Covariate Shift-Adaptation Using a Transductive Learning Model for Handling Non-Stationarity in EEG based Brain-Computer Interfaces

Haider Raza, Girijesh Prasad, Yuhua Li, Hubert Cecotti
Intelligent Systems Research Centre, University of Ulster, Londonderry, Northern Ireland, UK
raza-h@email.ulster.ac.uk, g.prasad@ulster.ac.uk, y.li@ulster.ac.uk, h.cecotti@ulster.ac.uk

***Abstract*—A major challenge to devising robust brain-computer interfaces (BCIs) based on electroencephalogram (EEG) data is the immanent non-stationary characteristics of EEG signals. Statistical properties of the signals may shift during inter-or-intra session transfers that often leads to deteriorated BCI performance. The shift in the input data distribution from training to testing phase is called a covariate shift. It can be caused by various reasons such as different electrode placements, varying impedances and other ongoing brain activities. We propose an algorithm to handle this issue by adapting to the covariate shifts in the EEG data using a transductive learning approach. The performance of the proposed method is evaluated on the BCI competition 2008-Graz dataset B. The results show an improvement in classification accuracy of the BCI system over a traditional learning method. The obtained results support the conclusion that covariate-shift-adaptation using transductive learning is helpful to realize adaptive BCI systems.**

Keywords*— *Non-stationary learning, covaraite shift adaptation, transductive learning, semi-supervised learning.

I. INTRODUCTION

A brain-computer interface (BCI) is an intelligently designed platform, which allows a user to express his or her will without muscle exertion, provided that the brain signals are properly translated into computer commands [1]. However, it may be difficult to satisfactorily classify the electroencephalography (EEG) patterns using fixed traditional classification algorithms because of inherent non-stationarity in the EEG signals. The non-stationarities in the EEG signal can be caused by various reasons such as user attention fluctuations, electrode placement, and user fatigue [2]–[4]. There are notable covariate shifts in the EEG signals during trial-to-trial, and session-to-session transfers [2], [5]. The covariate shift corresponds to a shift in the input data distribution, while the conditional distribution remains the same [6]. The covariate shift is one of the rationales behind the development of an adaptive learning BCI system. Moreover, a robust BCI system should provide visual/auditory feedback to a subject, and the system should be able to adapt to the subject, possibly with an adaptive algorithm. Various adaptive learning algorithms have been presented in the literature. Most of the adaptive learning algorithms made efforts to reduce the effect of non-stationarities in the extracted features [1]–[4], [7]–[9]. However, these require prior information about the ongoing EEG signal characteristics, and the prior information may not be available or maybe difficult obtain in real-world situations.

So far, constantly achieving a satisfactory classification accuracy has been one of the main challenges of the developed BCI systems, which directly affects the decision made from the BCI output [10]. To overcome this issue several feature extraction and feature classification techniques have been presented in the literature. A great variety of features have been attempted to make the BCI system more robust towards non-stationarities such as band powers, power spectral density, autoregressive, and time frequency features [1]. Moreover, some critical properties of the features must also be considered such as noise, outliers, high-dimensionality, and small calibration set for training phase.

To overcome the limitation of stationarity assumption in traditional supervised learning algorithms, we present a transductive learning algorithm. A k-nearest neighbour method is used to implement the proposed transductive learning based on Vapnik's principle [11]. Moreover, the method incorporates a semi-supervised learning (SSL), to extract the meaningful information from the unlabeled features, along with labelled samples. 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 must satisfy at least one of the following assumptions such as smoothness, cluster, or manifold assumption [12], [13]. Here, we have used the smoothness assumption to implement the transductive algorithm i.e., the points which are close to each other are more likely to share the same label. Furthermore, according to a discussion given in [14], it is said that the transductive learning method is a combination of SSL and inductive learning. The covariate shift problem is managed by initiating the adaptation based on the existing and new knowledge obtained through transductive learning. By considering state-of-the-art benchmark data set, we demonstrate the effectiveness of the proposed approach over traditional learning algorithms.

This paper proceeds as follows: Section II presents a background of BCI, non-stationarity and covariate shift in EEG signals. Section III consists of the proposed methodology involving the transductive learning algorithm. Section IV presents the data sets and feature analysis. The results are detailed and discussed in Section V and IV. Finally, Section VII presents the conclusion.

