A Combination of Transductive and Inductive Learning for handling Non-Stationarities in Motor Imagery Classification

H. Raza, H. Cecotti, G. Prasad
School of Computing and Intelligent Systems
Ulster University
Londonderry, Northern Ireland, UK

***Abstract*—A major issue for bringing brain-computer interface (BCI) based on electroencephalogram (EEG) recordings outside of laboratories is the non-stationarities of EEG signals. Varying statistical properties of the signals during inter- or intra-session transfers can lead to deteriorated BCI performances over time. These variations may cause the input data distribution to shift when transitioning from the training phase (calibration session) to the testing/operating phase resulting in a covariate shift. We propose to handle this issue using a novel hybrid learning method based on two classifiers, wherein the first classifier allows including new information in the training dataset, and the second classifier performs an overall classification. The proposed method is motivated by the smoothness assumption, i.e., the points that are closest to each other are more likely to share the same label, and may be added online to enrich the training dataset. The method is evaluated on two real-world datasets corresponding to motor imagery detection (BCI competition 2008 dataset 2A and 2B). The results support the conclusion that an improvement in the classification accuracy over traditional inductive learning and semi-supervised learning methods can be obtained.**

I. Introduction

A Brain-Computer Interface (BCI) is an alternative communication's means, which allows a user to express his will without muscle exertion, provided that the brain signals are properly translated into computer commands [1], [2], [3]. With an electroencephalography (EEG) based BCI that operates online in real-time non-stationary/changing environments, it is required to consider input features that are invariant to shifts of the data, or learning approaches that can be able to track the changes that can repeat overtime, to update the classifier in a timely fashion. In fact, it may be difficult to reliably classify the EEG patterns in BCI during long sessions using traditional inductive classification algorithms due to the non-stationarity characteristics of the EEG signal. The non-stationarities in the EEG may be caused by various reasons such as changing user attention level, electrode placement, and user fatigue [4], [5], [6]. There are therefore notable variations or shifts in the EEG signals during trial-to-trial, and session-to-session transfers [7]. These variations often appear as covariate shifts, wherein the input data distributions differ significantly between training/calibration and testing/operating phases, while the conditional distribution remains the same [8], [9].

The low classification accuracy has been one of the main concerns of the developed BCI systems based on motor imagery (MI) detection, which directly affects the reliability of the BCI decision making process and the information transfer rate. The traditional classification algorithms are mainly inductive. They deal with the development of a function that approximates output values using input data from the whole searching space (i.e. induction), and then uses this function to predict the output values for all the new input vectors (i.e. deduction). To enhance BCI performance, several feature extraction, feature selection, and feature classification techniques have been proposed in the literature [10], [11], [12], [13]. A large variety of features have been used in BCI based on motor imagery detection such as band powers, power spectral density, time frequency features, and common spatial patterns (CSP) based features [14]. However, the EEG non-stationary characteristics result in shifts in feature distributions for all types of features in varying degrees. An example showing the change of distribution between the training and test datasets is depicted in Figure 1.

To solve this issue, adaptive learning algorithms have been proposed for devising adaptive BCI systems with positive results [4], [15], [16]. Most of which have made efforts to reduce the non-stationarity aspect in the extracted features. In an adaptive learning technique, a priori information is required about the changes in the EEG signals. Additionally, the adaptive techniques are mostly based on supervised learning techniques, which need labeled data (i.e. a training dataset that can be obtained through a calibration session [7], [17]), which may not be available or difficult to acquire during the operating phase. Different strategies are possible, first the same classifier can be used over time, and the classifier is updated only when a significant shift is detected [18], [19]. Second, it can be assumed that the signal will always change over time in a continuous way, and not severe shift can be detected. In such a case, the classifier will be updated with all the new trials that can be correctly detected, and the updated training database will follow the data distribution. The later approach is chosen in the present study. It is also motivated by the time between two trials in motor imagery detection, which can be several seconds, and allows a classifier to be fully retrained thanks to the addition of new knowledge.

To overcome the issues discussed above, a transductive-inductive learning approach based on two classifiers is presented in this paper. The idea of the proposed learning algorithm is to handle the covariate shift by initiating the adaptation based on both the existing (training dataset) and new knowledge obtained through the testing phase. The transductive classifier is only used for adding new information in the existing training dataset, and the inductive classifier is used for predicting the BCI outputs, after being retrained each time the training dataset is updated. It is thus a learning approach wherein transductive and inductive learning are combined to update the training dataset, and to track the evolution of the features over time. The first classifier is a probabilistic weighted K-nearest neighbor (PWKNN) based transductive classifier. The output of the first classifier is used to determine if a trial and its corresponding estimated label can be added to the training dataset, and the inductive learning model is updated. Transductive learning combines induction and deduction in a single step, and is related to the field of semi-supervised learning (SSL), which uses both labeled and unlabeled data during learning [20]. By eliminating the need to construct a global model, transductive learning offers prospect to achieve higher accuracy. In order to make use of unlabeled data, it is necessary to assume some structure to the underlying distribution of data. Additionally, it is essential that the SSL approach must satisfy at least one of the following assumptions such as smoothness, cluster, or manifold assumption [21], [22], [23]. In SSL, the widely used type of algorithm is graph based label propagation, which propagates labels from labeled examples to the whole unlabeled data. In the present paper, we will make use of the smoothness assumption (i.e., the points which are closest to each other are more likely to share the same label) to implement a transductive learning algorithm. The second classifier is inductive, a linear support vector machine (SVM) classifier, its outputs are used to determine the BCI outputs. In this paper, we demonstrate the effectiveness of the approach on a synthetic dataset and on two real-world datasets, and we show that the covariate shifts can be tracked and adapted using the proposed method. Specifically, using the data from the BCI competition-IV 2A and 2B, we demonstrate the effectiveness of the proposed approach over traditional, adaptive, and semi-supervised learning algorithms. The contributions of the paper can be summarized as follows: First, a combination of a transductive and inductive classifier are used to track, and address the effect of covariate shifts in the non-stationary EEG signals. Second, the proposed learning strategy updates automatically the classifier online without making any a priori assumption about the distribution for the upcoming test data. The remainder of the paper is organized as follows: section II presents the system overview of the transductive-inductive learning method, including the probabilistic K-nearest neighbor. The datasets and the signal processing steps are described in section IV. The results are then presented in section V and discussed in section VI.

