processes involved in the proposed method.

3.1. Feature extraction

EEG signals resulting from motor imagery have been found to contain specific temporal, spectral and spatial features, that enable them to be recognized automatically [20]. For example, imagining a left hand movement is known to trigger a decrease of power, known as event related desynchronization (ERD), in the μ and β rhythms over the right motor cortex [20]. The increase of band power that occurs after the motor imagery is known as event related synchronization (ERS)[20].

The CSP algorithm, which is effective in computing spatial filters for detecting ERD/ERS effects [21, 22], was used to extract the features from the EEG data. CSP was extended to multi-class problems in [23] and further extension and robustification using the simultaneous optimization of spatial and frequency filters has been proposed in [24–26].

The CSP algorithm computes the transformation matrix W to yield features whose variances are optimal for discriminating two classes of EEG measurements by solving the eigenvalue decomposition problem

$$\Sigma_1 W = (\Sigma_1 + \Sigma_2)\, W \Delta, \tag{1}$$

where Σ_1 and Σ_2 are estimates of the covariance matrices of band-pass filtered EEG measurements of the respective motor imagery actions and Δ is the diagonal matrix that contains the eigenvalues of Σ_1 in descending order of magnitude. Spatial filtering is performed by linearly transforming the EEG measurements using

$$Z_i = W^T E_i, \tag{2}$$

where $E_i \in \mathbb{R}^{\mathrm{ch}\times n_t}$ denotes the single-trial EEG measurement of the ith trial, $Z_i \in \mathbb{R}^{\mathrm{ch}\times n_t}$ denotes E_i after spatial filtering, $W \in \mathbb{R}^{\mathrm{ch}\times \mathrm{ch}}$ denotes the CSP projection matrix, ch is the number of channels, n_t is the number of EEG samples and T denotes the transpose operator.

The CSP features of the ith trial are then given by

$$x_i = \log \frac{\mathrm{diag}(\bar{W}^T E_i E_i^T \bar{W})}{\mathrm{tr}[\bar{W}^T E_i E_i^T \bar{W}]}, \tag{3}$$

where $x_i \in \mathbb{R}^{2m}$ are CSP features, $\bar{W}$ represents the first m and the last m columns of W, $\mathrm{diag}(\cdot)$ returns the diagonal elements of the square matrix and $\mathrm{tr}[\cdot]$ returns the sum of the diagonal elements of the square matrix. Note that $m = 3$ was used in this study.

3.2. Clustering of EEG with the minimum entropy criterion

Since the features extracted using the CSP algorithm are the solutions of a generalized eigenvalue problem, a multiple of the extracted feature vectors is again a solution to the eigenvalue problem. It should be noted that the feature space is inherently non-Euclidean when comparing the extracted features. An appropriate comparison for two feature vectors x_1 and x_2 in this non-Euclidean space is the angle between these two vectors, measured by the cosine distance, $d(x_1, x_2) = 1 - (\frac{x_1 \cdot x_2^T}{|x_1| \cdot |x_2|})$. Clustering EEG data using the cosine distance between the feature vectors extracted by CSP has been shown to yield correct source signals in high-dimensional data [27].

In this work, the features extracted from the training data were initially clustered using the k-means algorithm with the cosine distance measure. The resulting initial clusters were optimized using the minimum entropy criterion[28]. The normalized information distance measures were used to quantify the amount of information shared between clusters.

Let a spatially filtered set of features $\mathcal{X} = \{x_1, \ldots, x_{n_t}\}$, where x_i is a feature vector such that $x_i \in \mathbb{R}^n$ and n_t is the number of trials in the training data used for clustering. If $\mathcal{C}$ is the space of all possible $K-$cluster partitions of $\mathcal{X}$, a partitional clustering $C = \{c_1, \ldots, c_K\}$ is a way to divide $\mathcal{X}$ into K non-overlapping subsets such that $C \in \mathcal{C}$. In the minimum entropy criterion, the optimal clustering $C^* \in \mathcal{C}$ would have maximum mutual information between the data and the clustering:

$$C^* = \arg\max_{\mathcal{C}} \{I(C; X)\}. \tag{4}$$

The entropy relation of (4) can be expressed as: $C^* = \arg\max_{\mathcal{C}}\{H(X|C)\}$, where $H(X|C)$ denote the conditional entropy of X for a given clustering C.

The minimum entropy criterion is based on the argument that optimal clustering would maximize the information shared between the clustering and data. It has been shown that, by using Havrda–Charvat structural entropy measure, the conditional entropy can be estimated without any assumptions about the distribution of the data. Havrda–Charvat structural entropy is defined as:

$$H_\alpha = (2^{1-\alpha} - 1)^{-1} \left[\sum_{k=1}^{K} p_k^\alpha - 1 \right], \alpha > 0, \alpha \neq 1. \tag{5}$$

Where α is the structural dimension, K is the number of partitions and p_k^α is the probability of a sample being included in the kth partition in the α-dimension [29].

