STATIONARY TRANSFER COMPONENT ANALYSIS FOR BRAIN COMPUTER INTERFACING

Sidath R.Liyanage1, Jialin S. Pan2, Haihong Zhang2, Kai Keng Ang2, Cuntai Guan2, Jian-Xin Xu1, Tong Heng Lee1
1 National University of Singapore, Singapore
2 Institute for Infocomm Research, A*STAR, Singapore
sidath@nus.edu.sg, {jsp, hhzh, kkang, ctguan}@i2r.a-star.edu.sg, {elexujx, eleleeth}@nus.edu.sg

ABSTRACT
Motion intention can be detected from human Electroencephalography (EEG) signals through BCI, which can facilitate motor motion control for disabled or paralyzed people. However, the continuous use of BCI is hindered by the non-stationarity of the EEG signals. This paper proposes a method to identify the EEG signal components that can be used to train a classifier to address the non-stationarity issue. The proposed method is based on Transfer Component Analysis (TCA). TCA seeks to locate components that can be transferred across domains in a Reproducing Kernel Hilbert Space (RKHS). The distributions associated with data are closer to each other in the subspaces spanned by the identified transfer components. Therefore, typical machine learning techniques can be applied in the subspace spanned by these transfer components. This results in classifiers that can be trained on the source domain and tested on the target domain. The proposed Stationary Transfer Component Analysis (STCA) method is compared with Stationary Sub-space Analysis (SSA) on the BCI competition IV dataset 2a. The results show significant improvements over the baseline case and the results are better than those produced by SSA.

KEY WORDS
Brain-computer interfaces (BCI), Transfer Learning, Transfer Component Analysis (TCA), electroencephalography (EEG).

1. Introduction

A Brain–Computer Interface (BCI) facilitates online communication between the human brain and peripheral devices. BCI's allow users to by-pass the natural neural pathways to motor neurons and muscles which can be employed to communicate with locked-in patients [1]. Wolpaw has defined a BCI as, “a system that measures central nervous system activity and converts it into artificial output that replaces, restores, enhances, supplements, or improves natural central nervous system output and thereby changes the ongoing interactions between the central nervous system and its external or internal environment” [2].

A typical BCI system includes the acquisition of brain signals, the processing and classification of the acquired signals, the feedback of the interpreted brain state, and the use of the classified signals to perform a task. It has been found that execution or imagination of limb movements generate changes in rhythmic EEG activity known as sensorimotor rhythms (SMR) [3]. BCI based on SMR extract features and translate the changes in EEG associated with motor imagery tasks and use the resulting output to control BCI applications [4].

A major challenge in classification of EEG signals is their non-stationarity. The brain signals substantially vary after the initial calibration, such that a classifier trained on one session can rarely be reused in the next experimental session. Long periods of low performance have often been observed even when the classifier is trained with data obtained on the same day.

The models developed using machine learning techniques are based on the assumption that underlying distributions of features are more or less static. However, EEG data are inherently non-stationary due to various factors underlying the distributions change between training and testing sessions. Non-stationary of EEG signals has been identified to be caused by factors such as, changes in the physical properties of the sensors, variabilities in neurophysiological conditions, psychological parameters, ambient noise and motion artifacts. This non-stationarity impedes the continuous use of BCI, particularly for the disabled. Therefore BCI is a difficult and inspiring application area with respect to non-stationarity.

A few methods, such as, Bayesian transduction, active learning and distribution matching have been suggested to address the non-stationarity issue [5-8]. Stationary Subspace Analysis (SSA) [9] is another unsupervised learning method that finds subspaces in which data distributions stay invariant over time.

In this paper we explore the use of Transfer Component Analysis (TCA) to reduce the difference between the distributions of session to session data. TCA tries to learn a set of common transfer components underlying both sessions such that the difference in distributions of data in the different sessions, when projected onto this subspace, can be dramatically reduced [10]. It has been successfully applied in domain adaptation in text classification and wifi localization.