II. BACKGROUND

A. Brain-Computer Interfacing (BCI)

With a BCI, a subject can control a device such as a cursor on the screen, by performing mental activities that are associated with actions, and are dependent on the BCI application. The system operates in two-phases, namely the calibration or training phase, and testing or evaluation (operation) phase. Figure 1 shows a block diagram of a generic BCI system. The EEG signals are acquired through a

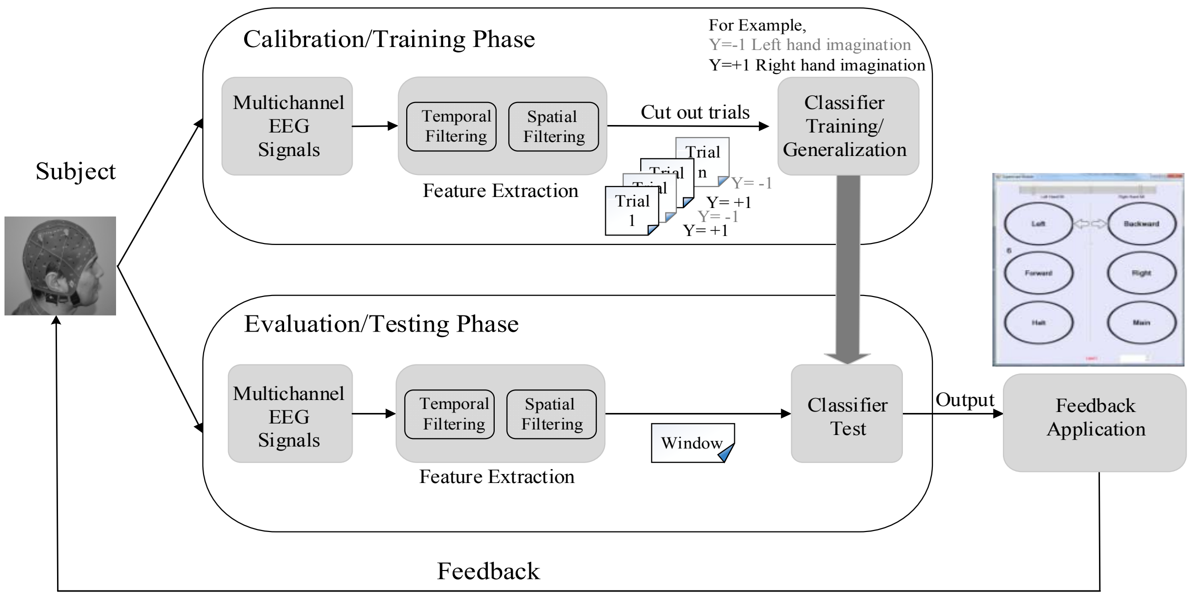
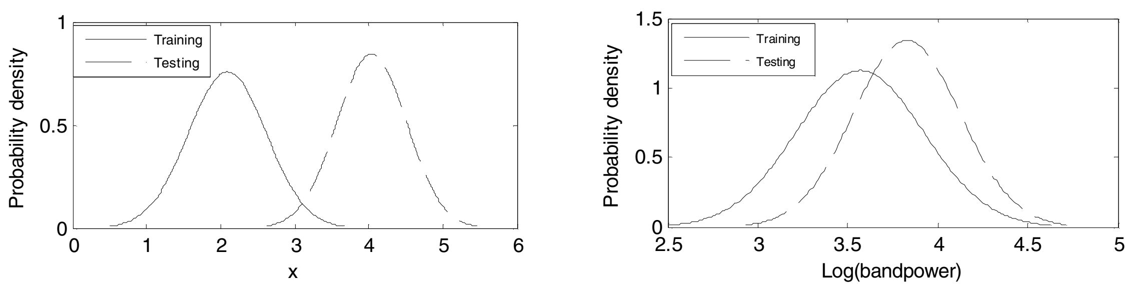


Fig. 1. *Block diagram of a BCI system. The systems runs in two phases. In the training phase, subjects are instructed to perform certain task and short segments of labeled EEG (trials) are collected. The features are extracted from the EEG signals using temporal and spatial filtering. Then, a classifier is trained on these labeled features. In the test phase, a window of features or a sliding window from continuous stream of EGG is passed to the classifier. The classifier outputs a real value that either belongs to class* -1 *or* +1 *according to the likeliness of the class membership. The output of the classifier is passed as input into the feedback/*application [15] and finally the subject receives the feedback from the screen.