The equation (5) can be simplified by discarding the constant coefficient and with $\alpha = 2$ to give: $H_2 = 1 - \sum_{k=1}^{K} p_k^2$.The conditional quadratic Havrda–Charvat entropy of $\mathcal{X}$ given C can be defined as:

$$H_2(\mathcal{X}|C) = \sum_{k=1}^{K} p(c_k) H_2(\mathcal{X}|C = c_k). \tag{6}$$

With the measure of conditional entropy (6), the objective function (7) can be expressed as:

$$C^* = \arg\min_{\mathcal{C}} \left\{ \sum_{k=1}^{K} p(c_k) H_2(\mathcal{X}|C = c_k) \right\}. \tag{7}$$

Estimating the conditional entropy without information about the underlying probability distributions is difficult. A solution is to use the Parzen window [30] method for density estimation as suggested in [31]. Principe *et al* used the Parzen window method in conjunction with quadratic Renyis entropy for density estimation [32]. In a similar manner we use the Parzen window [30] to estimate the conditional entropy. Given that a Gaussian kernel in n-dimensional space is

$$G(x - a, \sigma^2) = \frac{1}{(2\pi\sigma)^{\frac{n}{2}}} \exp\left(-\frac{\|x - a\|^2}{2\sigma^2} \right), \tag{8}$$

where σ is the kernel width parameter and a is the centre of the Gaussian window, the probability density estimation of $x \in \mathcal{X}$ can be expressed as

$$p(x) = \frac{1}{n_t} \sum_{i=1}^{n_t} G\left(x - x_i, \sigma^2\right). \tag{9}$$

The quadratic entropy of the features $\mathcal{X}$ can then be estimated by

$$\begin{aligned} H_2(\mathcal{X}) &= 1 - \int_x p^2(x)\, \mathrm{d}x \\ &= 1 - \frac{1}{(n_t)^2} \int_x \left(\sum_{i=1}^{n_t} G\left(x - x_i, \sigma^2\right) \right)^2 \mathrm{d}x. \end{aligned} \tag{10}$$

Since convolving two Gaussians yields a Gaussian, equation (10) can be expressed as

$$H_2(\mathcal{X}) = 1 - \frac{1}{(n_t)^2} \sum_{i=1}^{n_t} \sum_{j=1}^{n_t} G(x_i - x_j, 2\sigma^2). \tag{11}$$

In a similar manner, the conditional quadratic entropy can be estimated as

$$H_2(\mathcal{X}|\mathcal{C} = c_k) = 1 - \frac{1}{t_k^2} \sum_{x_i \in c_k} \sum_{x_j \in c_k} G(x_i - x_j, 2\sigma^2), \tag{12}$$

where t_k is the number of the data items in cluster c_k. Given the estimate in equation (12), the objective function (7) can be written as

$$C^* = \arg\max_{\mathcal{C}} \left\{ \sum_{k=1}^{K} p(c_k) \frac{1}{t_k^2} \sum_{x_i \in c_k} \sum_{x_j \in c_k} G\left(x_i - x_j, 2\sigma^2\right) \right\}. \tag{13}$$

Here the probability of encountering the cluster c_k in C is $\frac{t_k}{n_t}$. Therefore the conditional entropy ε based objective function becomes

$$C^* = \arg\max_{\mathcal{C}} (\varepsilon(\mathcal{C})), \tag{14}$$

where,

$$\varepsilon(\mathcal{C}) = \sum_{k=1}^{K} \frac{1}{t_k} \sum_{x_i, x_j \in C} \left(\exp\left(-\frac{\|x_i - x_j\|^2}{4\sigma^2} \right) \right). \tag{15}$$

Therefore, by maximizing $\varepsilon(\mathcal{C})$, the conditional entropy criterion is minimized.

3.3. Base classifier

The class-wise training data partitioned to clusters were used to train the ensemble. Individual SVM classifiers that make up the ensemble were trained independently. SVM has been found to yield the highest classification accuracies for synchronous BCI experiments [33]. Dara *et al* [34] have shown that the classification performance of a single SVM classifier can be surpassed by using an ensemble of SVM classifiers. It has also been shown that a combination of different SVM classifiers expands the regions of the test sample resulting in correct classifications. If there are L different SVM classifiers in an ensemble that has been trained independently on different training samples, then each SVM classifier would have a different generalization performance [35].

SVM classifiers have been known to show good generalization performance and provide easy-to-learn exact parameters for the global optimum [35]. Considering all these factors, SVM classifiers with linear kernels were used as base classifiers in the ensemble.

3.4. The DWEC method for non-stationary data

A classifier is any function $\Lambda : \mathbb{R}^n \rightarrow \Omega$, that maps a given object $x \in \mathbb{R}^n$, where $\mathbb{R}^n$ is the feature space, to a class label ω. Let the class label ω be a random variable that can take values in the set of class labels $\Omega = \{\omega_1, \ldots, \omega_\Gamma\}$, where Γ is the number of classes. The class with the highest posterior probability is the most natural choice for a given object $x \in \mathbb{R}^n$, where $\mathbb{R}^n$ is the feature space. In the canonical model of a classifier [36], a set of Γ discriminant functions, $G = \{g_1(x), \ldots, g_\Gamma(x)\}$, $g_i : \mathbb{R}^n \rightarrow \mathbb{R}, i = 1, \ldots, \Gamma$, each yielding a score for the respective class, is generated. The final output class label of the classifier is determined according to the maximum membership rule. The maximum membership rule can be given as,

$$\Lambda(x) = \omega_{i*} \in \Omega \leftrightarrow g_{i*}(x) = \max_{k=1,\ldots,\Gamma} \{g_i(x)\}.$$

In an ensemble consisting of L such classifiers where each classifier Λ_j, produces a class label $s_j \in \Omega$ where $j = 1, \ldots, L$. Thus for any object $x \in \mathbb{R}^n$ to be classified, the outputs from the L classifiers produce a vector $s = [s_1, \ldots, s_L]^T \in \Omega^L$.

