Hindawi
Computational Intelligence and Neuroscience
Volume 2018, Article ID 6323414, 12 pages
https://doi.org/10.1155/2018/6323414

Research Article

Multiclass Informative Instance Transfer Learning Framework for Motor Imagery-Based Brain-Computer Interface

Ibrahim Hossain, Abbas Khosravi, Imali Hettiarachchi, and Saeid Nahavandi

Institute for Intelligent Systems Research and Innovation, Deakin University, Geelong, VIC, Australia

Correspondence should be addressed to Ibrahim Hossain; ihossai@deakin.edu.au

Received 8 October 2017; Accepted 14 January 2018; Published 22 February 2018

Academic Editor: Amparo Alonso-Betanzos

A widely discussed paradigm for brain-computer interface (BCI) is the motor imagery task using noninvasive electroencephalography (EEG) modality. It often requires long training session for collecting a large amount of EEG data which makes user exhausted. One of the approaches to shorten this session is utilizing the instances from past users to train the learner for the novel user. In this work, direct transferring from past users is investigated and applied to multiclass motor imagery BCI. Then, active learning (AL) driven informative instance transfer learning has been attempted for multiclass BCI. Informative instance transfer shows better performance than direct instance transfer which reaches the benchmark using a reduced amount of training data (49% less) in cases of 6 out of 9 subjects. However, none of these methods has superior performance for all subjects in general. To get a generic transfer learning framework for BCI, an optimal ensemble of informative and direct transfer methods is designed and applied. The optimized ensemble outperforms both direct and informative transfer method for all subjects except one in BCI competition IV multiclass motor imagery dataset. It achieves the benchmark performance for 8 out of 9 subjects using average 75% less training data. Thus, the requirement of large training data for the new user is reduced to a significant amount.

1. Introduction

Brain-computer interface (BCI) is a system that establishes a communication channel between the brain and control devices without using the neuromuscular system of human body [1].

One of the noninvasive modalities of BCI is electroencephalography (EEG). BCI uses different types of EEG control signal from the external scalp of the brain. Some of the control signals used in BCI are visual evoked potential (VEP), P300 evoked potential, slow cortical potential (SCP), and sensory-motor rhythms (SMR) [2]. SMR can be modulated by actual as well as imagery motor task by user [3, 4]. Thus, SMRs are used in BCI as the control signal for translating motor task (hand and foot movement) [5] and referred to as motor imagery- (MI-) BCI. Hence, MI-BCIs are used for supporting patients with spinal cord injury and stroke [6–8]. MI-based BCI system possesses some drawbacks such as lack of robustness, complex setup, and long calibration time [1, 9, 10].

Generally, it is recommended to use at least five times more training data per class than the features [11, 12]. Channel-frequency-time information makes the feature vector of EEG signal very high-dimensional [13, 14]. These high-dimensional features necessitate the requirement for a large number of EEG epochs to be collected to train the classifier [15, 16]. But, EEG data acquisition is a lengthy and exhaustive task for the user. For motor imagery purpose, it is sometimes a day-long process [10, 16]. EEG signals recorded from the scalp are very subjective. It varies from one subject to another for same tasks. Even, it differs for same subject in different sessions [4]. Consequently, each individual has to go through this long data collection process in each attempt of using the system. It is most likely that long calibration time for a user has become one of the bottlenecks of BCI system. Calibration time reduction approaches reported in literatures also reflect the scenario well [17–22].

If labeled samples for certain tasks are available from other users, these samples can be used for a new user. The objective is to utilize the knowledge from data spaces of past

users to learn predictive function for a new user. This process of knowledge and information conveying from other domain is known as transfer learning (TL) [23].

TL has been applied for BCI in two types: domain adaptation and rules or knowledge sharing [24]. Some of the domain adaptation approaches are subject invariant common space [18, 19, 25–28], common stationary subspace transfer [29–31], conditional and marginal distribution adaptation [32, 33], and subject-to-subject adaptation [34]. Rule adaptation or sharing prior learning to learn new user prediction function has been attempted in [20, 26, 28, 35]. Active transfer learning (ATL) approach was proposed and implemented by Wu et al. in [36] for VEP based BCI. In their work, actively learned samples from the domain of new coming subject were combined with the samples of historical subjects. QBC was used as active learning method to select samples from the subject-specific domain. The authors used all samples from other subjects directly without any adaptation or selection. An improved version of ATL was proposed and implemented for binary MI-based BCI in our preliminary work [37, 38]. Both works implemented ATL on binary classification. In [36], authors did it for target and nontarget VEP while our preliminary work was done on left-hand and right-hand motor imagery classification with two different feature extraction processes in sequence. Since instances are transferred directly from the source to target domain, it is named as direct transfer with active learning (DTAL). DTAL needs to be investigated for multiclass BCI. Instead of direct transfer, an informative and functional subdomain transfer from source to target also needs to be introduced in DTAL. In addition to finding actively learned samples from target domain (in DTAL), active learning based on most uncertain samples from the source to target domain is introduced in this work. To serve these purposes, the following attempts are made in this paper:

(i) Multiclass direct transfer with active learning (mDTAL) is formulated and implemented. It is the multiclass extension of active transfer learning proposed in [36] for motor imagery BCI (Section 3.1).

(ii) Then, aligned instance transfer is introduced for multiclass MI-based BCI (Section 3.2).

(iii) After that, informative instances transfer framework is formulated and implemented with and without aligned subspace. Here, multiclass entropy as uncertainty criterion is applied in the source to target domain transfer (Section 3.3).

(iv) To address the subject-dependent performance variation of different methods, a generic optimized weighted ensemble of all proposed methods is constructed and applied (Section 3.4).

The main goal of this work is to develop an informative transfer learning framework for MI-BCI which is expected to perform better than direct transfer (mDTAL).

The rest of the paper is organized as follows: Section 2 will describe the concept of different terms and methods which are used for further algorithm's development. Section 3 will describe developed multiclass frameworks and optimized ensemble method. Section 4 will describe experimental setup. Then, Section 5 will analyze and discuss the results. Finally, Section 6 will conclude the paper with the scope of future improvement.

2. Methods

2.1. Transfer Learning (TL). At first, we need to define some terms to state our problem in the scope of transfer learning.

Domain. A domain D consists of $\{\chi, P(X)\}$. Here χ is features of n dimension $(x_1, x_2, \ldots, x_n)$ and $P(X)$ means marginal distribution. So, $D_S = D_T$ means $P_S(X) = P_T(X)$ and $\chi_S = \chi_T$. Similarly, $D_S \neq D_T$ means $P_S(X) \neq P_T(X)$ or/and $\chi_S \neq \chi_T$ [23].

Task. $T = \{Y, f(\cdot)\}$, where Y is set of all class label and $f(\cdot)$ is prediction function which is trained on $\{X, Y\}$. From probabilistic view point, $f(\cdot)$ will give conditional probability $P(Y \mid X)$. So $T_S \neq T_T$ means $Y_S \neq Y_T$ or /and $P_S(Y \mid X) \neq P_T(Y \mid X)$ [23].

Transfer Learning [23]. Given a source domain D_S and learning task T_S, a target domain D_T, and learning task T_T, transfer learning aims to help improve the learning of the target predictive function $f_T(\cdot)$ on D_T using the knowledge in D_S and T_S, where $D_S \neq D_T$ or $T_S \neq T_T$.

Dataset of EEG epochs from a new user is the target domain. EEG epochs with the label from past users are source domain. Same feature extraction method has been applied for both target and source EEG epochs. So, it can be implied that $\chi_T = \chi_S$. Same types of classes are labeled for imagery EEG epochs in both source and target domain. It implies that $Y_T = Y_S$. But, different subjects neural responses to same motor imagery action have different characteristics. As a result, marginal distribution and conditional distribution are different for source and target domain [33]. That means $P_S(X) \neq P_T(X)$ and $P_S(Y \mid X) \neq P_T(Y \mid X)$.