II. System overview

A. Problem Statement

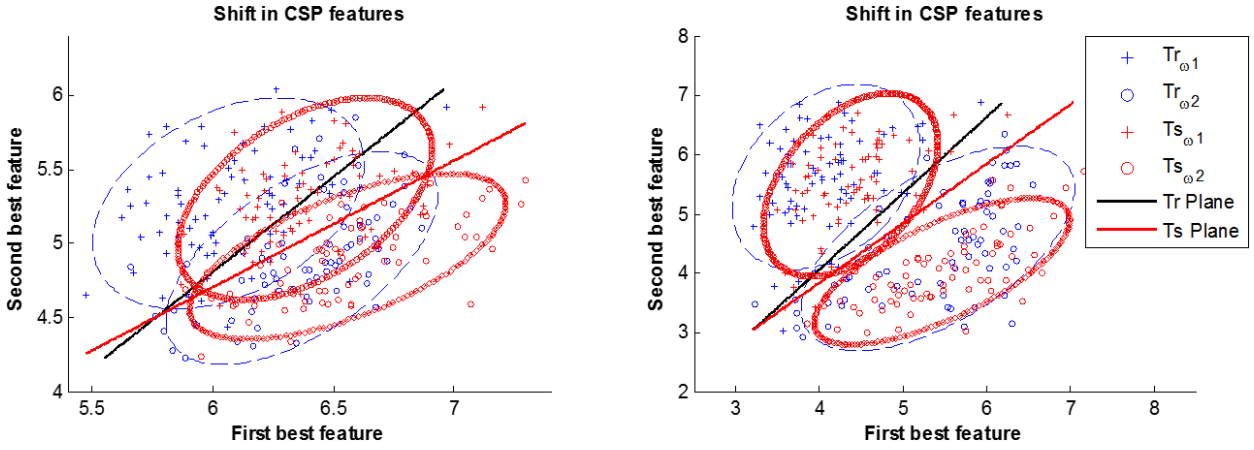
Let us consider a learning framework in which a training dataset is denoted by $X_{Tr} = \{(x_i, y_i)\}_{i=1}^{N}$, where N is the total number of observations, and a label y_i is associated with each input x_i. Depending upon the number of inputs and outputs, x_i and y_i may be a scalar or vector variables. Let us consider a two-class classification problem i.e., $y \in \omega_1, \omega_2$. The probability distribution of the inputs at time i can thus be defined as, $P(x_i) = P(\omega_1)P(x_i|\omega_1) + P(\omega_2)P(x_i|\omega_2)$ where $P(\omega_1)$ and $P(\omega_2)$ are the prior probabilities of getting a sample of the classes ω_1 and ω_2, respectively, while $P(x_i|\omega_1)$ and $P(x_i|\omega_2)$ are the conditional probability distribution for the time period i. The goal is to predict the labels $\hat{y_i}$ of upcoming samples resulting into $X_{Ts} = \{(\hat{y_i}|x_i)\}_{i=1}^{M}$, where M is the total number of observations in the testing phase.

B. Probabilistic K-Nearest Neighbor

Probability theory plays a vital role in the solution of pattern recognition problems, because most of the problems can be solved using a density estimation technique [24]. Its main task is to model a probability density function $P(x)$ of a random variable X, given X_{Tr}. There are two approaches for the density estimation, namely, parametric and non-parametric. One of the key limitations of the parametric approach is that it assumes a precise practical form for the distribution, which may lead to be incompatible for a specific application. An alternative approach is the non-parametric density estimation, which estimates density function without applying any assumption about underlying data distribution. Here, we consider a non-parametric method based on K-nearest-neighbors (KNN) because it is a transductive learning method that uses the test data point to determine a decision [25]. In a KNN method, we consider a small sphere centered at the point x at which, we wish to estimate the density $P(x)$. We allow the radius of the sphere to grow until it contains K data points, the estimate of the density is then given by:

$$P(x) \quad = \quad K/(N \cdot V) \tag{1}$$

where, V is set to the volume of the sphere, and N is the total number of points. The parameter K governs the degree of smoothing. The technique of KNN density estimation may be extended to the classification task in which the KNN density estimation is obtained for each class, and then the Bayes theorem is used to perform a classification task. Now, let's suppose that we have a data set comprising N_{ω_i} points in the class ω_i within the set of classes ω, where $i \in 1, 2$, so that $\sum_{\omega_i} N_{\omega_i} = N$. If we wish to classify a new point x, we draw a sphere centered on x containing precisely K points irrespective of their classes. Now suppose this sphere has the volume V, and contains $K_(\omega_i)$ from class ω_i. Then, an estimate of the density associated with each class or



(a) [8-12] Hz

(b) [14-30] Hz

Fig. 1. Covariate shift in the EEG dataset 2A subject A03, between training and testing input distribution for different frequency bands. (a) Mu band [8-12] Hz, and (b) Beta band [14-30]Hz. The red circle denote the features of the left hand motor imagery and blue crosses denote the features of the right hand motor imagery. The black and red lines represent the decision boundaries obtained by the training data and test data respectively.