The rest of the paper is organized as follows. Section 2 provides a description of the STCA method and the pre-processing steps for EEG data. In section 3, the data and experimental paradigm are presented followed by comparative results and discussions in Section 4. In Section 5 the conclusions are drawn up with a brief discussion on the results.
2. Methods

The goal of the current study is to identify the components of the calibration session that can be transferred to subsequent sessions which are most similar to the distributions of the subsequent sessions of EEG data. To this effect, the typical domain adaptation paradigm described in [10] was applied after the Common Spatial Patterns (CSP) algorithm for feature selection in BCI to perform session-to-session adaptation.

2.1 Common Spatial Patterns

One of the crucial factors that affect the success of machine learning is proper preprocessing of data. Non-informative dimensions of the data can be discarded and the features of interest for classification can be selected through suitable preprocessing techniques [11].

The Common Spatial Patterns (CSP) procedure was first used to remove abnormal components in EEG [12]. It was later extended to classification of movement related EEG signals [11]. An arbitrary number of first and last CSP components which maximizes the differences among the variances are selected in this procedure. Modern implementations of CSP such as filter bank CSP [13] consider frequency bands separately to identify the most discriminant features for a given subject more effectively.

The main concept of CSP is to apply a projection matrix that linearly transforms the multichannel EEG data into a lower dimensional spatial subspace. In this projection matrix, each row consists of weights for each channel.

Let the random variable \( \tilde{x} \in \mathbb{R}^N \) represent the EEG data, recorded from N electrodes. Let the intention of the BCI user be \( C = \{c_1, \ldots, c_M\} \) which needs to be inferred. The class probability can be represented as \( P(c_i), i = 1, \ldots, M \). It is further assumed that EEG data follows a Gaussian distribution with zero mean, i.e., \( P(\tilde{x}|C_i) = N(0, \Sigma_{x|i}) \), \( i = 1, \ldots, M \) irrespective of the class on which it is conditioned. A transformation matrix \( W \in \mathbb{R}_{N \times L} \) with reduced dimensions \( L << N \), leading to the linear transformation \( \hat{x} = W^T x \) which maximize the variances of class signal matrices can be found.

Let the raw EEG data be represented as \( \tilde{x} \). The normalized spatial covariance for each of the two distributions to be separated, i.e., \( C = \{c_1, c_2\} \) can be calculated as

\[
C = \frac{xx^T}{\text{trace}(xx^T)},
\]

where \( \text{trace}(x) \) denotes the sum of the diagonal elements of \( x \) and \( T \) is the transpose operator. The CSP algorithm solves the optimization problem

\[
\hat{W}^* = \arg \max_{\hat{W}} \left\{ \frac{\text{trace}(R_{x|c_1} \hat{W})}{\text{trace}(R_{x|c_2} \hat{W})} \right\},
\]

where \( R_{x|c_1} \) and \( R_{x|c_2} \) are covariance matrices of \( x \) with given \( c_1 \) and \( c_2 \). Solutions to (2), which is in the form of Rayleigh quotient, can be found by solving the generalized eigenvalue problem,

\[
R_{x|c_1} \hat{W} = \lambda R_{x|c_2} \hat{W}.
\]

The eigenvectors of (3) correspond to the desired spatial filters. The corresponding eigenvalues determine the value of the cost function

\[
\lambda^* = \left( \frac{\text{trace}(R_{x|c_1} \hat{W})}{\text{trace}(R_{x|c_2} \hat{W})} \right).
\]

The ratio of the variance between conditions of the component of the EEG data extracted by the spatial filter is represented by the associated eigenvalue. The eigenvalue associated with a spatial filter can also be considered as an indicator of its quality. The \( L \) eigenvectors of (3) is combined with the largest/smallest eigenvalues to form the projection matrix \( W \in \mathbb{R}^{N \times L} \). The spatially filtered EEG signal \( \hat{x} \), is obtained by the projection \( \hat{x} = W^T x \), where \( x \) denotes the original EEG signal.

The spatially filtered data is then applied to transfer component analysis in order to identify the features that are more stationary from session to session.