So, samples from source domain cannot represent the target domain correctly. Hence, it needs to get some subdomain from source efficiently which is related mostly to the target domain. The aim of TL is to learn a target prediction function $f(X_T) \rightarrow Y_T$ so that expected error on D_T is as low as possible while $P_S(X) \neq P_T(X)$ and $P_S(Y \mid X) \neq P_T(Y \mid X)$.

In this paper, our approach is to select the most informative instances from source domains with the help of few samples of the target domain. Then, we will add them to target domain samples to train a classifier for predicting the label of independent test data of target domain.

2.2. Active Learning (AL). Active learning method queries for unlabeled samples which have most uncertainty [40]. Trained hypothesis on labeled samples gets confused over some unlabeled samples. These samples are more close to decision line. So, labeling these uncertain samples will accelerate learning process of the model. Hence, these samples carry more information than other certain samples (Figure 1). In this work, active learning method is applied to two ends.

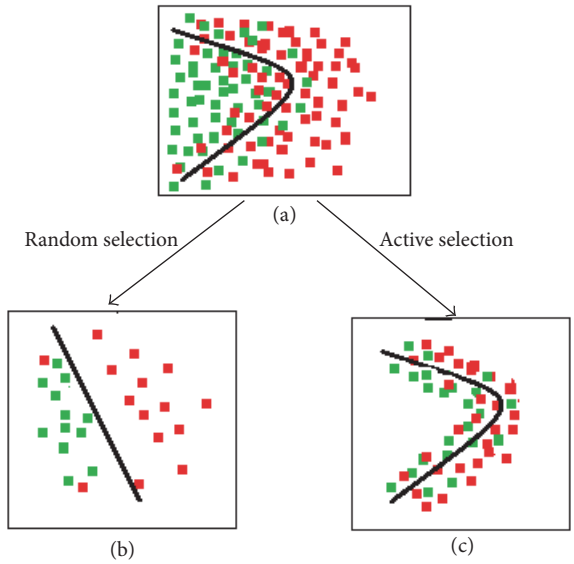

Figure 1: Visual presentation of AL. (a) 2D presentation of binary class dataset with expected decision boundary. (b) Learned decision line on randomly selected samples. (c) Learned decision line on actively selected samples which is more close to expected line.

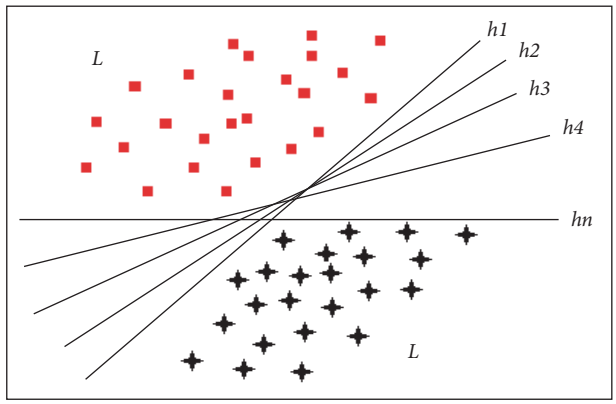

Figure 2: Illustration of linear version space. Each hypothesis in version space is consistent on L (labeled training set). Each of them represents different regions of version space.

At first, query by committee is applied to select the most informative samples from target domain.

Then, entropy is applied as uncertainty measure to select informative samples from the source domain.

Query by Committee (QBC) [41]. A hypothesis is a kind of particular set of parameters that tuned on some training set and it can make the prediction over new data. Hypothesis space is all possible set of hypotheses. Version space is a subset of these hypotheses which are consistent with the labeled training set L (Figure 2). Consistent means that the member of version space can make a correct prediction on all instances of L. One of the aims of AL is to select instances which can narrow down this version space. It will make the process of learning target prediction function more precise with fewer labeled instances.

QBC maintains a committee of hypothesis (version space) $C = \{C_0, C_1, C_2, \ldots, C_M\}$ (Figure 2). Each member of this committee is trained on labeled data L and represents a candidate hypothesis $(h_1, h_2, \ldots, h_n)$. Then, each member of committee votes for unlabeled samples about their label. The instances attaining the most disagreement about label among the members are considered as the most informative. In analytical perspective, QBC implementation has two steps:

(i) Construction of committee of hypotheses which depict various regions of version space from specific to general (Figure 2)

(ii) Quantification of disagreement among the members of the committee.

In this work, linear discriminant analysis (LDA) is our learning model. This model gives negative decision score for one class and positive for others. So, decision boundary ideally is zero scoreline. It is unlikely to get extreme negative (-1) at the same time extreme positive $(+1)$ score for a single sample. Certain instances will have the extreme sum of decision score for which most of the members are agreed. But, uncertain instances will not have the extreme score for any class. It makes the absolute value of the sum of the score for all classes close to zero. In case of LDA, ensemble sum of decision score close to zero represents more disagreement among the members. So, instances attaining the lowest absolute value of the algebraic sum of decision scores from members of the committee are the most informative.

Entropy. Entropy is the amount of information to encode a distribution [42]. It is used as the measurement of uncertainty. For binary classification, entropy enforces us towards posterior probability 0.5. For multiclass, entropy yields a central confusing area of posterior probability. It considers probability distributions for all classes.

$$x_{\text{Entropy}} = \max_{x_i} -\sum_{c=1}^{n_C} P_\theta\left(y_c \mid x_i\right) \log P_\theta\left(y_c \mid x_i\right). \quad (1)$$

Here, $P_\theta(y_c \mid x_i)$ is the predicted probability of ith sample x_i for class y_c by the model θ. n_C is the number of classes.

2.3. Feature Extraction: Common Spatial Pattern (CSP). This method maximizes the variance for one task and minimizes the variance for other task. Therefore, it yields to generate discriminating features of two classes for EEG classification [43–45]. Let us consider that $X_i \in R^{\text{ch}\times t}$ is the ith single-trial bandpass EEG signal and $Z \in R^{\text{ch}\times t}$ is the spatially filtered signal with CSP projection matrix $A \in R^{\text{ch}\times\text{ch}}$. Here, ch is the number of channels and t is the number of time points in single-trial bandpass EEG epoch.

$$Z = A^T X_i. \quad (2)$$

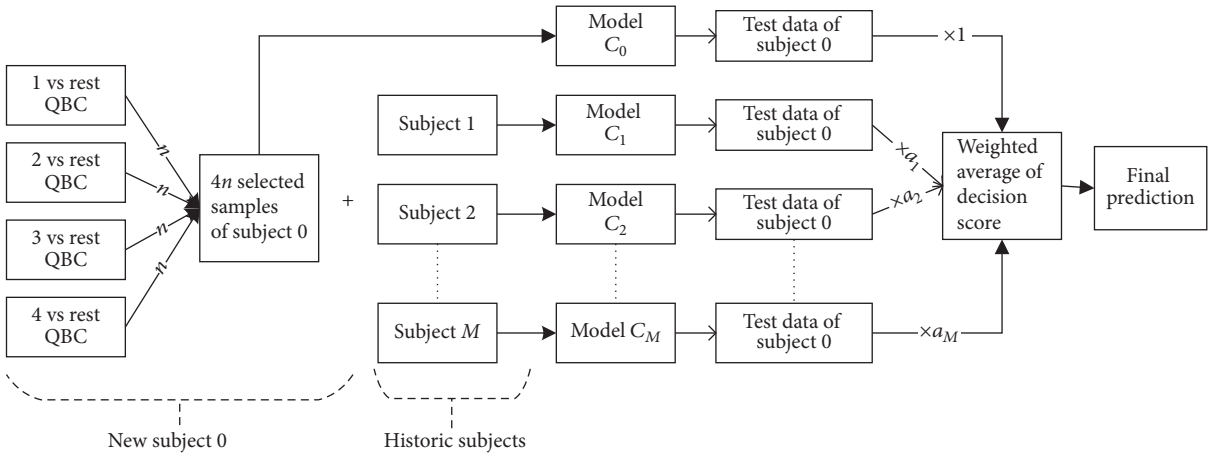

Figure 3: Multiclass direct transfer with active learning (mDTAL). Here, n is the number of samples to be selected from each class of new subject.