likelihood can be obtained by:

$$P(x\omega_i) = \frac{K_{\omega_i}}{N_i \cdot V} \quad (2)$$

Similarly, the unconditional density is given by $P(x) = K/(N \cdot V)$. The class prior probability is given by:

$$P(\omega_i) = N_i/N \quad (3)$$

Using the Bayes theorem, we can obtain the posterior probability of the class membership:

$$P(\omega_i|x) = \frac{P(x|\omega_i)P(\omega_i)}{P(x)} = \frac{K_{\omega_i}}{K} \quad (4)$$

If we wish to minimize the probability of misclassification, this is achieved by assigning the test point x to the class ω_i having the largest posterior probability, i.e. corresponding to the largest value of K_{ω_i}/K. Thus, to classify a new point, identify the K-nearest points from the training data set, and then assign the new point to the set having the largest number of representatives. This posterior probability is also known as the Bayesian belief or confidence ratio (CR). However, the overall estimate obtained by the KNN method may not be satisfactory, because the resulting density is not a true probability density since its integral over all the samples space diverges [26]. Another drawback is that it considers only the K points to build the density, and each neighbor has an equal weight. An extension to the above KNN method is to assign the weight to each sample that depends on the distance to x [27]. A radial basis function (RBF) kernel is used to obtain the weights. Using RBF Kernel, the nearest points have weights with higher value than furthest points. A probabilistic weighted K-nearest neighbors (PWKNN) approach based on an RBF kernel is thus proposed to devise the transductive classifier with $RBF_{(p,q)}$:

$$RBF_{(p,q)} = exp(-\frac{d^2_{(p,q)}}{2\sigma^2}) \quad (5)$$

where $d_{(p,q)}$ is the Euclidean distance from the unlabeled data point x_p to the labeled data point x_q is computed as given below:

$$d_{(p,q)} = \sqrt{\sum_{i=1}^{m}(x_p(i) - x_q(i))^2} \quad (6)$$

and $x(i)$ is the i-th feature of x and m is the number of features.

For binary detection, the confidence ratio of CR_{ω_i} of the class ω_i, for a data point x_p, is defined by:

$$CR_{\omega_1} = \frac{\sum_{j=1}^{K} RBF_{(p,j)} \cdot (l_j == \omega_1)}{\sum_{j=1}^{K} RBF_{(p,j)}} \quad (7)$$

$$CR_{\omega_2} = 1 - CR_{\omega_1} \quad (8)$$

where x_j, $1 \leq j \leq K$, corresponds to the j-th nearest neighbor of x_p. The outputs of PWKNN include the overall confidence of the decision, given by:

$$CR = max(CR_{\omega_1}, CR_{\omega_2}) \quad (9)$$

and the output class $\check{y}$ (1 if x_p is assigned to ω_1, 0 otherwise).

III. Classifier Combination

In the proposed system, a transductive classifier is used to determine if an example should be used or not to enrich a training dataset, which is used by an inductive classifier to determine the final output. Transductive learning was proposed and briefly studied more than 40 years ago by Vapnik and Chervonenk [25]. Later, it has been empirically acknowledged that transduction can often serve as a more efficient, or accurate learning method than traditional inductive learning approaches [22]. This appreciation is the rationale behind the development of the transductive learning model

for accounting non-stationarities in BCIs. Moreover, transductive learning methods estimate the value of a potential model (classification function) at a new data point using specific training cases related to that test point [25]. Using an SSL approach, every unlabeled point gives information about $P(x)$. The usefulness of an example depends on the condition whether the density of the input distribution has changed significantly, or the point is located very close to the classification boundary. For example, in SSL, the smoothness assumption states that the points closest to each other are more likely to share the same label. Hence, if the point satisfies the smoothness assumption, then a single observation provides some useful information to update the training data-set. Then, using the updated training data-set, an improved global model can be obtained to classify the new unlabeled data points.

In this paper, we propose a novel transductive-inductive learning model. The proposed approach consists of two steps: induction and transduction. Initially, an inductive classifier, denoted by $\mathcal{F}$, is trained on the features obtained from the calibration/training data X_{Tr}. Once the classifier $\mathcal{F}$ is trained, and a classification decision boundary is obtained, then the evaluation phase starts. In this phase, we set the parameters CR_α, K, and Δm, where CR_α is a confidence ratio threshold, K is the number of neighbors for preparing a transductive classifier (denoted by $\mathcal{T}$) through PWKNN algorithm, and Δm is the number of data points after which the adaptation starts. In addition, the classifier F will classify the input features obtained from the testing data X_{Ts} in the evaluation phase. The classifier $\mathcal{F}$ will initiate adaptation through transductive learning after every Δm consecutive trials (in this study, we have fixed $\Delta m = 5$). Considering a small number of trials Δm in each epoch, results in focusing on a trial-by-trial shift, which may be caused by various reasons such as muscular artifacts. Each time the classifier $\mathcal{F}$ initiates adaptation, it is considered as one epoch, and it assigns Δm data points from X_{Ts} to a variable ΔX_{Ts}, and predicts the labels through the transductive function $\mathcal{T}$. In this step, each observation is taken from ΔX_{Ts}, and the transductive function $\mathcal{T}$ is applied. Once all the ΔX_{Ts} points are processed through $\mathcal{T}$, the CR value is used for each trial to decide if the trial's features and the corresponding estimated output should be added to X_{Tr}, i.e. if $CR > CR_\alpha$. Thus, the labels $\mathcal{T}(\Delta X_{Ts})$ obtained through $\mathcal{T}$, which are above CR_α are inserted into the training dataset. Based on the updated training dataset, the inductive function $\mathcal{F}$ is updated (i.e. the classifier is retrained), and a new classifier is obtained. This process is repeated until all the M points in the testing phase are classified. The pseudo-code of the algorithm is given in the Algorithm 1.