2.2 Stationary Transfer Component Analysis

TCA algorithm has been typically developed to reduce the dissimilarity in the data distributions among the source domain and target domain in domain adaptation scenario. Domain adaptation seeks to adapt a regression model or classifier that has been trained on a source domain for operation on a target domain, where the source and target domains are different but related [10]. TCA is able to reduce the dissimilarity among different distributions pertaining to different domains while preserving important (geometric or statistical) properties of the original data [10]. TCA learns a set of transfer components which reduce the distance across domains in a Reproducing Kernel Hilbert Space (RKHS) [14].

In order to use TCA in a session to session transfer framework let the labelled data of the calibration session be \( D_s \) and the unlabelled data from a subsequent session be \( D_T \). Let \( D_s = \{(x_{s1}, y_{s1}), \ldots, (x_{sn}, y_{sn})\} \), where \( x_{si} \in X \) is the input and \( y_{si} \in Y \) is the corresponding output. Similarly, the data of the later session can be denoted as \( D_T = \{x_{t1}, \ldots, x_{tn}\} \), where the input \( x_t \in X \). \( Q(X_s) \) and \( R(X_t) \) denote the marginal distributions of \( X_s \) and \( X_t \) respectively. The key assumption is \( Q(X_s) \neq R(X_t) \), but \( P(Y_s|X_s) = P(Y_t|X_t) \). Based on the inputs \( x_{si} \) and outputs \( y_{si} \) from the calibration session, and the inputs \( x_t \) from the subsequent session, we attempt to predict the unknown outputs \( y_t \). We employ TCA to discover a hidden representation that is common for both \( X_s \) and \( X_t \) that conserves the data configuration of the two sessions after transformation.

Let \( \phi: X \rightarrow H \) be the nonlinear transformation from feature space to Reproducing Kernel Hilbert Space satisfying the condition \( Q^T(X_s^T) = R^T(X_t^T) \). If the transformed input from
source domain is $X_S^T = \{x_S^T\} = \{\phi(x_S)\}$, and the transformed target input is $X_T^T = \{x_T^T\} = \{\phi(x_T)\}$, the transformed input in the combined domain can be represented as $X^T = X_S^T \cup X_T^T$. The dissimilarity between the distributions of the two samples can be estimated based on the distance between the means of the two samples projected into RKHS [10,14]. This principle is used to empirically measure the distance between two distributions $Q$ and $R$ between the empirical means of the sessions as shown in equation (4).

$$
\text{Dist}(X_S^T, X_T^T) = \left\| \frac{1}{n_1} \sum_{i=1}^{n_1} \phi(x_S^i) - \frac{1}{n_2} \sum_{i=1}^{n_2} \phi(x_T^i) \right\|_2^2,
$$

where $x_S^i$ is from the first session and $x_T^i$ is from the subsequent session. The non-linear transformation $\phi: X \rightarrow H$ affected by the mapping $\phi$ can be found by minimizing equation (4).

The problem of searching for the non-linear transformation $\phi$ can be transformed to a kernel learning scenario as proposed in [15]. Based on the Kernel trick, $k(x_i, x_j) = \phi(x_i)^T \phi(x_j)$, the distance among the means of the two domains in (4) can be expressed as:

$$
\text{Dist}(X_S^T, X_T^T) = \text{tr}((KL)_L),
$$

where $K = \begin{bmatrix} K_{S,S} & K_{S,T} \\ K_{T,S} & K_{T,T} \end{bmatrix}$, (5)

is a $(n_1 + n_2) \times (n_1 + n_2)$ matrix. $K_{S,S}$, $K_{T,T}$, and $K_{S,T}$ are the kernels defined by $k$ on the data in the source domain, target domain, and cross domains respectively. $L = [L_{ij}] \geq 0$ with $L_{ij} = \frac{1}{n_1}$ if $x_i, x_j \in X_S$; $L_{ij} = \frac{1}{n_2}$ if $x_i, x_j \in X_T$ and $L_{ij} = -\frac{1}{n_1 n_2}$ otherwise.