Δ_1 and Δ_2 are the covariance matrixes of EEG signals X for two classes which can be obtained by

$$\Delta_Y = \frac{1}{n_Y} \sum_{i \in I_Y} X_i X_i^T \quad Y = [1, 2]. \tag{3}$$

Here, I_Y is the set of indices of trials corresponding to class Y and n_Y is the total number of trials from class Y. A is the transformation matrix satisfying below optimization.

$$\max_A \quad \frac{A^T \Delta_1 A}{A^T \Delta_2 A}. \tag{4}$$

This CSP filter matrix A can be obtained by solving

$$\Delta_1 A = (\Delta_1 + \Delta_2) AD. \tag{5}$$

Here, D is a diagonal matrix and it contains eigenvalues.

Generally, m first and m last rows of A (represented by $A^* \in R^{c \times 2m}$) make the spatial filtered signal Z^* [46]:

$$Z^* = X^T A^*. \tag{6}$$

Finally, logarithm of variance of Z will give the feature vector F.

$$F = \log\left(\operatorname{var}\left(Z^*\right)\right). \tag{7}$$

This CSP is for binary class. We have used four *one vs rest* binary CSP for four classes implementation [44].

2.4. Linear Discriminant Analysis (LDA). LDA is simple and fast to compute [47, 48] which is very successfully paired with CSP feature extraction for MI-based BCI. For binary classification, it deals with two scatter matrixes S_w and S_b which are named as *within-class* and *between-class* scatter. S_w and S_b are defined as follows:

$$S_w = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_i^j - \omega_i\right)\left(x_i^j - \omega_i\right)^T \tag{8}$$

$$S_b = \frac{1}{n} \sum_{i=1}^{k} n_i \left(\omega_i - \omega\right)\left(\omega_i - \omega\right)^T. \tag{9}$$

Here, ω_i denotes the mean vector of ith class and ω denotes the total mean vector. k and n are number of classes and total number of samples, respectively. Objective is set to find matrix G for transformation such that it can ensure maximization of between-class and minimization of within-class scatter. In this work, four *one vs rest* LDA classifiers are used for 4-class classification.

$$\max_G \quad \frac{\operatorname{tr}\left(G^T S_b G\right)}{\operatorname{tr}\left(G^T S_w G\right)}. \tag{10}$$

Decision score is calculated by

$$f(x) = Gx + b. \tag{11}$$

Here, b is the bias value and sign of $f(x)$ will give the class label.

3. Algorithms

3.1. Multiclass Direct Transfer with Active Learning (mDTAL). Multiclass extension of direct transfer learning with active learning or ATL [36] is formulated for MI-based BCI. CSP is used for feature extraction combined with LDA classifier since this combination is very successful for MI-based BCI [16, 46]. For mDTAL, we have considered *one vs rest* approach [49, 50] in three sections of this algorithm (Figure 3):

(i) *One vs rest* method for QBC while selecting most active samples from target domain

(ii) *One vs rest* CSP filter in feature extraction part

(iii) *One vs rest* method for LDA training and testing part.

Stepwise process of mDTAL algorithm is described as follows.

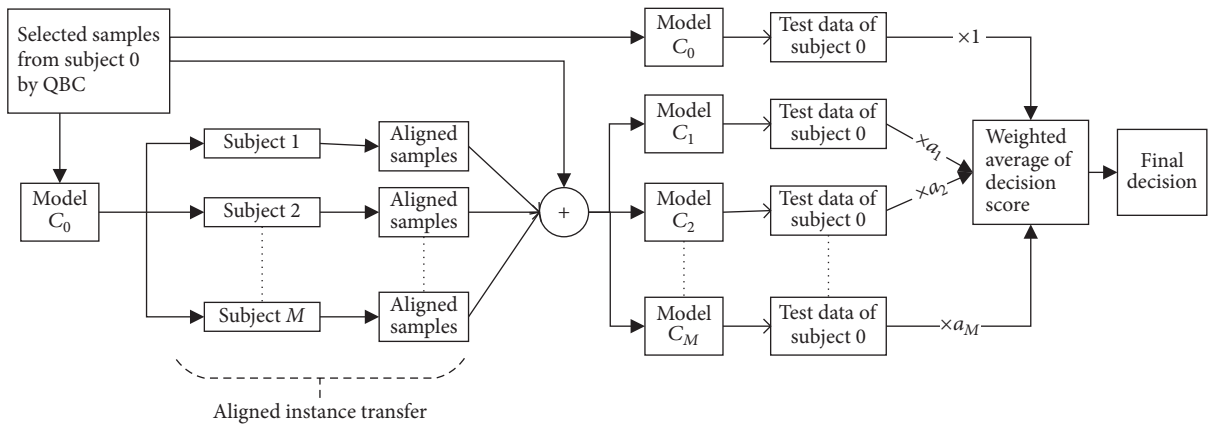

Figure 4: Multiclass aligned instance transfer with active learning (AITAL).

Algorithm: mDTAL

Step 1. Start with randomly chosen N^l labeled samples with equal class proportion and N^u unlabeled instances from target domain. M number of other subjects with N^m labeled instances from mth subject are available. N^t independent test samples of new subject are given for performance evaluation.

Step 2. Train classifier C_0 using N^l samples. Then, C_0 will calculate the decision score for each class of N^u samples as D_u^0.

Step 3. Train combined classifier C_m using $N^l \cup N^m$ combined training samples.

Step 4. Get 10-fold cross-validation accuracy a_m on $N^l \cup N^m$ training samples. Repeat Steps 2 and 3 for all historic subjects.

Step 5. Get ensemble weighted average decision score for each class on N^u using the following equation:

$$D_u = \frac{D_u^0 + \sum_{m=1}^{M} (a_m * D_u^m)}{1 + \sum_{m=1}^{M} a_m}. \tag{12}$$

Here, D_u^m is decision score calculated using classifier C_m on unlabeled samples N^u. For D_u^0, weight has been assigned as 1 to give subject-specific classifier higher priority over combined classifier. Similarly, ensemble weighted average decision score for test data set N^t is also calculated as follows:

$$D_t = \frac{D_t^0 + \sum_{m=1}^{M} (a_m * D_t^m)}{1 + \sum_{m=1}^{M} a_m}. \tag{13}$$

It is the ultimate output of the algorithm in each iteration.

Step 6. Linear classifier LDA has the negative score for one class and positive for other. So, decision score close to zero represents more uncertainty than others. Equation (12) calculates ensemble decision score of the $M + 1$ number of models or a committee of models. Unlabeled samples getting lowest or close to zero absolute decision score are more likely to learn decision boundary than others. Considering multiclass, $D_u(:, c)$ gives decision score for class *c vs rest*. So, the lowest absolute decision score of $D_u(:, c)$ will give most uncertain samples near class *c vs rest* boundary as follows (Figure 6):

$$S_c = \min_n ascend\left\{abs\left\{D_u(:, c)\right\}\right\}. \tag{14}$$

Here, $c = 1, 2, \ldots, n_C$ (number of class) and n is number of samples to be selected from each class (Figure 3).

Step 7. All selected unlabeled subject-specific samples S_c are queried for label. Then this newly labeled samples are added to N^l and removed from N^u. Steps 2 to 7 are repeated until maximum number of iteration.