Algorithm 1 Online training and evaluation.

```
Train F with X_Tr
Set the parameters CR_α, K, Δm
for each trial x_i from X_Ts do
    ŷ_i = F(x_i)
    if i mod Δm==0 then
        ΔX_Ts = X_Ts((i − m + 1) : i)
        for each trial x_j from ΔX_Ts do
            {T = PWKNN}
            (CR, ỹ) = PWKNN(X_Tr, x_j, CR_α, K)
            if CR > CR_α then
                Updated X_Tr = X_Tr + (x_j, ỹ)
        Train F with X_Tr
```

IV. Experimental protocol

A. Datasets

The BCI Competition IV dataset 2A is comprised of EEG data collected from nine subjects that were recorded during two sessions on separate days for each subject [28]. The data consists of 25 channels, and includes 22 mono-polar EEG channels, and 3 mono-polar EOG channels with a sampling frequency of 250 Hz. Among the 22 EEG channels, 10 channels are selected for this study, which are responsible for capturing most of the MI related activations: C3, FC3, CP3, C5, C1, C4, FC4, CP4, C2, and C6. Each session consists of six runs separated by short breaks, each run comprised of 48 trials of 7.5 seconds (12 for each class), resulting in a total of 288 trials. Only the classes corresponding to left hand and right hand were considered in the present study. The MI data from the session I was used to train the classifiers, and the MI data from the session-II was used for evaluation purposes.

The BCI competition 2008-Graz dataset 2B is a dataset consisting of EEG data from nine subjects [28]. Three bipolar channel recordings (C3, Cz, and C4) were acquired with a sampling frequency of 250 Hz. All signals were recorded mono-polarly with the left mastoid serving as reference and the right mastoid as ground. For each subject, data corresponding to five sessions are provided, with trials of 8 seconds. The MI data from session I and II (240 trials) were used to train the classifiers, data from the session-III (160 trials) was used to determine the hyper-parameters (i.e., K and CR_α), and the data from sessions IV and V (320 trials) were used to evaluate classifier performance.

B. Data Processing and Feature Extraction

The first stage of signal processing employs a filter bank that decomposes the EEG signals into multiple frequency bands [29]. A total of 10 band-pass filters are used, namely [8-12], [10-14], [12-16], [14-18], [16-20], [18-22], [20-24], [22-26], [24-28], [26-30] Hz [30]. This set of frequency bands is used because it covers the expected frequency range of motor imagery response. In the next sections, we consider a time segment of 3 s after the cue onsets for both data sets (BCI Competition IV dataset 2A and dataset 2B).

The second stage employs common spatial filters (CSP), a set of spatial filters that maximize the variance of spatially filtered signals under one condition, while minimizing it for the other condition. Raw EEG scalp potentials are known to have poor spatial resolution due to volume conduction. If the signal of interest is weak while other sources produce strong

signals in the same frequency range, then it is difficult to classify two classes of EEG measurements [14].

C. Evaluation

In order to evaluate the performance of the system, we have considered the classification accuracy as the measure of performance index. A Cohens kappa coefficient (kappa value) is used to compare the performance with other competing methods. The experiments are performed using a linear SVM pattern classifier ($\mathcal{F}$) and PWKNN classifier ($\mathcal{T}$) for transductive learning. The accuracy is given in percentage (%). The hyper-parameters K and CR_α needs to be carefully selected in order to limit the number of wrongly labeled elements that are added in the training database. Two variants of the proposed learning method, namely TI_1 and TI_2, are presented. In TI_1, the hyper-parameters are selected based on grid search to maximize the mean accuracy across subjects, with $K \in \{6, 12, 18\}$, and $CR_\alpha \in \{0.5, .., 1\}$. In TI_2, the hyper-parameters are determined for each subject, they are selected to maximize the accuracy of each subject. In the dataset 2A, the session I is divided into two parts, 80% for training the classifier and 20% is used to determine the hyper-parameters. The evaluation is then performed on the data from the session II. In the dataset 2B, the sessions I and II are used for training $\mathcal{F}$, session III is used to obtain the hyper-parameters, and sessions IV and V are used to evaluate the performance. For each dataset, the accuracy corresponding to a 10-fold-cross validation (10-CV) on the training dataset is presented. Additionally, two variants of the proposed method are presented and compared with the state-of-the-art methods. The baseline method uses a filter-bank with traditional inductive learning [29]. The baseline method does not adapt/re-train $\mathcal{F}$. It only obtains its global classification function once, and remains fixed during the evaluation phase. Moreover, to compare with other methods, variants of the SSL label propagation methods have been considered [31].

- SSL_1: It is a graph based SSL label propagation approach using the whole test dataset [31].
- SSL_2: the labels are propagated after every Δm trials, and the label prediction updates are continued until all the unlabeled data are assigned with the labels.
- SSL_3: the unlabeled test data are predicted using SSL_2, then $\mathcal{F}$ is trained on both the training data and the predicted test data using SSL_2, and the test data are classified using $\mathcal{F}$.

The classification accuracy obtained by training $\mathcal{F}$ on both the train and the test data, with the evaluation performed on the test data, is denoted by UB. Finally, we evaluate the performance by using Kappa value with other existing methods: the standard Common Spatial Pattern [32], Filter bank CSP for divide and conquer method [29], and recurrent quantum neural networks based EEG filtering for BCI [33].

V. Results

The features obtained with CSP are depicted for subject A03 in Fig 2. Each of sub-figures (a)-(j), represents a CSP feature corresponding to a frequency band. The blue crosses and red circles denote the features of the left hand and right hand MI, respectively. The black line represents a possible linear separation between the features of the two classes obtained from each frequency band. The hyperplan is plotted for an illustration purpose only.