The Bayesian optimal weighted majority voting for combining an ensemble of classifiers was defined in [18]. The label outputs produced by each classifier in the ensemble are represented as degrees of support for each class in the following manner:

$$\lambda_{j,k} = \begin{cases} 1, & \text{if } \Lambda_j \text{ labels } x \text{ in class } \omega_k \\ 0, & \text{otherwise.} \end{cases}$$

The discriminant function for class ω_k obtained through weighted voting is, $g_k(x) = \sum_{j=1}^{L} b_j \lambda_{j,k}$, where b_j is a coefficient for classifier Λ_j. Thus the value of the discriminant function would be the sum of the coefficients for these members of the ensemble whose output for x is ω_k. In this context, the optimal set of discriminant functions based on outputs of the L classifiers is

$$g_k(x) = \log P(\omega_k) P(x|\omega_k), k = 1, \ldots, \Gamma.$$

Kuncheva [18] has shown that in an ensemble of L classifiers with individual training accuracies $p_1, \ldots, p_L$ the optimal set of discriminant functions can be achieved by weighted majority voting with individual weights

$$b_j \propto \log \frac{p_j}{1 - p_j}, \tag{16}$$

where p_j is the training accuracy of the jth classifier where $j = 1, \ldots, L$.

The equation (16) is applicable only to stationary data where the distribution of the training data is similar to the distribution of the test data. In the presence of non-stationarity, using equation (16) with training accuracies would not lead to the optimal set of discriminant functions. Therefore under non-stationarity, the accuracies for each test sample should be considered individually to reach the optimal set of discriminant functions.

Since the performance of the classifiers is not known for the test samples, the weights b_j are actively calculated for each test sample based on estimated individual accuracies of classifiers in the ensemble in the proposed method. An estimate for the classification accuracy of each classifier is dynamically calculated based on the distances from test sample to the centres of the clusters consisting of the training data.

The proposed method, in principle, is applicable to classification problems with more than two classes. For simplicity, we will describe the algorithm for binary classification. In the proposed method, the training data are partitioned by clustering the features of the two classes separately. Let U and V be the number of clusters of class 1 and class 2 respectively. Let the clusters of class 1 be denoted by c_{1u}, where $u = 1, \ldots, U$ and clusters of class 2 by c_{2v}, where $v = 1, \ldots, V$. Each pair of clusters c_{1u} and c_{2v} correspond to a specific classifier Λ_j, in the ensemble, where $j = 1, \ldots, L$. Therefore, the number of classifiers in the ensemble, $L = U * V$.

Let the distance from the sample to the cluster centre c_{1u} be d_u and the distance to cluster centre c_{2v} be d_v. Let the ratio between the two distance measures be denoted by d_{uv}, where, $d_{uv} = \frac{d_u}{d_v}$.

A function to estimate the probability of correct classification based on the distance measures to the centres of clusters c_{1u} and c_{2v} consisting of training samples for the classifier is defined as

$$p_{uv}(x_t) = 1 - \frac{1}{2} \exp\left(-\frac{1}{\psi_{uv}^2} (\log(d_{tu}) - \log(d_{tv}))^2\right) \tag{17}$$

where $t = 1, \ldots, n_t$ denotes the index of the training samples in x_t and p_{uv} is the estimated accuracy of the classifier made from clusters c_{1u} and c_{2v}. d_{tu} is d_u for the tth training sample.

This function to estimate classification accuracy satisfies the following limits: $p_{uv} \rightarrow 1$, when $d_{uv} \rightarrow \infty$ and $p_{uv} \rightarrow 1$, when $d_{uv} \rightarrow \infty$. It should also be noted that $p_{uv} \in [0.5, 1]$ and $p_{uv} = 0$, when $d_{uv} = 1$. ψ_{uv} is a parameter whose optimal value should be found by optimizing the objective function given in equation (18) on the training data given by

$$f(\psi_{uv}) = \left[\frac{1}{n_t}\left[\sum_{t=1}^{n_t} p_{uv}(x_t)\right] - p_j\right]^2 \tag{18}$$

where p_j is the training accuracy of the jth classifier where $j = 1, \ldots, L$. In order to find an exact solution for the ψ_{uv} parameter by optimizing the objective function given in equation (18), it must be monotonically decreasing. It can be shown that

$$\frac{\partial p_{uv}}{\partial \psi_{uv}^2} = -\frac{1}{2} \exp\left(-\frac{1}{\psi_{uv}^2}\left(\log\left(\frac{d_{tu}}{d_{tv}}\right)\right)^2\right)\left(\log\left(\frac{d_{tu}}{d_{tv}}\right)\right)^2 (\psi_{uv}^2)^{-2} \leqslant 0. \tag{19}$$