3.2. Multiclass Aligned Instance Transfer with Active Learning (AITAL). There is no adaptation or selection from historic subjects in mDTAL method. Rather, it directly transfers all labeled samples from historic subjects. But, all samples from historic subjects may not be compatible with the domain of new subject. As a result, it may yield to negative transfer effect [51]. So, the idea is to transfer samples which are aligned with new subject decision boundary (Figure 4). Subject-specific model C_0 classifies some samples accurately from historic subjects. It can be assumed that these accurately classified samples agree with the decision boundary of target domain classifier C_0. So, these samples are considered as being aligned with target domain.

AITAL is similar to mDTAL algorithm except Step 3 where it will not take all of N^m samples from mth historic subject. Instead, it will take $N^{m'}$ aligned samples (see (15)) from mth historic subject which are determined by subject-specific model C_0 (Figure 4).

$$N^{m'} = N^m \mid \left\{L_0^m == Y\right\}, \quad Y = [1, \ldots, 4]. \tag{15}$$

Here, L_0^m is the label for samples from mth historic subjects which are predicted by subject-specific classifier C_0. Y is the true class label for these samples.

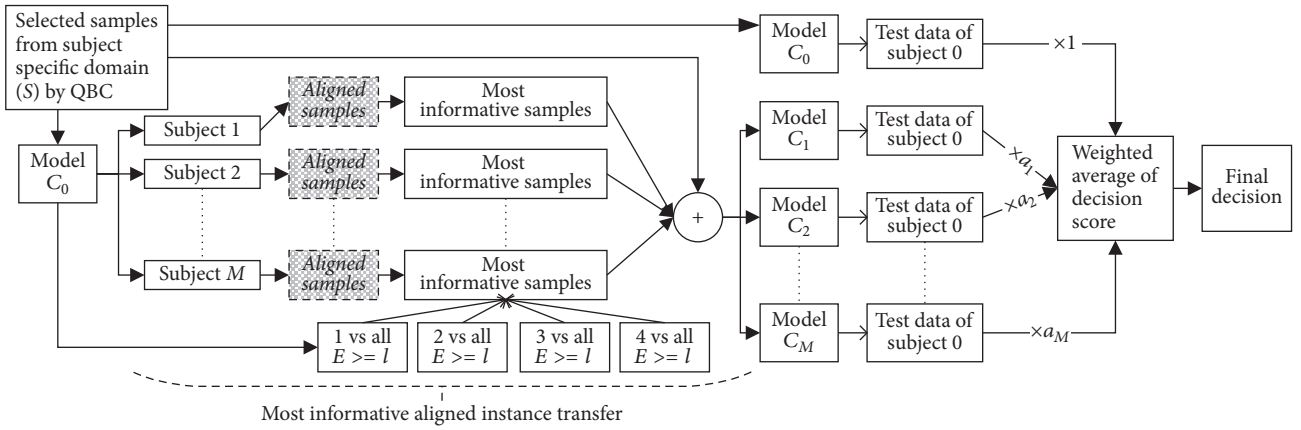

Figure 5: Multiclass most informative and aligned informative instance transfer with active learning (MIAITAL). Shaded box is obsolete for MIITAL.

3.3. Most Informative Instance Transfer with Active Learning (MIITAL). According to active learning query method, samples lying close to decision boundary are more likely uncertain to be predicted. It makes uncertain samples more informative to learn decision boundary than that of other samples. If informative samples from historic subjects are transferred to learn classifier for new user, it will be more effective. In this work, entropy of instances are used as the quantification of information carried by these samples. Entropy can be calculated by

$$E(i) = -\sum_{c} P_\theta(y_c \mid x_i) \log P_\theta(y_c \mid x_i) \qquad i = 1, 2, \ldots, N^m,\ c = [1, \ldots, 4]. \tag{16}$$

Here, $P_\theta(y_c \mid x_i)$ is probability of samples x_i to be in class y_c which is determined by classifier C_0 and represented as model θ.

For this work, we consider four *one vs rest* entropy calculation. Our goal is to find uncertain samples which are close to each *one vs rest* decision line. Ideally, samples having 50 : 50 probability ratio are most uncertain and have maximum entropy. We consider samples with probability ratio equal or more than 60 : 40 for this work. It yields to transfer samples that have entropy equal or greater than 0.29228 according to (16). This entropy limit is named information limit or cut-off (l).

There are two combinations of this algorithm:

(i) Transfer aligned and most informative samples (most informative and aligned instances transfer with AL (MIAITAL)) (Figure 5):

$$N^{m'} = N^m \mid \{L_0^m == Y,\ E \geq l\}, \quad Y = [1, \ldots, 4]. \tag{17}$$

(ii) Transfer most informative samples and ignore whether it is aligned or not (most informative instances transfer with AL (MIITAL) (Figure 5)):

$$N^{m'} = N^m \mid \{E \geq l\}. \tag{18}$$

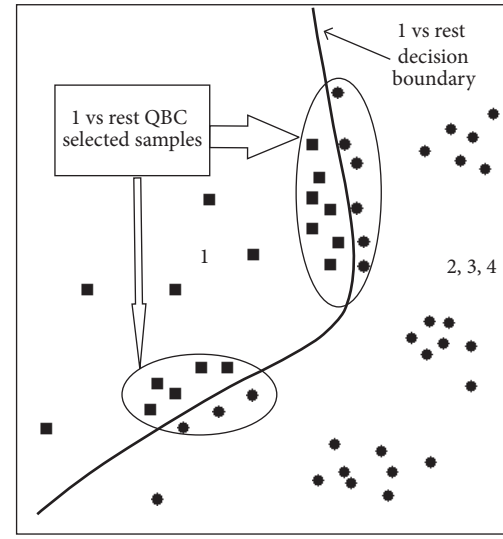

Figure 6: One versus rest presentation of QBC for multiclass.

Algorithm of MIAITAL or MIITAL is the same as mDTAL except Step 3. In Step 3, MIAITAL or MIITAL will take $N^{m'}$ according to (17) and (18) in place of N^m.

MIAITAL attempts to transfer most informative samples which are perfectly classified by classifier from previous iteration, whereas MIITAL attempts to transfer most informative samples (determined by entropy) and ignores alignment of those informative samples (Figure 5).

3.4. Optimized Ensemble for Multiclass Actively Learned Space Transfer. EEG epochs due to various motor imagery actions are not stable. So, finding prominent features followed by learning classifier does not always yield the expected result. As a result, performance is not generic for all subjects; it is subject-dependent. Some methods perform well for some subjects while not very good for others. The ensemble of different methods can give a general and steady performance

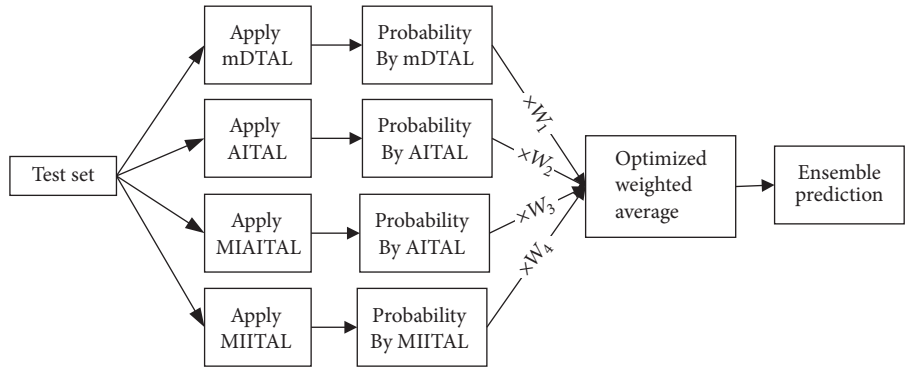

Figure 7: Optimized weighted ensemble of mDTAL, AITAL, MIAITAL, and MIITAL.