The performance results of the proposed learning algorithm are compared against the aforementioned methods, on the BCI Competition IV dataset 2A and dataset 2B. The accuracy for each subject, with the mean and standard deviation (std), is presented in Tables II and III. For the dataset 2A, with the baseline method, the accuracy is 77.78 ± 14.63 across subjects. The three variations of SSL provides an accuracy of 75.54 ± 12.99, 75.16 ± 15.03, and 75.95 ± 14.73 for SSL_1, SSL_2, and SSL_3. The proposed method provides an accuracy of 78.01 ± 12.86 for TI_1, and 79.17 ± 13.3 for TI_2. For the dataset 2B, the accuracy is 65.46 ± 25.25 with the baseline method, 65.59 ± 13.63, 65.59 ± 13.63, and 67.38 ± 13.75 for SSL_1, SSL_2, and SSL_3. Finally, TI_1 and TI_2 provide an accuracy of 69.56 ± 26.4 and 70.85 ± 25.52, respectively.

A two-sided Wilcoxon signed rank test is used to assess the statistical significance of the improvement at a confidence level of 0.05 in all the pairwise comparisons. The values of hyper-parameters CR_α and K are presented in Fig. 3. For the dataset 2A, they are set as $CR_\alpha = 0.95$ and $K = 12$. For the dataset 2B, they are set as $CR_\alpha = 0.95$ and $K = 18$. For the dataset 2A, there is an increment of 0.23% in the average accuracy for the TI_1 method over the baseline method. Additionally, the proposed method has shown an improvement of 2.06% in average accuracy in comparison with SSL based label propagation methods. Moreover, the improvement over the baseline method is 4.6% for dataset 2B, but it is not statistically significant (p=0.0938). The values of CR_α and K are given in Tables II and III for the dataset 2A and dataset 2B. The average accuracy for TI_2 is superior to the accuracy of the baseline method and the other reported methods for both the datasets 2A and 2B. However, for the dataset 2A, the improvement is only 1.39%, and it is not significant. However, there is statistically significant improvement of 5.39% ($p = 0.0391$) for the dataset 2B.

The effectiveness of proposed method for dataset 2A is presented in Table I, by comparing its classification results in terms of Kappa values. The TI_1 and TI_2 methods have better mean Kappa values over other competing methods. In addition, there is no statistically significant difference between TI_1 and TI_2 for both the databases 2A ($p = 0.25$) and 2B ($p = 0.21$). Finally, the average accuracy obtained with UB is only 84.57% and 79.02% for the dataset 2A and 2B, respectively. It shows that the use of the test datasets with the expected ground truth does not allow to obtain an accuracy above 90% but the variability across sessions leads to a drop of 6.79% for 2A, and 13.56% for 2B.

VI. Discussion

The proposed learning method for EEG-based BCI is based on Vapnik's principle of transduction i.e., given a test