Equation (19) implies $\frac{\partial \frac{1}{n_t}\left[\sum_{t=1}^{n_t} p_{uv}(x_t)\right]}{\partial \psi_{uv}^2} \leqslant 0$. Therefore an exact solution for the ψ_{uv} parameter can be found by optimizing equation (18). After the optimal ψ_{uv} parameter is found the accuracy can be estimated for each test sample by substituting the ψ_{uv} parameter value in equation (17) as,

$$p_{uv}(x) = 1 - \frac{1}{2} \exp\left(-\frac{1}{\psi_{uv}^2} (\log(d_u) - \log(d_v))^2\right) \tag{20}$$

Input

- Training data *Train* with correct labels $\omega_i \in \Omega = \{\omega_1, ..., \omega_\Gamma\}$ representing Γ classes.
- Integer K specifying number of clusters.
- Test data *Test*

Output

- Predicted class labels $s \in \Omega$ for *Test*.

Training

(i) Cluster training data of each class into K clusters.
(ii) Train classifiers on clustered training data.
(iii) Evaluate distances d_{tu} and d_{tv} from the cluster centres to training samples x_t, where $t = 1, \ldots, n_t$.
(iv) Optimize objective function based on distance ratio to find suitable ψ_{uv} parameter.

Test- Given a test instance $x \in$ *Test*

(i) Evaluate distances d_u and d_v from the cluster centres to sample x.
(ii) Estimate classification accuracies for each classifier using ψ_{uv} parameter found on the training data.
(iii) Calculate weights for each classifier j in the ensemble as $b_j = \log \frac{p_{uv}(x)}{1-p_{uv}(x)}$.
(iv) Choose the classifier decision that receives the highest weighted majority vote.

Figure 2. DWEC method. The inputs to the algorithm are the training data and the number of clusters to partition. The training step consists of clustering and training the classifier ensemble. In the testing step a previously unseen instance is presented to the classifier ensemble.

where x denotes a test sample and p_{uv} is the estimated classification accuracy of the classifier made from clusters c_{1u} and c_{2v}. Next, the weights for the jth classifier can be dynamically calculated for each test sample x as, $b_j = \log \frac{p_{uv}(x)}{1-p_{uv}(x)}$. Figure 2 summarizes the steps involved in the DWEC method.

4. Results and discussion

The proposed DWEC method was tested on the publicly available BCI Competition dataset 2A [19] and the ARTS12 dataset collected from 12 healthy subjects. For both datasets single-trial EEG data were extracted for training the CSP algorithm. Three pairs of CSP features for the 8–30 Hz band-pass filtered EEG measurements, extracted at the time segment of 0.5–2.5 s after the onset of the visual cue, were used.

The number of component classifiers in the ensemble depends on the number of clusters as too many clusters will result in smaller partitions leading to over-fitting and lower generalization accuracies for unseen data. Therefore only two to seven clusters, resulting in four to 49 individual classifiers respectively, were investigated.

4.1. Classification accuracies

The proposed DWEC method was evaluated on the dataset 2A of BCI Competition IV and ARTS12 dataset collected from 12 healthy subjects. Six separate ensembles of classifiers were developed consisting of four to forty nine individual classifiers. Their performances were compared against a single SVM classifier. The empirical results for the dataset 2A of BCI Competition IV are shown in table 1.

The highest classification accuracies for each subject are in boldface. A series of pairwise t-tests were carried out between the baseline results and each of the clustering approaches. The minimum conditional entropy criterion was used to find the optimal number of clusters. It considers only the diversity among the partitional clusters. The classification accuracies of the resulting classifiers, trained on a different number of clusters, were analysed. It was observed that the optimal numbers of clusters, seven, does not result in the best classification accuracy. It can be seen that the optimal number of clusters yielded a statistically significant improvement over the baseline result ($p = 0.048$). However, the ensemble of classifiers resulting from three clusters yielded the best overall classification accuracy (81.5%). A t-test between the ensemble built with three clusters and the ensemble built with seven clusters revealed that the two ensemble classifiers are not statistically different ($p = 0.93$). This could be attributed to the over-training of component classifiers and a lack of sufficient training data, as the sample numbers for training are reduced when more clusters are created.

The results obtained for the ARTS12 data set collected from 12 healthy subjects are shown in table 2. The training data were clustered only to three clusters based on the previous results. A pairwise t-test was carried out at a confidence level of 0.05 and the increase over the baseline results obtained with

Table 1. Results of dataset 2A.

		Number of clusters where training data are partitioned					
Subject	Baseline acc.	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	64.1	62.5	63.9	63.4
A3	93.1	96.9	96.8	96.2	96.5	95.2	95.9
A4	63.6	66.7	67.3	66.7	66.8	66.4	65.5
A5	54.8	75.9	75.9	75.6	75.4	75.7	75.6
A6	62.6	64.9	65.2	63.6	65.8	63.8	64.5
A7	77.1	78.1	78.1	77.9	78.1	78.5	78.7
A8	94.2	96.1	96.1	96.4	95.2	95.7	95.6
A9	93.8	92.6	93.2	92.8	93.25	92.8	93.2
Mean	75.9	81.3	81.5	81.0	80.9	80.8	80.9
Std. dev.	16.6	14.3	14.2	14.4	14.1	14.1	14.4
p value		0.039	0.032	0.047	0.047	0.059	0.048

The baseline results produced by a single SVM classifier are compared against ensembles created by combining multiple classifiers trained on clustered training data for the BCI Competition IV dataset 2A. The two sample Student's t-test is used to assess the statistical significance of the improvement at a confidence level of 0.05.