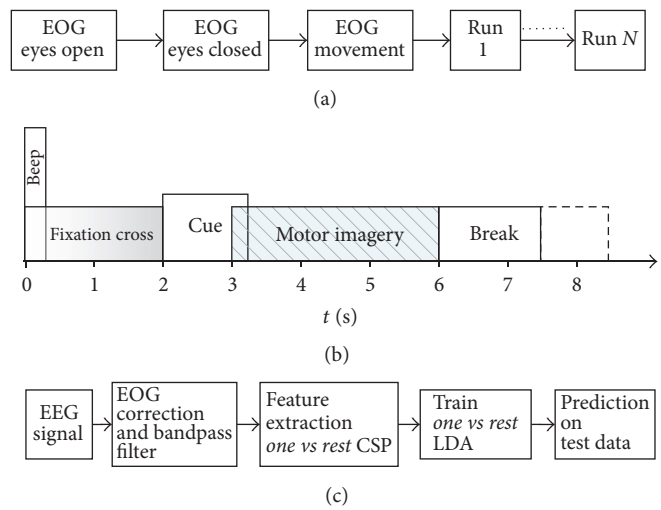

Figure 8: (a) Timing for one session [39]. (b) Timing for a single trial [39]. (c) Baseline method.

for all subjects. An optimized weighted ensemble is proposed to serve the purpose (Figure 7).

Some arbitrary weight (say W) is assigned for each class in each method. Then, these weights are optimized for minimum loss on a validation set. Loss function for nth subject is as follows:

$$\text{LOSS}_n = \sum_{i \in V} \left(Y_n^t(i) - W_m^c * P_m^c(i) \right)^2. \tag{19}$$

Here, the validation set V is 10 percent of the subject-specific training set and is randomly chosen from that data set. The initial value of weight W is some random value in the range of $[0, 1]$. Then, W is optimized by the genetic algorithm using the loss function from (19). Ensemble decision-making probability on test set T using optimized W is then obtained by

$$P_{\text{opt}}^c(x) = \frac{\sum_{m=1}^{M} W_m^c * P_m^c(x)}{\sum_{m=1}^{M} W_m^c}, \quad x \in T. \tag{20}$$

Here, P_m^c is probability generated by mth method for class c and W_m^c is optimized weight for corresponding class c and method m.

4. Experiment Setup

4.1. Experimental Data. Algorithms described in Section 3 are implemented for BCI competition IV dataset 2A [39]. This dataset consists of 9 subjects. In this dataset, each subject performs four types of motor imagery action for left hand, right hand, both feet, and tongue movement. Data is recorded in two sessions for each subject. In each session, a subject performs 72 trials per class which turns into 288 in total.

In each session, there are approximately 5 minutes of electrooculogram (EOG) recording keeping eyes open, close, and moving. Then, it is followed by the run of trials (Figure 8(a)). Each subject was facing a computer screen which was showing different indication guideline to the subject. Each single-trial starts ($t = 0$ s) with a fixation cross on the screen in front of the subject. After 2 seconds ($t = 2$ s), a cue appeared on the screen indicating arrow with the desired movement sign (left hand, right hand, foot, and tongue). After 1.25 seconds of cue appearing, subject starts to imagine the motor action and continues until $t = 6$ s. A short break (black screen) is given until next trial starts (Figure 8(b)). EEG epoch of 2 seconds after 0.5 seconds of cue appearing is taken as training data. 22 channels (Ag/AgCl) are used for

EEG signal recording and other three monopolar electrodes are used for EOG recording. The montage of the electrode was according to the international 10-20 system. Both of the EEG and EOG channels were sampled 250 Hz. After that, they had been filtered using 0.5 Hz to 100 Hz ranged bandpass filter. A 50 Hz notch filter was also enabled during recording to omit the line noise. The sensitivity of the amplifier was set to 100 uV and 1 mV for EEG and EOG recording, respectively.

4.2. Data Preprocessing and EOG Correction. Linear EOG correction method [52] is applied for artefact correction on raw EEG signals. β rhythms (12–30 Hz) of EEG signals are desynchronized with real movement or motor imagery [53]. μ rhythms (8–12 Hz) of EEG signals related to motor actions and sometimes correlate with β rhythms [54, 55]. For this reason, corrected EEG signal is bandpass filtered using casual Chebyshev Type II filter between 8 Hz and 32 Hz. After that, CSP is applied and features are extracted according to (6) and (7). Here, m is set to 2 for A^* in (6). So, 4 features are obtained from each EEG epoch.

4.3. Experiment and Simulation. For all method, first session of each subject is used as training set and second session is used as test set.

For comparison purpose, a baseline method is also implemented. In baseline method, the full training set of the respective subject is used to train LDA classifier. No sample from other users is used. After applying data preprocessing as described in Section 4.2, four *one vs rest* LDA classifiers are trained. These models are applied to predict label for respective independent test session (Figure 8(c)).

The accuracy achieved by this baseline method is the benchmark performance by an individual user. The purpose of other methods in this work is to achieve this performance using a reduced amount of training samples. This baseline process is followed for each internal model training and testing phase of other algorithms. As benchmark performance is a static value and does not depend on the iterative increment of subjective training samples, it is a straight line parallel to the horizontal axis.

Other methods in this work are iterative where samples from the new subject are added in training pool iteratively. Each subject is considered as the new user (target) while other 8 subjects are considered to be past users (source). Each simulation starts with 40 random samples (N^l in Step 1 of mDTAL algorithm) with equal class distribution from the target domain. Then, 2 samples per class (n in (14)) are added in each iteration until 20 iterations (maximum number of iteration in Step 7 of Section 3). So, maximum 200 subjective samples for each subject is added at final iteration. This amount of training samples from the new user is good enough to observe whether the new subject can reach the benchmark using a lower amount of training samples. For this reason, the maximum number of iterations is set to 20. This whole simulation is repeated 20 times for each subject to negate random starting samples effect. Then, the average of ten repeats in each iteration is taken as the performance of that iteration.

Only first session of each historical subjects is taken as source domain because label for the second session was kept closed in BCI competition IV. Training samples from the first session of target subject are added iteratively and the classification performance in each iteration is computed on the independent second session of the target subject.

5. Results and Discussion

The performance of proposed methods in this work is evaluated based on the following two criteria: first, investigation to find whether the method has reached the maximum baseline performance; second, the number of subjects for which intended method reaches the maximum baseline performance. Direct transfer method (mDTAL) is the multiclass extension of active transfer learning [36] for motor imagery BCI. Proposed informative space transfer algorithms (AITAL, MIAITAL, and MIITAL) will be compared with mDTAL based on the evaluation criteria mentioned above. Figure 9 presents the accuracy of all methods for comparison. The following observations can be drawn out from this result based on the above-mentioned evaluation criteria:

(i) mDTAL method fails to achieve the baseline performance except for subject A03, A06, and A08. But, it is showing gradual increment in accuracy as the training data from target domain increased. In mDTAL method, all samples from source domain are transferred to the target domain directly. The results reflect that, due to high variability among subjects, there is a high chance of completely different types of domain transfer in direct transfer method.

(ii) AITAL method reaches the baseline for subjects A01, A02, A03, and A06. It depicts that transferring solely aligned information does not always yield to transfer of discriminative features transformation. Widely sparse distribution may be aligned but might not have much information for target domain learning process. Moreover, aligned samples are not ensured to be equal in class distribution. So, there is a high possibility of introducing class-imbalance into the combined domain (source + target).

(iii) MIITAL and MIAITAL transfer most informative samples in each iteration. Both of them reach baseline or close to baseline for subjects A01, A02, A03, A06, A08, and A09. MIITAL shows better performance than MIAITAL. The reason behind this is that MIITAL emphasizes only on information carried out by samples while MIAITAL requires both informative and aligned samples. Some of the informative samples may not be aligned. These nonaligned informative samples with higher entropy are excluded in MIAITAL but are included in MIITAL. Thus, MIITAL outperforms MIAITAL with more informative samples.