(a) Synthetic data (b) log of EEG band power

Fig. 2. *Covariate shift* [6]*: (a) Training dataset has normal distribution with* $\mathcal{N}(x; 2, 1.5)$*, and test dataset also has normal distribution with* $\mathcal{N}(x; 4, 1.5)$*. Thus the mean of the testing data distribution has changed from that of training, resulting in covariate shift. (b) Covariate shift in EEG data between training and testing input distribution for real-world dataset.*

multichannel EEG amplifier, and a pre-processing is performed. In this stage the undesired effects of artifacts, e.g., eye movement or noises, are removed from the signal. In the next stage, BCI task related features are extracted from the signals using feature extraction techniques such as band power filtering and spatial filtering. Spatial filters such as common spatial pattern (CSP) are typically used to extract discriminative spatial filters for the classification of EEG signals. A CSP [3] is a data-driven technique for supervised decomposition of the EEG signals. The new features obtained after CSP are vectors of numbers that are extracted from each trial of the experiment. The classifier then uses the best extracted feature vector to train itself. After training, the users can actually transfer information through their brain activity and control application. Finally, the user interface translates the classifier output into control signals.

B. Non-Stationarity in BCI

Non-stationarity is a very common issue in EEG based BCI. Due to the non-stationarity effects, the covariate shift can be often found in trial-to-trial, and session-to-session transfers of EEG signals. The covariate shift corresponds to a change in the density of the input data distribution over time (Figure 2.). It is well known that due to covariate shift, the generalization performance of the classifier degrades over time [2]. The optimal classification boundary obtained from the calibration session may not be able to correctly classify the upcoming data having very different distribution.

Most of the traditional BCI classification algorithms are built upon the common assumption that the density distribution during training and testing stages remains the same. Unfortunately, this assumption is often violated. To cope up with this issue, several non-stationary learning algorithms have been proposed in the literature. Non-stationary BCI learning algorithms include co-adaptation [1], covariate shift-adaptation using density ratio estimation approach [2], bias adaptation [6], dynamically weighted ensemble classification [16], principal component based covariate shift-adaptation [17], covariate shift-minimization [11], and unsupervised short-term covariate shift-minimization [18]. Most of the methods are focused on reducing the non-stationarity effects or by estimating the density of the upcoming data.

C. Inductive and Transductive learning

Most of the learning model developed and implemented thus far are based on an inductive inference method. A model in an inductive inference is derived from data representing the problem space during the training phase, and this model is further applied to new data. This model is usually created without taking into account any information about the particular new data i.e., test data. The models, in most of the cases, are global models, covering the whole searching space. These models are difficult to update on new data without using old (previous) data. Some global models may consist of multiple local models that collectively cover the whole searching space, and can be adjusted incrementally on new data. In contrast to an inductive inference, a transductive learning claims that it is more powerful to solve a simpler task than inductive learning [14] . In an inductive leaning, one learns a function that makes prediction on the whole search space. A transductive approach asks for less, and it only concerns itself with predicting the values of the function at the test point of interest.

A transductive learning approach is different from a semi-supervised learning (SSL). An SSL uses the information contained in unlabeled data which we have in addition to the training (labelled) data. Transduction on other hand, claims that it is powerful because it is solving a simpler task than the inductive learning. Moreover, a transductive learning algorithm is a combination of SSL and inductive learning [14].

The basic idea implemented in this work concerns a novel transductive learning algorithm. The k-nearest neighbors approach is used to devise the transductive learning, so as to locally fit the model on the selected k-nearest neighbors. The proposed approach is given in the next section.

III. Transductive Learning Model

Transductive inference was already proposed and briefly studied more than 40 years ago by Vapnik and Chervonenkis [19]. Later, it was empirically acknowledged that transduction can often serve as more efficient or accurate learning than traditional inductive supervised learning approaches [13]. This appreciation is the main rationale behind the development of a transductive learning model for the non-stationary BCI. Moreover, transductive inference methods estimate the value of a potential model (function) only at a single point of space (new data) using additional information related to that point [20]. This approach is based on the Vapnik's principle, *"one should avoid solving more difficult intermediate problems when solving a target problem*" [11].

It can be formally stated that the transductive learning is a combination of an SSL and an inductive learning [14]. Using an SSL approach, the unlabeled points may give information about $P(x)$ at the desired point, where P is the probability of the input distribution. The usefulness of the point will depend on whether the distribution has changed significantly. For example, in an 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 point provides useful information. The information extracted from every single point is useful to update the knowledge base (KB). Then, using the inductive learning, the best global model can be obtained, to classify the unlabeled data points from the test distribution. This approach seems more appropriate for clinical and medical applications of learning systems, where the focus is not to fit the global model, but on the individual patient/subject.