Table 2. Results of data collected from 12 healthy subjects.

Subject	Baseline acc.	Acc. from DWEC with three clusters
1	60.7	65.0
2	62.1	65.2
3	52.7	57.5
4	69.4	70.7
5	67.2	69.3
6	82.2	87.9
7	81.1	84.3
8	95.2	97.5
9	73.0	75.0
10	57.2	61.9
11	49.4	56.6
12	82.7	84.7
Mean	69.4	73.0
t-test (P value)		2.67×10^{-5}

This table compares the baseline accuracy given by a single SVM classifier against the ensemble classifier trained on three clusters of training data for the data collected from twelve healthy subjects. The two sample Student's t-test is used to assess the statistical significance of the improvement at a confidence level of 0.05.

a single SVM classifier was found to be statistically significant ($p = 2.67 \times 10^{-5}$).

4.2. Addressing non-stationarity

The presence of non-stationarity in the session-to-session data can be clearly identified by the clustering analysis. Figure 3 highlights the presence of non-stationarity in dataset 2A. The clusters of the feature space from the second session spanned a different space compared to the clusters of the feature space from the first session. A classifier trained on the first session will not be able to classify the data from subsequent sessions due to the presence of this non-stationarity. The feature space consists of the two best features ($m = 1$ in CSP algorithm) selected after the CSP algorithm.

Figure 4 shows two examples that are correctly classified only by the proposed method. A reduced two-dimensional feature space consisting of the two best features, by setting $m = 1$ in the CSP algorithm, is used for the plot. Three base classifier hyperplanes are shown in the figure by dashed lines. The classifier L_{11} is trained on cluster 1 of class 1 and cluster 1 of class 2. L_{22} is trained on cluster 2 of class 1 and cluster 2 of class 2 and L_{33} is trained on cluster 3 of class 1 and cluster 3 of class 2. The baseline ensemble without dynamic weighting is also shown as a dashed line. The black dots represent features from the second session. Test sample x_1 belongs to class 1, but it is classified wrongly to class 2 by classifiers L_{22} and L_{33}, however L_{11} classifies it correctly and because the decision of L_{11} is magnified by the weighting method, the effective hyperplane of the ensemble for x_1, shown as $EL1$, correctly classifies the sample x_1 in class 1.

Test sample x_2 also belongs to class 1, but it is incorrectly classified to class 2 by classifiers L_{11} and L_{33}, however L_{22} classifies it correctly and because the decision of L_{22} is magnified by the weighting method, the effective hyperplane of the ensemble for x_2 shown as $EL2$ correctly classifies the sample x_2 in class 1.

A further analysis was carried out on the BCI Competition dataset 2A to ascertain whether the proposed DWEC method is capable of accounting for non-stationarity in EEG data. In this study, part of the test data was also included in the training data. The hypothesis, that the clustering based classifier ensemble is capable of accounting for non-stationarity when there is more variability in the data, was statistically analysed for significance. Table 3 summarizes the results of the analysis.

Two baseline cases were considered in the analysis (case 1 and case 2). In the first case, the classifiers were trained with all the training data similar to the standard procedure and evaluated on only half of the randomly chosen test data. In the second set-up (case 2), half of the test data were randomly selected to be incorporated into the training data and were tested on the other half of the test data. The clustering-based ensemble was also trained in a similar manner and tested on a randomly chosen half of the original test samples. Q4

Two statistical tests were carried out to compare the mean results of this study. First, the baseline cases where the