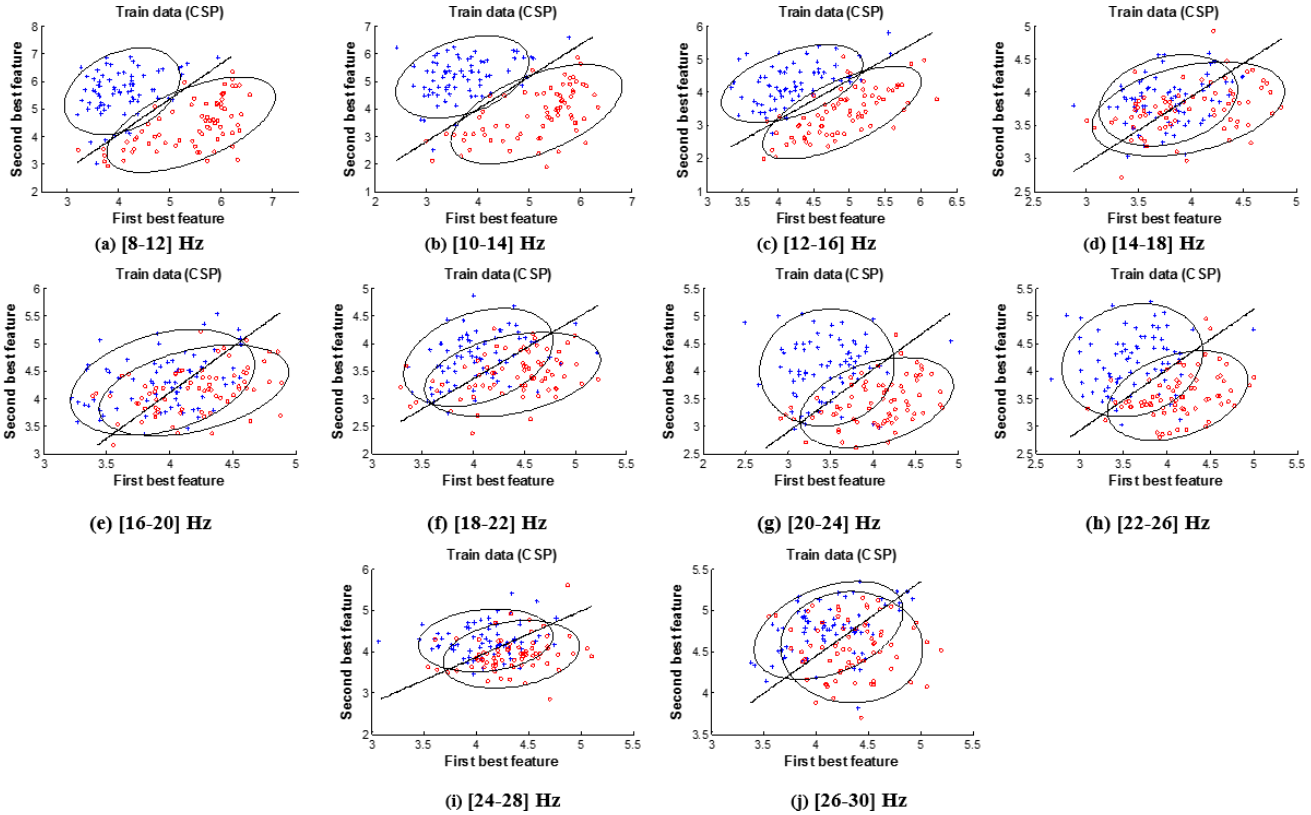


Fig. 2. Distribution of the two best features obtained by CSP for Subject A03. The plots (a-j) represent the CSP features for each band. The red circles denote the features of the left hand MI and blue crosses denote the features of the right hand MI. The black line represents the separation plane for illustration purpose only.

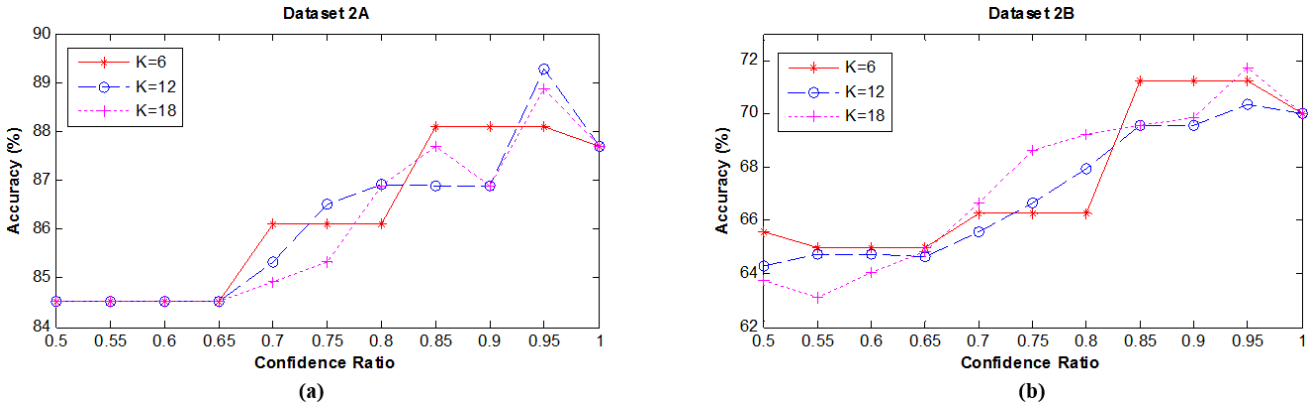


Fig. 3. Accuracy obtained for the different values of CR_α and K. The lines correspond to the mean accuracy across the 9 subjects.

point, one should focus on the training points which are in a neighborhood of this test point. A PWKNN based on an RBF kernel is used to construct a local decision rule, and predict the label of the test point according to this transduction rule. The proposed approach has satisfied the smoothness assumption: If two instances are close to each other in a high-density region; they are likely to share the same label. The new information/knowledge obtained through transduction is used to update the training dataset of the inductive classifier. The main classification function is still inductive because the transductive knowledge is only used to add more information into training dataset.

A certain minimum value of confidence ratio (CR_α) is used to decide whether the information is useful or not to enrich the classifier. If it is the case then it is added to the training dataset. The discarded information may come from artifacts or an abrupt change in the input distribution. The values of CR_α and K needs to be carefully chosen in order to achieve better performance. With subject dependent hyper-parameters values, the proposed TI_2 method shows a statistically significant improvement over the baseline method ($p < 0.05$).

TABLE II
CLASSIFICATION ACCURACY (IN %) RESULTS FROM BCI COMPETITION IV DATASET 2A.

	10-CV	Baseline	SSL_1	SSL_2	SSL_3	TI_1	TI_2			UB
Subject	Training	Eval	Eval	Eval	Eval	Eval	K	CR_α	Eval	Eval
A01	87.86	90.97	86.81	86.43	87.86	93.75	6	0.85	93.75	97.22
A02	82.14	61.11	56.94	55.71	55.00	61.11	18	0.85	61.11	63.89
A03	92.86	90.97	90.97	90.71	90.71	90.97	12	0.8	93.75	99.31
A04	87.86	69.44	72.92	72.86	75.00	69.44	6	1	69.44	76.39
A05	89.29	70.83	66.67	67.14	69.29	70.14	6	0.55	76.39	76.39
A06	85.71	65.28	67.36	66.43	65.71	65.28	6	0.85	66.67	70.14
A07	90.00	68.75	56.25	55.00	57.86	68.75	12	0.95	68.75	88.19
A08	96.43	92.36	91.67	92.14	92.14	92.36	12	0.95	92.36	96.53
A09	75.71	90.28	90.28	90.00	90.00	90.28	6	0.85	90.28	93.06
Mean	87.54	77.78	75.54	75.16	75.95	78.01			79.17	84.57
Std	6.01	14.63	12.99	15.03	14.73	12.86			13.3	13.11

TABLE III
CLASSIFICATION ACCURACY (IN %) RESULTS FROM BCI COMPETITION IV DATASET 2B.