In this paper, we propose a transductive learning model based on the k-nearest neighbors principle. Initially, the classifier is trained on the data obtained from the calibration/training stage. Once the classifier is trained and optimal classification boundary is obtained, then the evaluation phase starts. In the evaluation phase, after every five consecutive trials, the classifier will initiate adaptation based on the transductive learning approach. The reason for considering a small number of trials in each epoch is that it results in focusing on trial-by-trial shift, which may occur due to muscular artifacts. The incremental step of five trials was chosen empirically, and in relation to the size of the data set. Each time the classifier initiates adaptation, it is considered as one epoch, also there is no overlap between the epochs. Once the adaptation is initiated, the Euclidean distance (d_i) from the unlabeled data point (p) to the labelled data point (q) in the initial training database is computed as given below:

$$d_{(p,q)} = \sqrt{\sum_{j=1}^{m}(q_j - p_j)^2} \qquad (1)$$

where in Cartesian coordinates q and p are the two different points in Euclidean m-space. The row vector $d = [d_1, \dots\dots\dots, d_n]$ is a vector of Euclidean distances between p and each of the n points in the initial training database. Then, the k-nearest neighbors are selected (for example, if the number of neighbors k=6, then $dist = [d_1, d_2, d_3, d_4, d_5, d_6]$ is a vector of the k shortest distances from d and $l = [l_1, l_2, l_3, l_4, l_5, l_6]$ is the vector of the corresponding labels of the nearest points in the training data set, $l\,(i) \in \{1,2\}$). Next, the inverse of the Euclidean distance is computed as given below:

$$dist_{inv}(i) = \frac{1}{dist(i) + \varepsilon} \qquad (2)$$

where, $\varepsilon = 10^{-4}$, is a appended to avoid the problem of division by zero. A high value of $dist_{inv}(i)$ implies that the corresponding pattern in the training data-base is closer to the current unlabeled feature set. In order to estimate a measure of confidence for the prediction using kNN, the sum of the inverse of Euclidean distances of the selected neighbors in a particular class are computed and divided by the sum of inverse of Euclidean distances of all the k-nearest neighbors as given in equation (3)

$$CR_j = \frac{\sum_1^k dist_{inv}(i) * (l(i) == j)}{\sum_1^k dist_{inv}(i)} \qquad (3)$$

where $j \in \{1,2\}$.

$$CR_{Max} = MAX(CR_1, CR_2) \qquad (4)$$

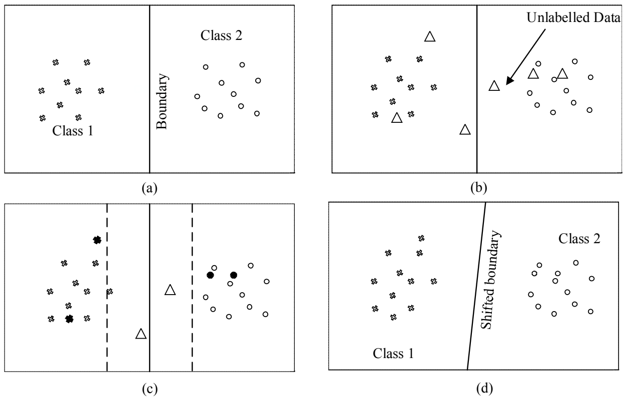


Fig. 3. Schematic diagram for transductive learning. (a) Initial classification boundary *for* class 1 (cross) and class 2 (circle), (b) Triangles represent the unlabeled data points. (c) For each unlabeled point, k-nearest neighbors are selected and the class is assigned to each of them based on confidence ratio. If the score CR_{Max} is more than confidence ratio threshold, then the information is inserted into KB. (d) The classifier is retrained on the updated KB and new shifted boundary is obtained.

This confidence ratio (CR_{Max}) acts as a belief or confidence for a data sample to belong to a particular class. CR_{Max} is used to decide if the current trial's features and estimated label should be added to the existing knowledge base, i.e., labeled data. Figure 3. shows the transductive inference model, where after every n-consecutive trials, the learning function/model and the class separation plane are updated.

We investigate two variants of the proposed transductive learning in this study, namely *Trans₁* and *Trans₂*. In method *Trans₁*, the confidence ratio threshold is empirically set to 0.70. The value of CR threshold is fixed based on an empirical study, later described in the experimental section. If CR_{Max} >0.70, then the information is added to the existing KB, otherwise the information is discarded. Once the new information is added into the training base, the classifier's parameters are re-estimated, and its decision boundary is shifted as shown in Figure 3.