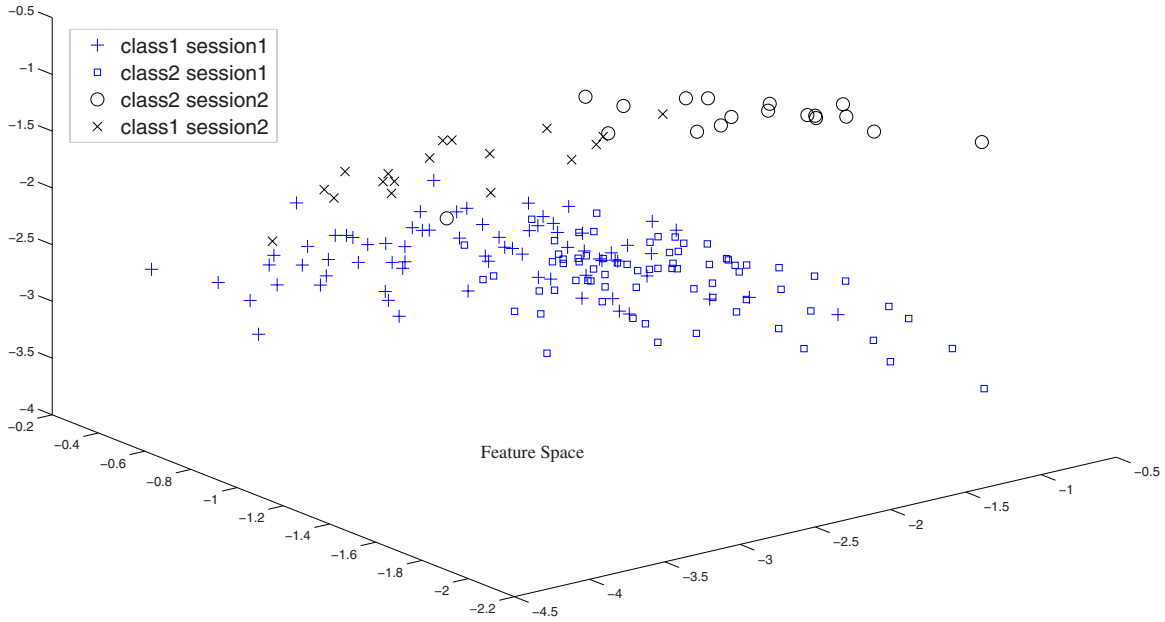

Figure 3. Session-to-session non-stationarity in BCIC IV dataset 2A subject A1.

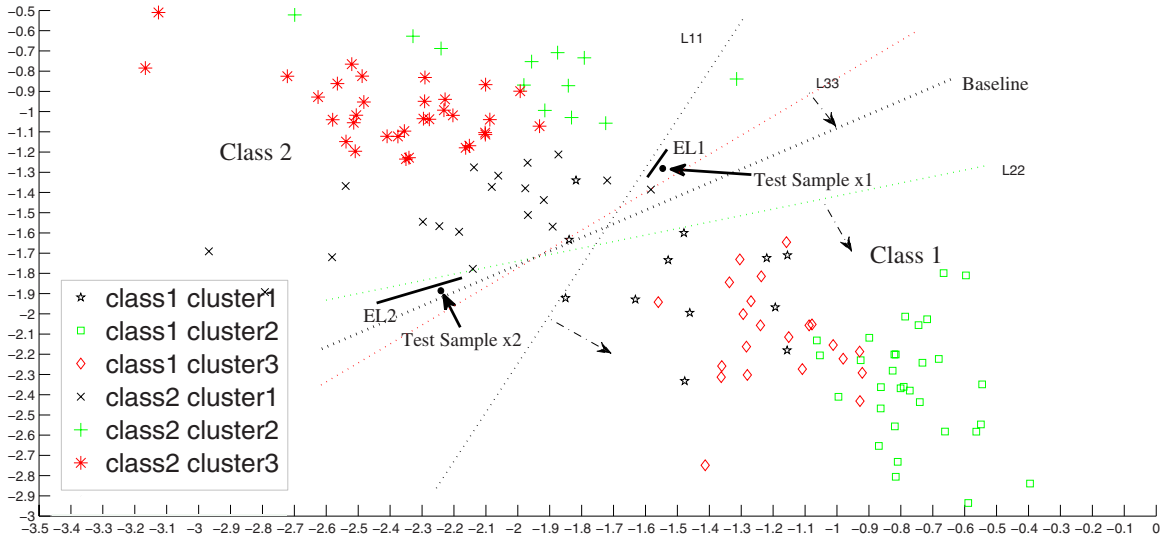

Figure 4. Examples of two test samples from ARTS12 dataset subject 3. Three clusters of each class are combined resulting in nine classifiers. Only three classifier hyperplanes $L11$, $L22$ and $L33$ are shown in the figure. The baseline classifier hyperplane is also shown by a dashed line. The chosen test samples shown as black dots are correctly classified by the proposed method but misclassified by other combination methods. The effective hyperplanes, resulting from dynamic weighting, for each of the test samples are shown as solid lines $EL1$ and $EL2$. The dashed arrows perpendicular to the classifier hyperplanes indicate the direction of class 1 by each classifier.

ensemble of classifiers was not used were compared against the corresponding cases with the ensembles. The probability values of the pairwise t-tests are denoted as $P1$ in table 3. The tests suggest that the proposed DWEC method results in statistically significant improvements over the respective baseline cases under both settings ($P1 = 0.013$ and 0.031).

The second comparison was carried out between the case where half of the test samples were included for training without the proposed classifier combination method, against the case where the classifier ensemble was trained with only the training data and tested on half of the test data. The test indicates that the mean accuracies resulting from the two cases are not different at a 0.05 level of significance ($P2 = 0.068$). Even if our proposed method did not use any data from the test set (case 3), the DWEC method achieved at least equal performance to that of the baseline method with additional data taken on the same day as the test samples (case 2).

4.3. Complexity analysis

The complexity of the proposed framework depends on the complexities of the main components: the CSP algorithm, clustering mechanism, classifier ensemble and optimal weights calculation.

Pre-processing and feature extraction steps depend mostly on the complexity of the CSP algorithm. The CSP algorithm

Table 3. Comparison of effects of including data from the second session.