	10-CV	Baseline	SSL_1	SSL_2	SSL_3	TI_1	TI_2			UB
Subject	Training	Eval	Eval	Eval	Eval	Eval	K	CR_α	Eval	Eval
B01	71.67	63.13	66.56	66.56	66.56	63.75	6	1	63.75	70.31
B02	60.42	50.83	50.00	50.00	50.83	62.08	12	0.95	62.08	61.25
B03	62.92	48.13	49.69	49.69	49.38	53.13	12	0.95	54.38	59.69
B04	88.85	90.63	85.31	85.31	87.81	90.63	6	0.85	93.75	97.19
B05	85.38	61.88	61.25	61.25	63.44	61.88	18	0.75	56.25	90.94
B06	76.67	76.56	71.88	71.88	75.94	89.69	18	0.95	89.69	85.63
B07	65.00	66.56	70.00	70.00	71.56	64.06	12	1	66.56	78.44
B08	58.93	65.94	61.56	61.56	64.69	71.25	18	0.85	80.31	88.75
B09	67.50	71.88	74.06	74.06	76.25	71.88	6	0.85	79.06	78.44
Mean	70.81	65.46	65.59	65.59	67.38	69.56			70.85	79.02
Std	10.78	25.25	13.63	13.63	13.75	26.4			25.52	28.43

TABLE I
COMPARISON OVER TYPES OF EEG-BASED CLASSIFICATION METHODS ON KAPPA VALUES FROM BCI COMPETITION IV DATASET 2A.

Subjects	CSP[32]	FBCSP (DC)[34]	RQNN[33]	TI_1	TI_2
A01	0.56	0.71	0.22	0.88	0.88
A02	0.31	0.37	0.22	0.22	0.22
A03	0.70	0.66	0.58	0.76	0.88
A04	0.44	0.47	0.21	0.39	0.39
A05	0.22	0.41	0.43	0.43	0.53
A06	0.20	0.26	0.22	0.33	0.33
A07	0.61	0.73	0.17	0.38	0.38
A08	0.76	0.58	0.35	0.85	0.85
A09	0.72	0.50	0.58	0.81	0.81
Mean	0.50	0.52	0.33	0.56	0.58
Std	0.22	0.16	0.16	0.26	0.27

Another important issue is the number of trials after which an adaptation is initiated. Considering a small number of trials in each epoch results in focusing on trial-by-trial shift, which can be due to muscular artifacts. However, the long term non-stationarities may be accounted for by considering a large number of trials in each epoch. We chose a small number of trials for updating the classifier, as our aim was to track the covariate shift trial-by-trial. The proposed transductive-inductive learning technique makes use of on-line data to extract features, and thus adapts to non-stationarities in the streaming EEG. The experimental results demonstrated the effectiveness of the proposed learning strategy, showing better performance than traditional learning methods.

VII. CONCLUSION

In this paper, we have proposed a hybrid classifier combination using transductive and inductive classifiers to address the effect of covariate shifts in non-stationary EEG signals associated with motor imagery detection in the EEG signal. The labels from the unlabeled data are estimated with a probabilistic weighted K-nearest neighbor approach. The proposed TI_2 method that has subject dependent hyper-parameters provides a statistically significant superior classification accuracy over a purely inductive baseline method on the dataset-2B. Particularly this learning approach provides a foundation for combining the transductive learning and inductive learning for EEG based BCI. Further work will include the detection of the covariate shift in the data, and only retrain the inductive classifier when a relevant shift is detected. This may reduce the processing time and complexity in re-training the classifier in the absence of any real change in the data.

VIII. ACKNOWLEDGMENTS

H.R. was supported by Ulster University, Vice-Chancellor's Research Scholarship (VCRS). G.P. and H.C. were supported by the Northern Ireland Functional Brain Mapping Facility project (1303/101154803), funded by InvestNI and the Ulster University. G.P. and H.R. were also supported by the UKIERI DST Thematic Partnership project "A BCI operated hand exoskeleton based neuro-rehabilitation system" (UKIERI-DST-2013-14/126).

REFERENCES

[1] M. Thulasidas, C. Guan, S. Member, J. Wu, and A. P. Speller, "Robust classification of eeg signal for brain computer interface," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 14, no. 1, pp. 24–29, 2006.

[2] 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. 767–91, Jun. 2002.

[3] J. d. R. Millán, R. Rupp, G. R. Müller-Putz, R. Murray-Smith, C. Giugliemma, M. Tangermann, C. Vidaurre, F. Cincotti, A. Kübler, R. Leeb, C. Neuper, K.-R. Müller, and D. Mattia, "Combining brain-computer interfaces and assistive technologies: state-of-the-art and challenges," *Frontiers in Neuroscience*, vol. 4, no. 161, pp. 1–15, 2010.

[4] Y. Li, H. Kambara, Y. Koike, and M. Sugiyama, "Application of covariate shift adaptation techniques in brain-computer interfaces," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 6, pp. 1318–1324, Jun. 2010.

[5] H. Raza, G. Prasad, and Y. Li, "Ewma model based shift-detection methods for detecting covariate shifts in non-stationary environments," *Pattern Recognition*, vol. 48, no. 3, pp. 659–669, Aug. 2015.

[6] J. R. Wolpaw, "Brain-computer interface research comes of age: Traditional assumptions meet emerging realities," *Journal of Motor Behavior*, vol. 42, no. 6, pp. 351–353, 2010.

[7] M. Arvaneh, C. Guan, and C. Quek, "Eeg data space adaptation to reduce intersession non-stationary in brain-computer interface," *J. Neural Comput.*, vol. 25, pp. 1–26, 2013.

[8] A. Satti, C. Guan, D. Coyle, and G. Prasad, "A covariate shift minimization method to alleviate non-stationarity effects for an adaptive brain-computer interface," in *Proc of the Int. Conf. on Pattern Recognition*, 2010, pp. 105–108.

[9] M. Sugiyama, M. Krauledat, and M. K. R., "Covariate shift adaptation by importance weighted cross validation," *J. Mach. Learn. Res.*, vol. 8, pp. 985–1005, 2007.