In method *Trans₂*, the number of neighbors (k) and confidence ratio (CR_{Max}) are optimized to maximize the accuracy on the data in session-2. The optimized hyper parameters obtained from the session-2 are fixed, and then the evaluation is performed on the rest of the three sessions. In the evaluation phase, the value of k is fixed for each subject. If the confidence value of predicted sample is greater than the CR threshold then, the information is added to the existing KB, otherwise the information is discarded. After adding the new information, the classifier's parameters are re-estimated and its hyper-plane is shifted.

The main difference between the *Trans₁* and *Trans₂* method is the choice of the parameters (i.e. the number of neighbors (k) and confidence ratio threshold). In the method *Trans₁*, the parameters are fixed based on an empirical study given in experimental section. Whereas in the method *Trans₂*, the parameters are optimized to maximize the accuracy on the data in session-2.

Moreover, to adapt to the covariate shift in the EEG, the data are normalized by removing the mean and dividing by standard deviation in the training phase as given in equation (5),

$$z = \frac{x - \mu}{\sigma} \tag{5}$$

where μ is the mean of the population and σ is the standard deviation of the population. However, in the testing phase for the first epoch (i.e., 5 initial trials), the mean and standard deviation obtained from the training phase are used to normalize the data. After each epoch the new knowledge is inserted in the training base. Hence, the mean and standard deviation are re-estimated and new statistics are obtained, then further standardization is performed using new-statistics. This helps in adapting the covariate shift after each epoch, because the data are normalized to zero-mean.

IV. Datasets and Feature Analysis

A. Data Discription

BCI competition 2008-Graz dataset B [21] is used to evaluate the performance of the proposed method. This dataset consists of EEG data from 9 subjects, namely [B01-B09]. We have used a subset of 7 subjects to evaluate the performance of the proposed method, except subjects B05 and B08. Three channel bipolar recordings (C3, Cz and C4) were made with a sampling frequency of 250 Hz, the montage is shown in Figure 4. All signals were recorded monopolarly with the left mastoid serving as reference and the right mastoid as ground. For each subject, 5 sessions are provided. Each trial is a complete paradigm of 8 seconds, for more detail refer to [22]. For the experimentation in this paper, the session-1 is used for training, the session-2 is used to set additional hyper parameters for method *Trans₂*, and the three remaining sessions are used for testing/evaluation for both the methods *Trans₁* and *Trans₂*.

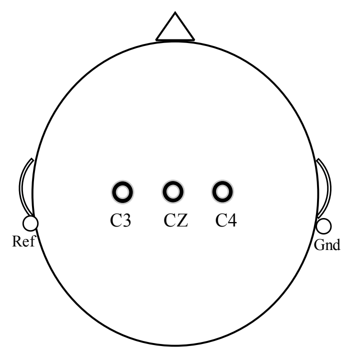


Fig. 4. Electrode montage corresponding to the international 10-20 system.

B. Temporal Filtering

Band power features for all the channels are extracted for mu (μ) (8-12 Hz) and beta (β) (12-30 Hz) bands. During each trial, a specific time over the imagery period is selected to obtain the temporal features for the classification. The information obtained from event-related de-synchronization (ERD) at certain frequency bands are useful for classification. ERD can be seen as a correlate of an activated cortical area, and presents itself as a decrease in power in certain frequency bands in the EEG, and is observable in the μ and central β rhythms.

C. Spatial Filtering

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, effective filtering becomes a problem [3]. Common Spatial Patterns (CSP) is a technique to analyse multichannel data based on the recording from two classes. It is a data-driven supervised decomposition of signals parameterized by a matrix $W \in \mathbb{R}^{C \times C}$ that projects the signal $x(t) \in \mathbb{R}^{C}$ in the original sensor space to $x_{CSP}(t) \in \mathbb{R}^{C}$, which lives in the surrogate sensor space, as given in following equation (6):

$$x_{CSP}(t) = W^{T}\, x(t) \tag{6}$$

In a nutshell, CSP filter maximizes the variance of spatially filtered signal under one condition while minimizing it for the other condition. CSP analysis is applied in order to obtain an effective discrimination of mental states that are characterised by ERD/ERS effects.

V. Experiments and Results

The performance of the proposed transductive learning is evaluated using support vector machines (SVMs), and linear discriminant analysis (LDA) pattern classifiers. The average classification accuracy obtained from the sessions 3, 4 and 5 is presented in Table I and Table II, for SVM and LDA, respectively. The proposed methods are compared with a traditional learning algorithm when CSP is applied for feature extraction and is mentioned in Tables I and II as baseline method.