From above observation, it is clear that informative transfer approaches (MIITAL and MIAITAL) have reached the baseline for six out of nine subjects while direct transfer

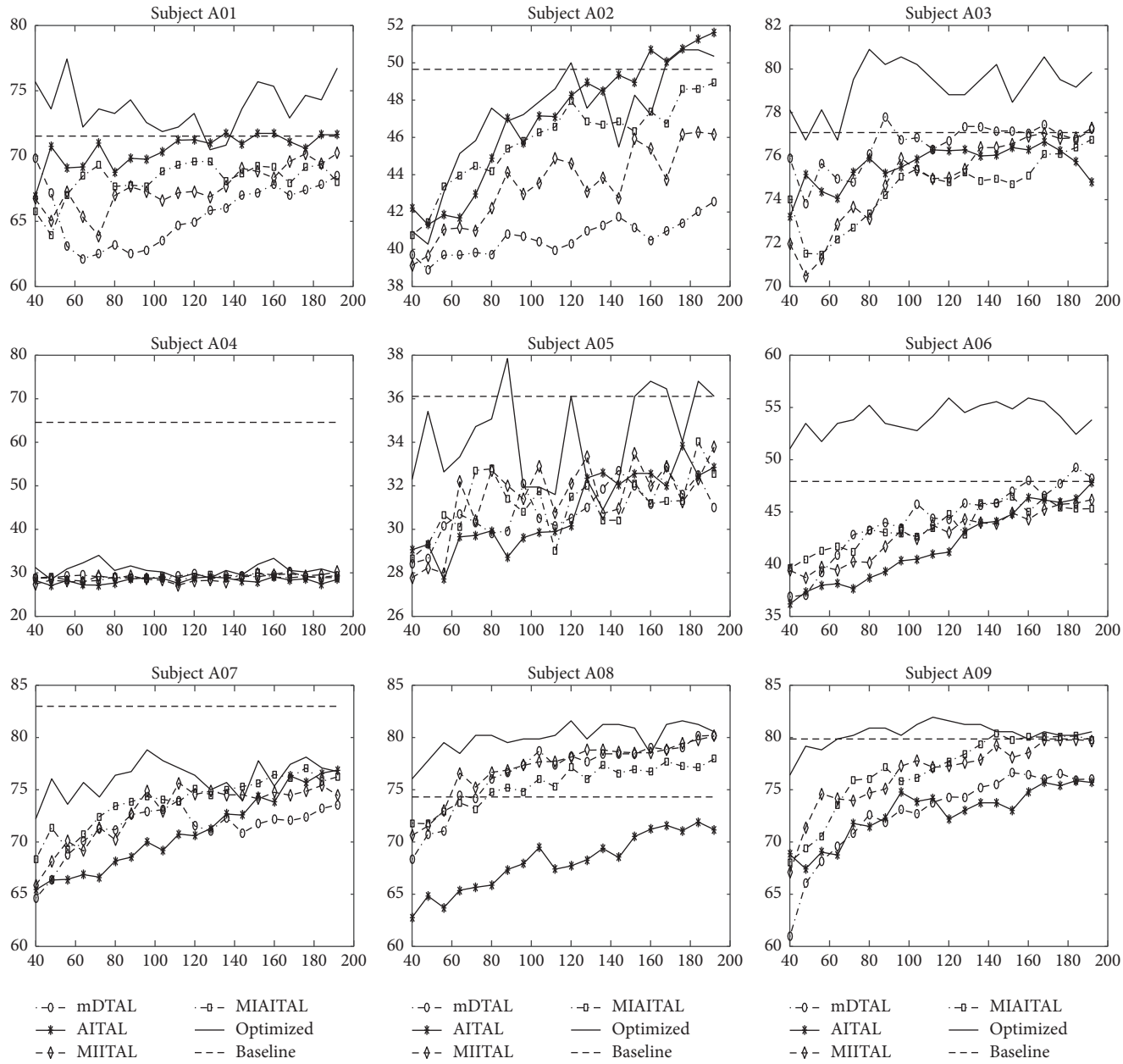

Figure 9: Performance of mDTAL, AITAL, MIITAL, MIAITAL, and optimized ensemble method on BCI competition IV dataset 2A. Accuracy is along the y-axis and the number of subject-specific training samples is along the x-axis.

(mDTAL) reaches baseline only for three out of nine subjects. It implies that informative subspace transferring enables the new subject to achieve the baseline performance with a reduced number of training data for more number of subjects compared with direct transfer methods. Table 1 shows the number of subject-specific training samples required to reach baseline or close to baseline. It also implies that MIITAL method reaches baseline or close to baseline using average 49% less subject-specific data for 6 out of 9 subjects.

Though informative instance transfer achieves better performance for most of the subjects, this is not a generic outcome for all subjects. Subject A05 and subject A07 are much closer to baseline, but they do not reach it. Exceptionally, subject A04 is very far from the expected line for all the methods. To find a generic solution for all subjects, an optimized weighted ensemble of the proposed four methods is applied (Figure 7). Performance of optimized weighted ensemble method is shown in Figure 9 (solid black line).

Optimized ensemble of all methods achieves the baseline and sometimes better than baseline with less amount of subject-specific samples for 8 out of 9 subjects. As per results in Table 1, optimized ensemble method reaches the baseline or close to baseline using average 75.5% less subject-specific training samples for 8 out of 9 subjects.

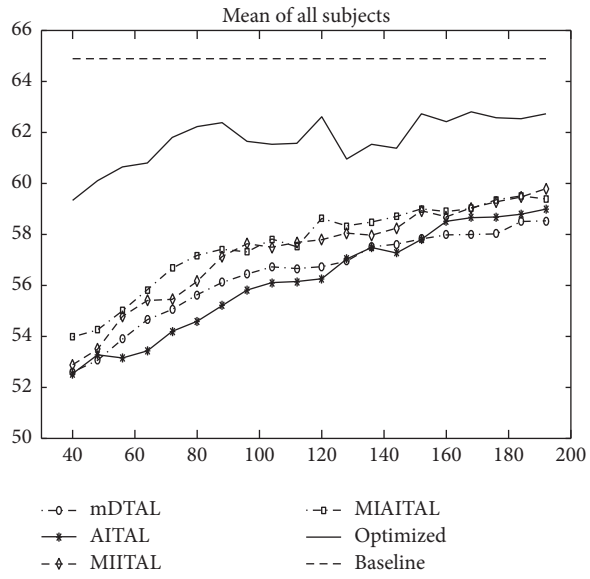

Figure 10: Mean performance of all subjects for mDTAL, AITAL, MIITAL, MIAITAL, and optimized ensemble methods. Accuracy is along the y-axis and number of training samples added from the new subject is along the x-axis.

Table 1: Number of samples to reach baseline for different methods.

Sub	Base	MIITAL	Reduction by MIITAL	Optimized	Reduction by optimized
A01	288	180	37.5%	40	86%
A02	288	180	37.5%	90	68.75%
A03	288	140	51.4%	40	86%
A04	288	×	×	×	×
A05	288	×	×	120	58%
A06	288	170	41%	40	86%
A07	288	×	×	100	65.3%
A08	288	60	79%	40	86%
A09	288	150	48%	70	75.7%
Mean	288		49%		76.56%

*×: baseline cannot be reached.

To get a generic view irrespective of subjects, mean of the accuracy of all subjects is presented in Figure 10. It shows that proposed informative transfer learning methods MIITAL and MIAITAL are performing always better than that of direct transfer (mDTAL). It infers that informative transfer has advantages over the direct transfer. However, mean performance of both algorithms is behind the baseline performance by 4-5%. On the other hand, mean of the optimized ensemble is much closer to mean baseline of all subjects (differs by only 1-2%). Subjective combination adaptation would have yielded better results in comparison to the optimized ensemble of the methods. However, this will be considered in a future study.

