ABSTRACT: Transfer Learning is a critical topic of research in the BCI field. Its goal is to reuse data gathered in a previous session (source session) in order to reduce, or completely bypass, calibration in a new session (target session). Although many methods have been proposed to tackle this problem, little is known about what characteristics of the datasets should be taken into account in order to ensure good performance. In this paper, we perform an exploratory analysis to study the influence of some simple descriptors of the source and target datasets over the classification scores obtained with Transfer Learning. We observe that the discriminability of the data points in the target session plays an important role in determining how well the Transfer Learning will work, as opposed to that of the source session, which has no statistically significant role in most cases.

INTRODUCTION

Reducing calibration time is an important challenge in Brain-Computer Interface (BCI) research [1]. Several Transfer Learning (TL) approaches have been proposed in the literature for doing so and most are based on the idea of using data from a previous recording session of a subject (the source session) to classify the data of a new session from the same subject or a different one (the target session). However, reusing data directly from previous sessions in general yields poor results. This comes from the fact that the statistical distributions of data from different sessions (same subject or not) are rarely the same [1].

A typical TL approach in BCI is to transform the data points from both the source and target datasets so that the discrepancy between their statistical distributions is reduced [2, 3]. In this paper, we match the statistics of the source and target datasets via the recently proposed Riemannian Procrustes Analysis (RPA) [2], a method that adapts the classical Procrustes analysis to a Riemannian geometry framework.

It is well known that, although any pair of source–target subjects can go through a Transfer Learning procedure, some pairs of subjects yield better results in classification than others. Our main goal in this paper is to investigate some factors that might explain this variability and how one could try to predict beforehand (i.e., before doing any matching of the datasets or classifying the data points) the "compatibility" between two datasets.

Our exploratory analysis relies on the estimation of linear models and the study of statistical significance of the coefficients estimated for those models. We use as explanatory factors the intra-scores for the source and target subjects (cross-validated classification score using the subject’s dataset as training and testing dataset), and the Maximum Mean Discrepancy (MMD [4]) between the two datasets, which is a common measure of discrepancy between statistical distributions. We observe that the intra-score for the target subject plays an important role in determining how well the Transfer Learning will work, as opposed to the intra-scores of the source subjects, which play no statistically significant role in most cases. We also observe that before doing any transformation on the data points of the source and target datasets, the MMD between their statistical distributions plays a statistically significant role over the performance of the Transfer Learning. However, once the RPA is applied, the MMD between the datasets becomes very small and no longer carries statistical information to describe the variability of the cross-subject scores. This confirms the relevance of the RPA method.

MATERIALS AND METHODS

This section begins with a formal definition of the Transfer Learning problem. Then, we give a brief introduction to concepts of Riemannian geometry and describe the RPA method. Finally, we present the statistical tools used in our exploratory analysis of Transfer Learning as well as the dataset chosen for our investigations.

Transfer Learning: We formulate the problem of Transfer Learning by first defining two datasets, the source (S) and the target (T) dataset. They are comprised of couples

\[ S = \left\{ (C_i^S, y_i^S) \mid i = 1, \ldots, K_S \right\}, \]
\[ T = \left\{ (C_i^T, y_i^T) \mid i = 1, \ldots, K_T \right\}, \]

with \( C_i^S \) and \( C_i^T \) being data points, and \( y_i^S \) and \( y_i^T \) their corresponding class labels; \( K_S \) and \( K_T \) are the number of trials in the source and target sessions respectively. In this paper, the data points in \( S \) and \( T \) are not Euclidean feature vectors as is usu-
ally done, but symmetric positive definite (SPD) matrices, which are used to parametrize the statistics of EEG multivariate time series [5].

Transfer Learning concerns the case when the statistical distributions $\mu_S$ and $\mu_T$, describing the source and target datasets respectively, are different. In this context, one might want to train a classifier $h$ using the information in $S$ and apply it to data points in $T$ (or vice-versa). A common approach is to define a transformation for the data points in the target dataset so that their new statistical distribution is the same as that of the source dataset. To do so, most algorithms define an optimization procedure that tries to minimize some notion of distance between the statistical distributions of $S$ and $T$, such as the MMD. In [6], a theoretical analysis of the Transfer Learning problem has shown that methods reducing the distance between statistical distributions are mathematically well justified, since they reduce the upper bounds of the classification error of $h$ in $T$.

Riemannian geometry of SPD matrices: We denote by $X_k\in\mathbb{R}^{n \times T}$ the recording of $T$ samples on $n$ electrodes of the $k^{th}$ trial of a zero-mean time series and $\mu_k$ the class associated to $X_k$. The spatial covariance matrix $C_k$ associated to $X_k$ is an $n \times n$ matrix estimated as usual by

$$C_k = \frac{1}{T-1}X_kX_k^T.$$  

Covariance matrices are symmetric positive definite (SPD) and form a manifold $P(n)$. When associated to a metric, one can define fundamental geometric notions in $P(n)$, such as geodesics (shortest curve joining two points), distance between two points (length of the geodesic connecting them), the center of mass of a set of points, etc. We endow $P(n)$ with the affine-invariance Riemannian metric, which induces the distance

$$\delta^2_{\rho}(C_i, C_j) = ||\log(C_i^{-1/2}C_jC_i^{-1/2})||_F^2,$$

for $C_i, C_j \in P(n)$. This distance is more natural for the $P(n)$ manifold as compared to the Euclidean distance and has been instrumental in several BCI classification algorithms developed in recent years [5]. The geometric mean $M$ according to distance (3) of a set of covariance matrices $\{C_1, \ldots, C_K\}$ is defined as [5]

$$M = \arg\min_{X \in P(n)} \sum_{k=1}^K \delta^2_{\rho}(X, C_k),$$

where the cost function in (4) is the dispersion of the set of matrices around a matrix $X$. The above definitions suffice for the intents of this paper. The interested reader will find a thorough treatment of the subject in the monography of R. Bhatia [7] and its applications to BCI in [5].