TABLE I: CLASSIFICATION ACCURACY (%) FOR DIFFERENT SUBJECTS USING SVM CLASSIFIER

Support Vector Machine (SVM)											
		Non-Normalized					**Normalized**				
	Baseline	**Trans_1**		**Trans_2**			**Trans_1**		**Trans_2**		
Subject	**SVM**	**K**	**Acc %**	**K**	**CR**	**Acc %**	**K**	**Acc %**	**K**	**CR**	**Acc %**
B01	51.88	18	54.38	6	0.75	***54.58***	18	52.71	6	0.6	53.33
B02	54.51	6	54.72	12	0.85	***57.29***	18	57.22	18	0.7	54.17
B03	45.42	18	47.50	6	0.60	46.88	6	***50.00***	6	0.75	48.96
B04	94.58	12	***94.67***	6	0.85	93.54	6	89.17	12	0.75	90.42
B06	59.79	6	61.88	6	0.75	***62.08***	18	60.63	18	0.7	58.96
B07	52.71	6	56.88	6	0.85	52.50	6	56.67	12	0.75	***57.08***
B09	61.67	12	63.33	18	0.65	59.17	6	***67.50***	18	0.7	62.71
Mean	**60.08**		**61.91**			**60.86**		**61.98**			**60.80**
STD	**16.13**		**15.36**			**15.22**		**13.24**			**13.77**

TABLE II: CLASSIFICATION ACCURACY (%) FOR DIFFERENT SUBJECTS USING LDA CLASSIFIER

Linear Discriminant Analysis (LDA)											
		Non-Normalized					**Normalized**				
	Baseline	**Trans_1**		**Trans_2**			**Trans_1**		**Trans_2**		
Subject	**LDA**	**K**	**Acc %**	**K**	**CR**	**Acc %**	**K**	**Acc %**	**K**	**CR**	**Acc %**
B01	45.63	18	51.46	6	0.55	50.00	6	***55.00***	6	0.55	***55.00***
B02	56.94	12	55.76	6	0.70	54.51	18	***58.40***	6	0.85	54.79
B03	47.29	6	47.71	6	0.70	46.88	6	***52.08***	6	0.75	51.46
B04	***93.13***	18	92.67	6	0.75	90.63	6	91.67	6	0.65	91.67
B06	61.25	18	62.08	12	0.90	***66.46***	18	62.29	6	0.75	59.18
B07	55.21	6	56.88	6	0.80	53.33	12	55.42	6	0.8	***58.54***
B09	***72.50***	12	68.75	18	0.85	61.46	18	67.92	18	0.8	65.63
Mean	**61.71**		**62.19**			**60.47**		**63.25**			**62.32**
STD	**16.51**		**15.09**			**14.87**		**13.59**			**13.69**

A. Performance Evaluation for Trans1

In the proposed $Trans_1$ method, two parameters are required to be selected, namely, the confidence ratio (CR_{Max}) threshold and the number of neighbors (k). The classifier is initially trained on the data from the session-1 and the parameters (CR_{Max} threshold and k) are selected from the session-2 based on an empirical analysis. The analysis is performed different values of $k \in \{6, 12, 18\}$, and the CR_{Max} threshold is selected in the range [0.50-1]. The evaluated accuracy from the session-2 is used to decide the threshold for the CR_{Max} as presented in Figure 5.

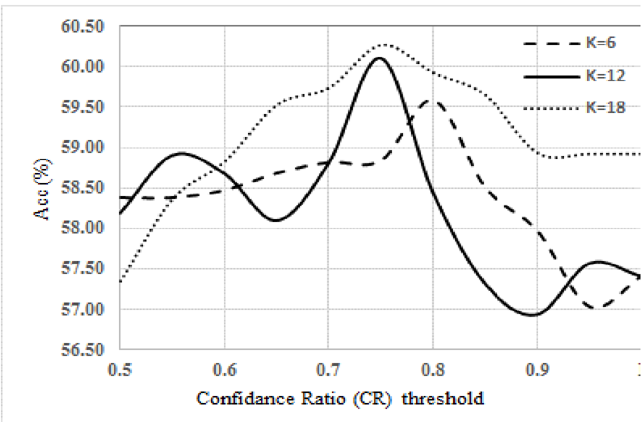


Fig. 5. Classification accuracy obtained for the different values of CR threshold in the range [0.50-1] from session-2 of the dataset. The lines are the mean accuracy of for 7 subjects. The accuracy is evlauted for different values of $k \in \{6,12,18\}$. For all values of k, the best accuracy is obtained in the range between [0.7-0.8].