Another observation is that subject A04 has no improvement by all these methods. Any of the methods used in this study is unable to improve the performance of subject A04. This can be due to the fact that EEG response of some subjects have complete dissimilarity with others [55]. When a completely dissimilar subspace is transferred and combined with the target subject (A04), it does not much effect towards the improvement of predictive function for the target domain. A remedy for this issue could be achieved by clustering closely related subject [28]. Closely related subjects or domains form a cluster. Nonrelated or dissimilar subjects are excluded from this cluster. Then, informative subspace from this close group or cluster can make the transfer more effective. For EEG epochs consisting large number of channels, EEG channels selection could be a better addition for robustification [56–60].

Presented results infer that a single method is working well for some subjects and not up to the mark for others. It implies that performance of proposed TL methods is subject-dependent. Automatic selection of the best approach for a subject is an open question to be investigated. One of the possible causes behind the performance variation is CSP applied for extracting features from a broad range of μ and β rhythms (8–32 Hz). Subjective frequency ranges can be yielded into better feature extraction and selection [49]. Incorporating this subject adaptive frequency ranges will ensure feature transfer from subjective range. Thus, it will lead to better features transfer into proposed TL algorithm. One concerning matter is the mean baseline performance of multiclass BCI that is not up to the mark. Advance feature extraction and learning algorithm could be applied to raise up this baseline which leads to subsequent incorporation into MIITAL and consequent performance raise of the MIITAL algorithm.

In summary, this paper presents two slightly different informative subspace transfer frameworks (MIAITAL and MIITAL) on multiclass BCI. Though MIITAL has achieved the expected result for a good number of subjects, still it is lagging behind the baseline in general. The optimized ensemble of these methods has overcome the gap. The primary goal of this work is to investigate the functionality of informative subspace transferring over the direct transfer for multiclass BCI. Though it succeeded for most of the subjects, there are many scopes to improve in the proposed framework. Secondary goal is to find comparatively better informative transfer approach. From empirical results, it is clear that MIITAL is serving the purpose better than MIAITAL.

6. Conclusion

In this work, we applied direct transfer learning with active learning on multiclass motor imagery BCI. To improve the performance, an informative instances transfer framework is proposed. Its key advantage compared with direct transfer methods is transferring informative instances that narrow down the search spaces more precisely around the decision line. Hence, it reduced training data significantly for most of the subjects (6 out of 9). A generic optimized ensemble of proposed methods is also implemented. It has achieved expected accuracy with fewer subject-specific samples (using average 75% less training samples) for 8 out of 9 subjects.

Results achieved in this paper point out some directions for future work as well. Subject adaptive method selection could give a more fine-tuned performance. Cluster base transfer combined with informative transferring could also lead to better performance for the underperforming subject. Another scope is filtering subject and subspace based on distribution similarity. Domain adaptation based on marginal and conditional distribution could introduce more generalize adaptation in the proposed TL framework. All these improvements can reduce the calibration effort remarkably and lead us towards a generic TL framework for BCI application.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] J. R. Wolpaw and E. W. Wolpaw, "Preface," *Brain-Computer Interfaces: Principles and Practice*, 2012.

[2] L. F. Nicolas-Alonso and J. Gomez-Gil, "Brain computer interfaces, a review," *Sensors*, vol. 12, no. 2, pp. 1211–1279, 2012.

[3] G. Pfurtscheller, C. Neuper, D. Flotzinger, and M. Pregenzer, "EEG-based discrimination between imagination of right and left hand movement," *Electroencephalography and Clinical Neurophysiology*, vol. 103, no. 6, pp. 642–651, 1997.

[4] G. Pfurtscheller and F. H. L. da Silva, "Event-related EEG/MEG synchronization and desynchronization: basic principles," *Clinical Neurophysiology*, vol. 110, no. 11, pp. 1842–1857, 1999.

[5] R. Scherer, A. Schloegl, F. Lee, H. Bischof, J. Janša, and G. Pfurtscheller, "The self-paced graz brain-computer interface: Methods and applications," *Computational Intelligence and Neuroscience*, vol. 2007, Article ID 79826, 2007.

[6] S. Silvoni, A. Ramos-Murguialday, M. Cavinato et al., "Brain-computer interface in stroke: a review of progress," *Clinical EEG and Neuroscience*, vol. 42, no. 4, pp. 245–252, 2011.

[7] G. Pfurtscheller, G. R. Müller-Putz, R. Scherer, and C. Neuper, "Rehabilitation with brain-computer interface systems," *The Computer Journal*, vol. 41, no. 10, pp. 58–65, 2008.

[8] K. K. Ang and C. Guan, "Brain-computer interface in stroke rehabilitation," *Journal of Computing Science and Engineering*, vol. 7, no. 2, pp. 139–146, 2013.

[9] B. Blankertz, M. Tangermann, C. Vidaurre et al., "The Berlin brain-computer interface: non-medical uses of BCI technology," *Frontiers in Neuroscience*, vol. 4, article 198, 2010.

[10] J. B. F. Van Erp, F. Lotte, and M. Tangermann, "Brain-computer interfaces: beyond medical applications," *The Computer Journal*, vol. 45, no. 4, pp. 26–34, 2012.

[11] A. K. Jain and B. Chandrasekaran, "39 Dimensionality and sample size considerations in pattern recognition practice," *Handbook of Statistics*, vol. 2, pp. 835–855, 1982.

[12] S. J. Raudys and A. K. Jain, "Small sample size effects in statistical pattern recognition: recommendations for practitioners," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 3, pp. 252–264, 1991.

[13] I. T. Hettiarachchi, T. T. Nguyen, and S. Nahavandi, "Motor imagery data classification for BCI application using wavelet packet feature extraction," in *Proceedings of the the International Conference on Neural Information Processing*, pp. 519–526, Springer, 2014.

[14] I. T. Hettiarachchi, T. T. Nguyen, and S. Nahavandi, "Multivariate adaptive autoregressive modeling and Kalman filtering for motor imagery BCI," in *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, SMC 2015*, pp. 3164–3168, Hong Kong, October 2015.

[15] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: a review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4–37, 2000.

[16] B. Blankertz, G. Dornhege, M. Krauledat, K.-R. Müller, and G. Curio, "The non-invasive Berlin brain-computer interface: fast acquisition of effective performance in untrained subjects," *NeuroImage*, vol. 37, no. 2, pp. 539–550, 2007.

[17] M. Krauledat, M. Schröder, B. Blankertz, and K.-R. Müller, "Reducing calibration time for brain-computer interfaces: A clustering approach," in *Proceedings of the 20th Annual Conference on Neural Information Processing Systems, NIPS 2006*, pp. 753–760, can, December 2006.

[18] M. Krauledat, M. Tangermann, B. Blankertz, and K.-R. Müller, "Towards zero training for brain-computer interfacing," *PLoS ONE*, vol. 3, no. 8, Article ID e2967, 2008.

[19] S. Fazli, F. Popescu, M. Danóczy, B. Blankertz, K.-R. Müller, and C. Grozea, "Subject-independent mental state classification in single trials," *Neural Networks*, vol. 22, no. 9, pp. 1305–1312, 2009.

[20] M. Alamgir, M. Grosse-Wentrup, and Y. Altun, "Multitask learning for brain-computer interfaces," in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pp. 17–24, 2010.

[21] W. Tu and S. Sun, "A subject transfer framework for EEG classification," *Neurocomputing*, vol. 82, pp. 109–116, 2012.

[22] F. Lotte, "Signal processing approaches to minimize or suppress calibration time in oscillatory activity-based brain-computer interfaces," *Proceedings of the IEEE*, vol. 103, no. 6, pp. 871–890, 2015.