A unified kernel learning method was proposed to find a nonlinear mapping $\phi$ based on kernel feature extraction in [10]. Following this method, the kernel matrix $K$ in (5) can be decomposed as, $K = \left( KK^{-1/2} \right) \left( K^{-1/2} K \right)$. A $(n_1 + n_2) \times m$ matrix $\bar{W}$ can be used to transform the features vectors to m-dimensional space, which is generally, $m \ll (n_1 + n_2)$.

The resulting kernel is $\bar{K} = \left( KK^{-1/2} \bar{W} \right) \left( \bar{W}^T K^{-1/2} \bar{W} \right) = KWKW^T K$, (6)

where $W = K^{-1/2} \bar{W} \in \mathbb{R}^{(n_1+n_2) \times m}$. The distance between the empirical means of the two domains can be expressed as:

$$
\text{Dist}(X_S^T, X_T^T) = \text{tr}((KWKW^T K)L) = \text{tr}(W^T KLKW),
$$

based on the definition of $\bar{K}$ in (6). A regularization term $\text{tr}(W^T W)$ is applied to limit the complexity of $W$. The problem of rank deficiency of the denominator in the generalized eigen decomposition is alleviated with this regularization term [10]. The optimization problem needed to extract the transfer components is reduced to:

$$
\min_W \text{tr}(W^T W) + \mu \text{tr}(W^T KLKW)
$$

s.t. $W^T KLKW = I$.

where $\mu$ is a trade-off parameter, $I \in \mathbb{R}^{m \times m}$ is the identity matrix, $H = I_{n_1+n_2} - \frac{1}{n_1+n_2} 11^T$ is the centering matrix, where $1 \in \mathbb{R}^{n_1+n_2}$ is the column vector consisting of ones, and $I_{n_1+n_2} \in \mathbb{R}^{(n_1+n_2) \times (n_1+n_2)}$ is the identity matrix. The constraint $W^T KLKW = I$ ensures that the trivial solution ($W = 0$) is avoided. More details of this algorithm can be found in [10].

### 3. Data and Experimental Procedure

The data set 2A of the fourth BCI Competition [16] was used to compare the proposed method. EEG data from 9 different subjects had been collected on two different days for this data set. Four motor imagery tasks: left hand, right hand, both feet and tongue, had been carried out in a cue-based manner during data collection. There had been 6 runs separated by short breaks in each session. 48 trials had been collected in each run, made up of 12 trials for each of the four imagery tasks. A total of 288 trials had been collected in each of the two sessions.

All the subjects had been comfortably seated comfortably in armchairs during the data collection. Subjects had been instructed on the different cue signals associated with the four motor imagery tasks and had been asked to continue the motor imagery task until the fixation cross sign had disappeared after 6 seconds. Subjects had not been instructed on any specific strategies for the motor imagery tasks. For example, a subject could have imagined movement of the fingers, imagine writing or any other movement of the specified hand. Imagination of feet movement could have been achieved by imagining any form of movements of the feet. Similarly, the tongue movement task could have involved imagination of any movement of the tongue. Further details about the data set can be found in [17].

The fixation cross had been first presented at the beginning of the trial along with short warning tones. The cue, indicating to start performing the specific motor imagery task had been presented two seconds after the fixation cross. The cue had been displayed on screen for 1.25 seconds. No feedback had been provided to the subjects during the data collection. Six such runs had been completed in each session which had been interleaved with short breaks. The timing scheme associated with this data collection is shown in the figure 1.

The data had been collected at a sampling rate of 250Hz and bandpass filtered in the range of 0.5-100Hz. A 50Hz...
notch filter had also been applied to remove line noise. The amplifier sensitivity had been set at 100μV. Signals had been acquired from 22 Ag/AgCl electrodes with inter-electrode distances of 3.5 cm, positioned mostly over the bilateral primary motor cortical areas. The collected signals were subject to again to a bandpass filter between 8Hz and 30Hz prior to applying the CSP algorithm for spatial filtering. 12 best features resulting after CSP transformation were input to TCA method to identify 4 transfer components as the final features to train an SVM classifier.