Figure 5 shows, the maximum accuracy is obtained when the value of CR_{Max} threshold is in between 0.70 and 0.80. So, the value of CR_{Max} threshold is fixed at 0.70 for the evaluation phase. The accuracy and the values of k given in Table I and Table II are the best accuracies for the corresponding values of k. For the SVM, the accuracies given by the method $Trans_1$ are better for all subjects, in both cases of normalized and non-normalized data. But, for LDA, two subjects, namely B04 and B09 in the baseline method are slightly better on both normalized and non-normalized data.

B. Performance Evaluation for $Trans_2$

In the proposed $Trans_2$ method, two parameters are required to be optimally selected, namely, the confidence ratio (CR_{Max}) threshold and the numbers of neighbours (k). The evaluation is performed on the rest of the sessions 3, 4 and 5. The best value of the k is selected from the set $\{6,12,18\}$. The evaluation is performed on each element of the k, whereas *the* CR_{Max} threshold is evaluated and selected in the range [0.50-1]. The accuracies and the values of k given in Table I and Table II are the best accuracies for the corresponding values of k. For the SVM, the accuracies given by the method $Trans_2$ are better for all subjects, in both cases of normalized and non-normalized data. Moreover, the subjects B01, B02, and B06 show an improvement for non-normalized data, however, the subject B07 is better for the normalized data. But, for LDA, the subjects B04 and B09 are slightly better for the baseline method. The subjects B01 and B07 have shown a significant improvement when the data is normalized. Whereas, the subject B06 is best for non-normalized data.

VI. Discussion

The proposed idea is based on Vapnik's principle that given a test point, one should focus on the training points which are in a neighborhood of this test point. The k-nearest neighbors are used to construct a local decision rule (also called local learning), and predict the label of the test point according to this transduction rule. However, the local learning is still inductive because there exists an implicit decision function, even though it is never explicitly constructed. The proposed approach also falls in the category of SSL because, the unlabeled data points $P(x)$ are used to extract the meaningful information. The CR_{Max} threshold is used to measure whether the information in hand is useful or not. If the information is useful then it is added to the existing KB.

The estimation of CR_{Max} threshold and k is an important issue. These parameters are required to be carefully selected in order to achieve better performance. For method $Trans_1$, the value of CR_{Max} threshold is empirically selected in the range [0.50-1]. Based on the analysis presented in Figure 5, it shows that the CR_{Max} threshold ranging between [07-0.8] has maximum accuracy. This implies that the performance of the proposed method mostly depends upon the optimal choice of CR_{Max} threshold. Hence, to choose the CR_{Max} threshold optimally, a detailed study was done, so that the subject specific parameters are selected. Based on the selected parameters the proposed method shows an improvement over the baseline method. Another important issue is to decide the number of trials after, which the adaptation is initiated. The reason of considering a small number of trials in each epoch is that it results in focusing on trial-by-trial shift. However, the long term non-stationarities may be accounted for by considering larger number of trials in each epoch. As discussed earlier, this approach seems more appropriate for clinical and medical applications of learning systems, where the focus is not to fit the global model, but on the individual subject.

The experimental results demonstrated the effectiveness of the proposed transductive learning strategy for EEG based BCI. The proposed method with CSP filters outperforms the traditional learning methods.

VII. Conclusion

We proposed the transductive learning strategies to reduce the effect of covariate shifts in non-stationary EEG signals. Our approach is based on Vapnik's principle "one should avoid solving more difficult intermediate problems when solving a target problem". The information from the unlabeled data are extracted based on semi-supervised learning and transductive learning approaches. The extracted information is then added to the existing knowledge base. Based on the updated knowledge base, the classifier's parameters are re-estimated and the classification boundary is shifted. Experimental results demonstrated that the new techniques are especially effective for non-stationary EEG signals. The proposed methods have better classification accuracy over the baseline method. This framework provides a foundation for transductive learning and semi-supervised learning for EEG based BCI. Following the proposed framework, more advanced strategies can be developed. A promising extension of the strategies would be to detect the covariate shift in the data and then, the corrective adaptive initiative can be taken. This approach may reduce the unnecessary efforts needed to retrain the classifier.

REFERENCES

[1] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi, "A Review of Classification Algorithms for EEG-Based Brain-Computer Interfaces.," *J. Neural Eng.*, vol. 4, no. 2, pp. R1–R13, Jun. 2007.