[23] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[24] V. Jayaram, M. Alamgir, Y. Altun, B. Scholkopf, and M. Grosse-Wentrup, "Transfer Learning in Brain-Computer Interfaces," *IEEE Computational Intelligence Magazine*, vol. 11, no. 1, pp. 20–31, 2016.

[25] H. Kang, Y. Nam, and S. Choi, "Composite common spatial pattern for subject-to-subject transfer," *IEEE Signal Processing Letters*, vol. 16, no. 8, pp. 683–686, 2009.

[26] H. Kang and S. Choi, "Bayesian common spatial patterns for multi-subject EEG classification," *Neural Networks*, vol. 57, pp. 39–50, 2014.

[27] F. Lotte and C. Guan, "Regularizing common spatial patterns to improve BCI designs: unified theory and new algorithms," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 2, pp. 355–362, 2011.

[28] D. Devlaminck, B. Wyns, M. Grosse-Wentrup, G. Otte, and P. Santens, "Multisubject learning for common spatial patterns in motor-imagery BCI," *Computational Intelligence and Neuroscience*, vol. 2011, Article ID 217987, 2011.

[29] W. Samek, F. C. Meinecke, and K. Muller, "Transferring subspaces between subjects in brain—computer interfacing," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 8, pp. 2289–2298, 2013.

[30] W. Samek, C. Vidaurre, K.-R. Müller, and M. Kawanabe, "Stationary common spatial patterns for brain-computer interfacing," *Journal of Neural Engineering*, vol. 9, no. 2, Article ID 026013, 2012.

[31] P. von Bünau, F. C. Meinecke, F. Király, and K.-R. Müller, "Finding stationary subspaces in multivariate time series," *Physical Review Letters*, vol. 103, no. 21, Article ID 214101, 2009.

[32] M. Long, J. Wang, G. Ding, S. J. Pan, and P. S. Yu, "Adaptation regularization: A general framework for transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 5, pp. 1076–1089, 2014.

[33] D. Wu, V. J. Lawhern, and B. J. Lance, "Reducing Offline BCI Calibration Effort Using Weighted Adaptation Regularization with Source Domain Selection," in *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, SMC 2015*, pp. 3209–3216, Hong Kong, October 2015.

[34] M. Arvaneh, I. Robertson, and T. E. Ward, "Subject-to-subject adaptation to reduce calibration time in motor imagery-based brain-computer interface," in *Proceedings of the 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2014*, pp. 6501–6504, USA, August 2014.

[35] W. Samek, M. Kawanabe, and K.-R. Muller, "Divergence-based framework for common spatial patterns algorithms," *IEEE Reviews in Biomedical Engineering*, vol. 7, pp. 50–72, 2014.

[36] D. Wu, B. Lance, and V. Lawhern, "Transfer learning and active transfer learning for reducing calibration data in single-trial classification of visually-evoked potentials," in *Proceedings of the 2014 IEEE International Conference on Systems, Man and Cybernetics - SMC*, pp. 2801–2807, San Diego, CA, USA, October 2014.

[37] I. Hossain, A. Khosravi, and S. Nahavandhi, "Active transfer learning and selective instance transfer with active learning for motor imagery based BCI," in *Proceedings of the 2016 International Joint Conference on Neural Networks, IJCNN 2016*, pp. 4048–4055, Canada, July 2016.

[38] I. Hossain, A. Khosravi, I. T. Hettiarachchi, and S. Nahavandhi, "Informative instance transfer learning with subject specific frequency responses for motor imagery brain computer interface," in *Proceedings of the 2017 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pp. 252–257, Banff, AB, October 2017.

[39] C. Brunner, R. Leeb, G. Müller-Putz, A. Schlögl, and G. Pfurtscheller, *Bci competition 2008–graz data set a, Institute for Knowledge Discovery (Laboratory of Brain-Computer Interfaces)*, vol. 16, Graz University of Technology, 2008.

[40] B. Settles, *Active Learning Literature Survey*, vol. 52, University of Wisconsin, Madison, Wis, USA, 2010.

[41] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," in *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pp. 287–294, ACM, July 1992.

[42] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*, The University of Illinois Press, Urbana, Ill, USA, 1949.

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

[44] J. Müller-Gerking, G. Pfurtscheller, and H. Flyvbjerg, "Designing optimal spatial filters for single-trial EEG classification in a movement task," *Clinical Neurophysiology*, vol. 110, no. 5, pp. 787–798, 1999.

[45] T. Nguyen, I. Hettiarachchi, A. Khosravi, S. M. Salaken, A. Bhatti, and S. Nahavandi, "Multiclass EEG data classification using fuzzy systems," in *Proceedings of the 2017 IEEE International Conference on Fuzzy Systems, FUZZ 2017*, Italy, July 2017.

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

[47] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, pp. 179–188, 1936.

[48] J. Ye, R. Janardan, and Q. Li, "Two-dimensional linear discriminant analysis," *Advances in Neural Information Processing Systems*, vol. 17, pp. 1569–1576, 2005.

[49] 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," *Frontiers in Neuroscience*, vol. 6, article 39, 9 pages, 2012.

[50] G. Dornhege, B. Blankertz, G. Curio, and K.-R. Müller, "Increase information transfer rates in BCI by CSP extension to multi-class," in *Proceedings of the 17th Annual Conference on Neural Information Processing Systems, NIPS 2003*, can, December 2003.

[51] M. T. Rosenstein, Z. Marx, L. P. Kaelbling, and T. G. Dietterich, "To transfer or not to transfer," in *NIPS 2005 Workshop on Inductive Transfer: 10 Years Later*, vol. 2, p. 7, 2005.

[52] A. Schlögl, C. Keinrath, D. Zimmermann, R. Scherer, R. Leeb, and G. Pfurtscheller, "A fully automated correction method of EOG artifacts in EEG recordings," *Clinical Neurophysiology*, vol. 118, no. 1, pp. 98–104, 2007.

[53] G. Pfurtscheller and C. Neuper, "Motor imagery and direct brain-computer communication," *Proceedings of the IEEE*, vol. 89, no. 7, pp. 1123–1134, 2001.

[54] J. A. Pineda, "The functional significance of mu rhythms: translating 'seeing' and 'hearing' into 'doing'," *Brain Research Reviews*, vol. 50, no. 1, pp. 57–68, 2005.

[55] G. Pfurtscheller, C. Brunner, A. Schlögl, and F. H. Lopes da Silva, "Mu rhythm (de)synchronization and EEG single-trial classification of different motor imagery tasks," *NeuroImage*, vol. 31, no. 1, pp. 153–159, 2006.

[56] A. Ghaemi, E. Rashedi, A. M. Pourrahimi, M. Kamandar, and F. Rahdari, "Automatic channel selection in EEG signals for classification of left or right hand movement in Brain Computer Interfaces using improved binary gravitation search algorithm," *Biomedical Signal Processing and Control*, vol. 33, pp. 109–118, 2017.

[57] S. Park, J. Kim, and K. Sim, "EEG electrode selection method based on BPSO with channel impact factor for acquisition of significant brain signal," *Optik - International Journal for Light and Electron Optics*, vol. 155, pp. 89–96, 2018.

[58] C. Zhang, X. Deng, Y. Tang, G. Wang, and D. Li, "Optimization of electrode electroencephalography channel selection based on UPS-EMOA algorithm," *Journal of Medical Imaging and Health Informatics*, vol. 7, no. 5, pp. 1093–1098, 2017.

[59] A. Franklin Alex Joseph and C. Govindaraju, "Channel selection using glow swarm optimization and its application in line of sight secure communication," *Cluster Computing*.

[60] J. Yang, H. Singh, and E. L. Hines, "Channel selection and classification of electroencephalogram signals: an artificial neural network and genetic algorithm-based approach," *Artificial Intelligence in Medicine*, vol. 55, no. 2, pp. 117–126, 2012.