	Methods			
	Baseline without clustering		DWEC with three clusters	
Subject	Case 1	Case 2	Case 3	Case 4
A1	87.49	90.06	96.17	97.42
A2	56.85	60.24	66.08	68.51
A3	93.25	96.91	97.13	98.47
A4	63.64	64.99	68.72	70.34
A5	55.03	57.09	76.39	78.47
A6	64.75	64.87	68.87	69.17
A7	77.11	78.35	78.82	80.17
A8	94.27	96.34	97.95	98.11
A9	93.92	95.71	95.01	96.59
Mean	76.26	78.51	82.79	84.14
Std Dev	16.47	16.57	13.65	13.41
$P1$	–	–	0.013	0.031
$P2$	–	0.068		–

Case 1 and case 3: train classifiers on all training data and test on half of test data. Case 2: train with half of training data and half of test data and test on the other half of test data. Case 4: train with half of training data and half of test data and test on the other half of test data. $P1$ compares the significance between the baseline cases (case 1 and case 3) against the corresponding approaches with the ensemble built by three clusters (case 2 and case 4). The $P2$ statistic compares the case where half of the test samples were included for training without the proposed classifier combination method (case 2) against the case where classifiers are trained with only the training data and tested on half of the test data (case 3).

needs to compute covariance matrices, which are in the order $O(N * \text{ch}^3)$, where N is the dimensionality of the data and ch is the number of components (channels). In this study, the dimensionality of the data was 6.

The complexity of the clustering algorithm depends on the initialization step and the iterative updates. The initialization step costs $O((n_t)^2 * N)$ as the complete kernel matrix needs to be set up. Finding the best target cluster for each datum costs $O(K)$ time and the update procedure costs $O(n_t)$ time. K is the number of clusters and n_t is the number of data samples. The cost of the main loop of the algorithm is therefore $O(I * n_t(K + \mu * n_t))$ where I is the number of iterations and $0 < \mu < 1$ is the expected ratio of data items that change membership. The number of membership changes is large for the first few iterations, then quickly reduces as the algorithm converges. Overall, the time complexity of clustering is dominated by the quadratic cost of computing the kernel matrix. The maximum number of iterations was set to 50 for increased efficiency.

The complexity of the ensemble depends partly on the number of SVM classifiers and on the SVM classification algorithm. The complexity of one SVM classifier depends on the number of features and support vectors. When a linear kernel is used the time complexity depends only on the feature dimensionality [35]. Therefore, the complexity for one SVM classifier is in $O(N)$, where N is the dimensionality of the data. The complexity of the whole ensemble is $O(N * K^2)$.

The calculation of optimal weights involves $O(K^2)$ distance measures and their optimization. The optimization function is smooth and convex with a complexity of $O(K^2)$. Each gradient computation complexity is also $O(K^2)$, so if all of them have to be computed during an iteration that adds $O(K^4)$. If the total number of iterations for the optimization is I, the complexity of optimization adds up to $O(I * K^4)$.

5. Conclusion

In this study, we proposed a novel method to partition EEG data using clustering, and multiple classifiers were trained using the partitioned datasets. The final decision of the classifier ensemble was then obtained by weighting the classification decisions of individual classifiers. A combination method based on the distances from the test sample to the constituent cluster centres that form the specific classifier was subsequently used to weigh the classifier decisions. The proposed DWEC method was applied on the publicly available dataset 2A from BCI Competition IV and the dataset ARTS12 collected from 12 healthy subjects. The classification accuracies obtained showed that the proposed method yielded statistically significant improvements. The analysis carried out in section 4.2 showed that the proposed DWEC approach can be used to address non-stationarity in EEG data.

References

[1] Wolpaw J R, Birbaumer N, McFarland D J, Pfurtscheller G and Vaughan T M 2002 Brain–computer interfaces for communication and control *Clin. Neurophysiol.* **113** 767–91

[2] Gribkov D and Gribkova V 2000 Learning dynamics from non-stationary time series: analysis of electroencephalograms *Phys. Rev.* E **61** 6538–45

[3] Sykacek P, Roberts S and Stokes M 2004 Adaptive BCI based on variational Bayesian Kalman filtering: an empirical evaluation *IEEE Trans. Biomed. Eng.* **51** 719–29

[4] Guger C, Ramoser H and Pfurtscheller G 2000 Real-time EEG analysis with subject-specific spatial patterns for a brain–computer interface (BCI) *IEEE Trans. Rehabil. Eng.* **8** 447–56

[5] Shenoy P, Krauledat M, Blankertz B, Rao R P and Muller K R 2006 Towards adaptive classification for BCI *J. Neural Eng.* **3** 13–23

[6] Sugiyama M, Krauledat M and Muller K 2007 Covariate shift adaptation by importance weighted cross validation *J. Mach. Learn. Res.* **8** 985–1005

[7] Kaplan A Y, Fingelkurts A A, Borisov S V and Darkhovsky B S 2005 Nonstationary nature of the brain activity as revealed by EEG/MEG: methodological, practical and conceptual challenges *Signal Process.* **85** 2190–212

[8] Pascual J, Vidaurre C and Kawanabe M 2011 Investigating EEG non-stationarities with robust PCA and its application to improve BCI performance *Int. J. Bioelectromagnetism* **13** 50–51