In order to compare how the proposed algorithm, the baseline case without TCA and application of Stationary sub-space analysis [9] method were also implemented. The dimensionality of stationary subspace was set at 20 after cross-validation studies. Before applying SSA the EEG data was bandpass filtered in the range of 8–30 Hz and the CSP features on the stationary subspace were subsequently computed. The number of iterations was limited to 5 due to the heavy computational requirements.

4. Results and Discussions

Results obtained using the proposed method and the SSA method are depicted in table 1. The baseline case where only CSP is applied for spatial filtering without STCA is considered as the baseline. The mean accuracies were compared with the Student’s t-test to find statistical significance. The results obtained for BCIC IV data set 2a are shown in table 1. The subjects are denoted as A1 to A9. All classifiers were trained on the training data obtained on the previous date and were evaluated on test data which had been collected on a subsequent date. The mean accuracies and standard deviations calculated for all the subjects are denoted as mean and S.D. in the table 1. Highest mean accuracy of 78.64% is achieved by the STCA method.

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

The mean accuracy from STCA based method is found to be significantly higher than the baseline. It should be noted that accuracies of 6 subjects show improvements over the baseline compared to only 3 subjects improving with SSA. However, for subject A4, the improvement achieved by SSA is higher than the improvement achieved using STCA based method.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Baseline</th>
<th>STCA</th>
<th>SSA</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>87.3</td>
<td><strong>91.4</strong></td>
<td>87.3</td>
</tr>
<tr>
<td>A2</td>
<td>56.8</td>
<td><strong>59.9</strong></td>
<td>56.9</td>
</tr>
<tr>
<td>A3</td>
<td>93.1</td>
<td>93.1</td>
<td>93.1</td>
</tr>
<tr>
<td>A4</td>
<td>63.6</td>
<td><strong>70.1</strong></td>
<td>70.5</td>
</tr>
<tr>
<td>A5</td>
<td>54.8</td>
<td><strong>59.6</strong></td>
<td>57.9</td>
</tr>
<tr>
<td>A6</td>
<td>62.6</td>
<td><strong>65.4</strong></td>
<td>62.6</td>
</tr>
<tr>
<td>A7</td>
<td>77.1</td>
<td>77.1</td>
<td>77.1</td>
</tr>
<tr>
<td>A8</td>
<td>94.2</td>
<td><strong>95.3</strong></td>
<td>94.2</td>
</tr>
<tr>
<td>A9</td>
<td>93.8</td>
<td>93.8</td>
<td>93.8</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td><strong>75.92</strong></td>
<td>78.64</td>
<td>77.04</td>
</tr>
<tr>
<td><strong>S.D.</strong></td>
<td>16.66</td>
<td><strong>15.17</strong></td>
<td>15.66</td>
</tr>
<tr>
<td><strong>P value</strong></td>
<td>-</td>
<td>0.013</td>
<td>0.19</td>
</tr>
</tbody>
</table>

Figure 2. Features before STCA
Figure 2 highlights the session to session non-stationarity present in data from subject A1 in a three-dimensional feature space. After applying STCA method, features that are stationary across the sessions are selected as illustrated in figure 3. The reduction in dimensionality by selecting a less number of features also helps to improve the classification accuracies. Therefore, STCA is able to address the non-stationarity present in the EEG data by transforming the features to a latent space where the stationary components can be identified.

The Results suggest that the proposed STCA method can significantly improve classification accuracies in most cases. However, it cannot be concluded that the STCA method is better than SSA as the overall improvement is not significantly different among the two methods. However, the computational time needed for STCA is far less than the time required for SSA. In a 2.83GHz processor with 4GB RAM STCA on average required less than 10000 seconds where as the SSA method needed more than 20 hours of processing time.

5. Conclusion

In this study STCA was applied in a BCI setting to identify transfer components to address the non-stationarity issue. The method was compared with the baseline case where the classifier trained on the calibration data is tested on data obtained on a later session. SSA method was also implemented for comparison.

Results indicate that STCA is effective in addressing non-stationarity and is faster than SSA. Further research is needed to combine the pre-processing and feature selection methods with the STCA method.

References