[2] 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. 1318–24, Jun. 2010.

[3] B. Blankertz and R. Tomioka, "Optimizing spatial filters for robust EEG single-trial analysis," *IEEE Signal Process. Mag.*, no. January 2008, pp. 41–56, 2008.

[4] H. Cecotti, M. P. Eckstein, and B. Giesbrecht, "Single-Trial Classification of Event-Related Potentials in Rapid Serial Visual Presentation Tasks Using Supervised Spatial Filtering," *IEEE Trans. Neural Networks Learn. Syst.*, pp. 1–1, 2014.

[5] H. Raza, G. Prasad, and Y. Li, "Dataset Shift Detection in Non-Stationary Environments using EWMA Charts," in *IEEE International Conference on Systems, Man, and Cybernetics*, 2013, pp. 3151–3156.

[6] H. Raza, G. Prasad, and Y. Li, "EWMA Based Two-Stage Dataset Shift-Detection in Non-stationary Environments," in *Artificial Intelligence Applications and Innovations.*, 2013, pp. 625–635.

[7] K.-R. Müller, M. Krauledat, G. Dornhege, G. Curio, and B. Blankertz, "Machine Learning Techniques for Brain-Computer Interfaces," *J. Biomed. Eng.*, vol. 49, pp. 11–22, 2004.

[8] M. Arvaneh, G. Cuntai, K. Kai, and Q. Chai, "Optimizing Spatial Filters by Minimizing Within-Class Dissimilarities in Electroencephalogram-Based Brain-Computer Interface," *IEEE Trans. neural …*, vol. 24, no. 4, pp. 610–619, 2013.

[9] B. Rivet, H. Cecotti, M. Perrin, E. Maby, and J. Mattout, "Adaptive training session for a P300 speller brain-computer interface.," *J. Physiol. Paris*, vol. 105, no. 1–3, pp. 123–9, 2011.

[10] H. Cecotti, "Spelling with non-invasive Brain-Computer Interfaces--current and future trends.," *J. Physiol. Paris*, vol. 105, no. 1–3, pp. 106–14, 2011.

[11] V. Vapnik, "An Overview of Statistical Learning Theory," *IEEE Trans. Neural Networks*, vol. 10, no. 5, pp. 988–99, Jan. 1999.

[12] X. Zhu, "Semi-supervised learning literature survey," *Comput. Sci. Univ. Wisconsin-Madison*, 2006.

[13] O. Chapelle, B. Scholkopf, and A. Zien, *Semi-supervised learning*. MIT Press, 2006.

[14] O. Chapelle, B. Schölkopf, and A. Zien, "A discussion of semi-supervised learning and transduction," *Semi-Supervised Learn.*, pp. 473 – 478, 2006.

[15] V. Gandhi, G. Prasad, D. Coyle, L. Behera, and T. M. McGinnity, "EEG-Based Mobile Robot Control Through an Adaptive Brain-Robot Interface," *IEEE Trans. Syst. Man, Cybern. Syst.*, pp. 1–1, 2014.

[16] S. R. Liyanage, C. Guan, H. Zhang, K. K. Ang, J. Xu, and T. H. Lee, "Dynamically Weighted Ensemble Classification for Non-Stationary EEG Processing.," *J. Neural Eng.*, vol. 10, no. 3, p. 036007, Apr. 2013.

[17] W. Rosenstiel, M. Bogdan, and M. Sp, "Principal Component Based Covariate Shift Adaption to Reduce Non-Stationarity in a MEG-Based Brain-Computer Interface," *EURASIP J. Adv. Signal Process.*, pp. 2–8, 2012.

[18] R. Mohammadi, A. Mahloojifar, and D. Coyle, "Unsupervised Short-term Covariate Shift Minimization for Self-paced BCI," in *IEEE Symposium on Computational Intelligence, Cognitive Algorithms, Mind, and Brain (CCMB)*, 2013, no. i, pp. 101–106.

[19] V. N. Vapnik and A. J. Chervonenkis, "Theory of pattern recognition (1974).," 1974.

[20] N. Kasabov, *Evolving connectionist systems*. Springer-Verlag London Limited, 2007.

[21] R. Leeb, C. Brunner, Müller-Putz, A. G. R., Schlögl, and G. Pfurtscheller, "BCI Competition 2008–Graz data set B. Graz University of Technology, Austria." 2008.

[22] 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, no. July, p. 55, Jan. 2012.