[10] M. Arvaneh, C. Guan, K. K. Ang, and C. Quek, "Optimizing spatial filters by minimizing within-class dissimilarities in electroencephalogram-based brain-computer interface," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 24, no. 4, pp. 610–619, 2013.

[11] S. Shahid and G. Prasad, "Bispectrum-based feature extraction technique for devising a practical brain-computer interface," *J. Neural Eng.*, vol. 8, no. 2, p. 025014, Apr. 2011.

[12] H. Il Suk and S. W. Lee, "A novel bayesian framework for discriminative feature extraction in brain-computer interfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 2, pp. 286–299, 2013.

[13] D. Coyle, G. Prasad, and T. M. McGinnity, "Faster self-organizing fuzzy neural network training and a hyperparameter analysis for a brain-computer interface," *IEEE Trans. Syst. Man, Cybern.*, vol. 39, no. 6, pp. 1458–71, Dec. 2009.

[14] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K. R. Müller, "Optimizing spatial filters for robust eeg single-trial analysis," *IEEE Signal Proc. Mag.*, vol. 25, no. 1, pp. 41–56, 2008.

[15] A. Llera, V. Gómez, and H. J. Kappen, "Adaptive multiclass classification for brain computer interfaces," *Neural computation*, vol. 26, no. 6, pp. 1108–27, 2014.

[16] P. Shenoy, M. Krauledat, B. Blankertz, R. P. N. Rao, and K.-R. Müller, "Towards adaptive classification for bci," *J. Neural Eng.*, vol. 3, no. 1, pp. R13–R23, Mar. 2006.

[17] M. A. Oskoei, J. Q. Gan, and O. Hu, "Adaptive schemes applied to online SVM for BCI data classification," in *Proc. of the 31st Annual Int. Conf. of the IEEE Eng. in Medicine and Biology Soc.*, 2009, pp. 2600–2603.

[18] H. Raza, H. Cecotti, Y. Li, and G. Prasad, "Learning with covariate shift-detection and adaptation in non-stationary environments: Application to brain-computer interface," in *Proc. of the Int. Joint Conf. on Neural Net.*, 2015, pp. 1–8.

[19] ——, "Adaptive learning with covariate shift-detection for motor imagery-based braincomputer interface," *Soft Computing*, pp. 1–12, 2015.

[20] C. Vidaurre, A. Schlögl, R. Cabeza, R. Scherer, and G. Pfurtscheller, "Study of on-line adaptive discriminant analysis for eeg-based brain computer interfaces," *IEEE Trans. Biomed. Eng.*, vol. 54, no. 3, pp. 550–556, 2007.

[21] X. Zhu, "Semi-supervised learning literature survey," University of Wisconsin-Madison, Tech. Rep. Computer Science Technical Report 1530, 2008.

[22] B. S. O. Chapelle and A. Zien, Eds., *Semi-supervised learning*. MIT Press, 2006.

[23] H. Cecotti, "Greedy multi-class label propagation," in *Proc. of the Int. Joint Conf. on Neural Net.*, 2015, pp. 1–7.

[24] M. Sugiyama, T. Kanamori, T. Suzuki, M. C. du Plessis, S. Liu, and I. Takeuchi, "Density-difference estimation," *Neural Compututation*, vol. 25, no. 10, pp. 2734–75, Oct. 2013.

[25] A. Gammerman, V. Vovk, and V. Vapnik, "Learning by transduction," in *Proc. 14th Conf. Uncertain. Artif. Intell.*, no. 1, 1998, pp. 148–155.

[26] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.

[27] S. A. Dudani, "The distance-weighted k-nearest-neighbor rule," *IEEE Trans. Syst. Man Cybern.*, vol. 6, no. 4, pp. 325–327, 1976.

[28] M. Tangermann, K. R. Müller, A. Aertsen, N. Birbaumer, C. Braun, C. Brunner, R. Leeb, C. Mehring, K. J. Miller, G. R. Müller-Putz, G. Nolte, G. Pfurtscheller, H. Preissl, G. Schalk, A. Schlögl, C. Vidaurre, S. Waldert, and B. Blankertz, "Review of the bci competition iv," *Front. Neurosci.*, vol. 6, p. 55, 2012.

[29] K. K. Ang, Z. Y. Chin, C. Wang, C. Guan, and H. Zhang, "Filter bank common spatial pattern algorithm on bci competition iv datasets 2a and 2b," *Front. Neurosci.*, vol. 6, p. 39, 2012.

[30] H. Raza, H. Cecotti, and G. Prasad, "Optimising frequency band selection with forward-addition and backward-elimination algorithms in EEG-based brain-computer interfaces," in *Proc. of the Int. Joint Conf. on Neural Net.*, 2015, pp. 1–7.

[31] X. Zhu and Z. Ghahramani, "Learning from labeled and unlabeled data with label propagation," Carnegie Mellon University, Tech. Rep. Technical Report CMU-CALD-02-107, 2002.

[32] H. Ramoser, J. Müller-Gerking, and G. Pfurtscheller, "Optimal spatial filtering of single trial eeg during imagined hand movement," *IEEE Trans. Rehab. Eng.*, vol. 8, no. 4, pp. 441–446, Dec. 2000.

[33] V. Gandhi, G. Prasad, D. Coyle, L. Behera, and M. T. M., "Quantum neural network-based eeg filtering for a brain-computer interface," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 25, no. 2, pp. 278–288, 2014.

[34] K. K. Ang, Z. Y. Chin, H. Zhang, and C. Guan, "Filter bank common spatial pattern (fbcsp)," in *Proc. Int. Joint Conf. Neural Networks (IJCNN)*, 2008, pp. 2390–2397.