[9] Pascual J, Kawanabe M and Vidaurre C 2011 Modelling non-stationarities in EEG data with robust principal component analysis *Hybrid Artificial Intelligent Systems (Lecture Notes in Computer Science* vol 6679*)* (Berlin: Springer) pp 51–58

[10] Kawanabe M, Samek W, von Bünau P and Meinecke F 2011 An information geometrical view of stationary subspace analysis *Artificial Neural Networks and Machine*

Learning—ICANN 2011 ed T Honkela, W Duch, M Girolami and S Kaski (Berlin: Springer) pp 397–404
[11] Quinonera-Candela J, Sugiyama M, Schwaighofer A and Lawrence N D 2009 *Dataset Shift in Machine Learning* (Cambridge, MA: MIT Press)
[12] Hastie T and Tibshirani R 1996 Discriminant adaptive nearest neighbour classification *J. R. Stat. Soc.* B **58** 155–76
[13] Dawid A P 1992 Applications of a general propagation algorithm for probabilistic expert systems *Stat. Comput.* **2** 25–36
[14] Ghahramani Z and Kim H C 2003 *Bayesian Classifier Combination* (University College London)
[15] Erdogan H and Åžen M U 2010 A unifying framework for learning the linear combiners for classifier ensembles *Proc. 20th Int. Conf. on Pattern Recognition (Istanbul, Turkey)*
[16] Simpson E, Roberts S J, Smith A and Lintott C 2011 Bayesian combination of multiple, imperfect classifiers *Neural Information Processing Systems* (Spain:
[17] Fazli S, Popescu F, Danczy M, Blankertz B, Muller K R and Grozea C 2009 Subject-independent mental state classification in single trials *Neural Netw.* **22** 1305–12
[18] Kuncheva L I 2004 *Combining Pattern Classifiers: Methods and Algorithms* (Hoboken, NJ: Wiley)
[19] Blankertz B 2008 BCI Competition IV Fraunhofer FIRST (IDA) http://ida.first.fraunhofer.de/projects/bci/competition-iv/
[20] Pfurtscheller G, Brunner C, Schlogl A and Lopes da Silva F H 2006 Mu rhythm (de)synchronization and EEG single-trial classification of different motor imagery tasks *NeuroImage* **31** 153–9
[21] Koles Z J and Soong A C K 1998 EEG source localization: implementing the spatio-temporal decomposition approach *Electroencephalogr. Clin. Neurophysiol.* **107** 343–52
[22] Ramoser H, Muller-Gerking J and Pfurtscheller G 2000 Optimal spatial filtering of single trial EEG during imagined hand movement *IEEE Trans. Rehabil. Eng.* **8** 441–6
[23] Dornhege G, Millan J d R, Hinterberger T, McFarland D and Müller K-R (ed) 2007 *Towards Brain-Computer Interfacing* (Cambridge, MA: MIT Press)
[24] Dornhege G, Blankertz B, Krauledat M, Losch F, Curio G and Müller K-R 2006 Combined optimization of spatial and temporal filters for improving brain–computer interfacing *IEEE Trans. Biomed. Eng.* **53** 2274–81
[25] Lemm S, Blankertz B, Curio G and Muller K-R 2005 Spatio-spectral filters for improved classification of single trial EEG *IEEE Trans. Biomed. Eng.* **52** 1541–8
[26] Ang K K, Chin Z Y, Wang C C, Guan C T and Zhang H H 2012 Filter bank common spatial pattern algorithm on BCI competition IV datasets 2A and 2B *Front. Neurosci.* **6**
[27] Krauledat M, Schrder M, Blankertz B and Muller K-R 2007 Reducing calibration time for brain–computer interfaces: a clustering approach *Advances in Neural Information Processing Systems* vol 19 ed B Schlkopf, J Platt and T Hoffman (Cambridge, MA: MIT Press) pp 753–60
[28] Vinh N X, Epps J and Bailey J 2010 Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance *J. Mach. Learn. Res.* **11** 2837–54
[29] Havrda J and Charvat F 1967 Quantification method of classification processes: concept of structural entropy *Kybernetika* pp 30–5
[30] Parzen E 1962 On estimation of a probability density function and mode *Ann. Math. Stat.* **33** 1065–76
[31] Rosenblatt M 1956 Remarks on some nonparametric estimates of a density function *Ann. Math. Stat.* **27** 832–7
[32] Principe J C, Xu D and Fisher J 2000 Information theoretic learning *Unsupervised Adaptive Filtering* ed S Haykin pp 265–321
[33] Lotte F, Congedo M, Lcuyer A, Lamarche F and Arnaldi B 2007 A review of classification algorithms for EEG-based brain–computer interfaces *J. Neural Eng.* **4** R1–13
[34] Dara R A, Makrehchi M and Kamel M S 2010 Filter-based data partitioning for training multiple classifier systems *IEEE Trans. Knowl. Data Eng.* **22**
[35] Burges C J C 1998 A tutorial on support vector machines for pattern recognition *Data Min. Knowl. Discovery* **2** 121–67
[36] Duda R O, Hart P E and Stork D G 2001 *Pattern Classification* 2nd edn (New York: Wiley)