Riemannian Procrustes Analysis: In this paper, we use RPA [2] for transforming data points in $T$ so that their new distribution is as close as possible to $\mu_S$. RPA works by considering the distributions of data points in $S$ and $T$ as shapes in a high-dimensional space. It performs rigid geometric operations over the ensemble of data points, such as translation, stretching and rotation, to make their shapes as similar as possible (see Figure 1 for a visual representation of these operations). The transformations in RPA are done respecting the intrinsic geometry of $P(n)$, which is where the data points of $S$ and $T$ are defined. The steps involved in this procedure are summarized as follows:

1. Estimate the geometric means of $S$ and $T$, denoting them by $M_S$ and $M_T$, respectively, and the dispersions around the mean of each dataset, denoted by $d_S$ and $d_T$.

2. Re-center the data points in $S$ and $T$ by doing

$$C_i^{S(\text{rct})} = M_S^{-1/2}C_i M_S^{-1/2},$$

$$C_i^{T(\text{rct})} = M_T^{-1/2}C_i M_T^{-1/2},$$

and forming new datasets $S^{(\text{rct})} = \{C_i^{S(\text{rct})}\}$ and $T^{(\text{rct})} = \{C_i^{T(\text{rct})}\}$.

Note that the geometric mean of these two new datasets is the Identity matrix.

3. Stretch the dispersion around the mean of the data points in $T^{(\text{rct})}$ so that it matches the dispersion around the mean of $S^{(\text{rct})}$ by creating new data points

$$C_i^{T(\text{str})} = \left(C_i^{T(\text{rct})}\right)^s,$$

where we require $s \in \mathbb{R}$ to verify

$$s^2 = d_S/d_T.$$
4. The last step consists in rotating the matrices from $T^{(m)}$ around the origin and matching the orientation of its point cloud with that of $S^{(m)}$. We have then

$$C_i^{T(m)} = U^T C_i^{T(m)} U,$$

where $U$ is determined via an optimization procedure that minimizes the distance between the class means of each dataset after the rotation. Note that this step is a semi-supervised one, since it requires knowledge of at least a few labels of the target dataset for estimating its class means (see [2] for more details).

_Layer Transfer Learning classification:_ In this paper, whenever we want to do a classification task with data points that live in $P(n)$, we use the Minimum-Distance to Mean classifier (MDM) [1, 5, 8], which is a generalization of the nearest-centroid classifier to the space of SPD matrices. It works by first estimating the geometric mean of the elements of each class in the training dataset (the class means). Then, it assigns to each unlabeled data point the class of the nearest class mean according to the $\delta_R$ distance. In the context of transfer learning, we will always consider the source dataset $S$ as the training dataset and the target dataset $T$ as the testing dataset. The classification score is simply the accuracy of the classifier.

In the following analysis, we will consider the results of Layer Transfer Learning classification on three different cases:

- **DCT:** the source and target datasets are used directly as training and testing datasets, that is, without any transformation.
- **RCT:** the source and target datasets are both re-centered and then used as training and testing dataset.
- **RPA:** the source and target datasets go through the full RPA procedure (re-centering + stretching + rotating) and are then used as training and testing datasets.

For this analysis we will assume that all labels in the target dataset are available for the estimation of the class means. In fact, our intent here is not to evaluate the performance of the RPA method in a realistic situation when only a few labels from the target dataset are available (this has been done in [2]), but rather to understand how its performance might be influenced by other factors.

_Seriation procedure:_ Given a dataset, all cross-subject TL scores are summarized in a matrix $S^{(m)}$, where the $S_{ij}^{(m)}$ element contains the accuracy of the classification with method $m \in \{\text{DCT, RCT, RPA}\}$ using subject $i$ as target and subject $j$ as source. We use a tool from combinatorial data analysis named seriation [9] to rearrange the lines and columns of $S^{(m)}$ in order to make relevant patterns emerge. The rows and columns of $S^{(m)}$ are sorted in decreasing order of their marginals. The output of this procedure is a new representation where the pairs of source-target subjects with the best accuracy are located at the top-left region of the matrix, while the worst pairs are at the bottom-right region.

_Statistical analysis procedure:_ Our quantitative analysis is based on the estimation of linear regression models to describe the variability on the values of $S_{ij}^{(m)}$ as defined in the previous subsection. We estimate a different linear model $L_i^{(m)}$ for each target subject $i$ and method $m$. We do this because the cross-subject scores for two target subjects with the same source subject are statistically dependent, which would undermine the estimation of a linear model mixing all scores from all source-target pairs. Moreover, the results after the RPA method are related to those for the RCT one, since the latter includes the former as a processing step.

We define the linear model $L_i^{(m)}$ for target subject $i$ on method $m$ as:

$$S_{ij}^{(m)} = \beta_{1,i}^{(m)} S_i + \beta_{2,i}^{(m)} S_j + \beta_{3,i}^{(m)} \eta_{ij}^{(m)} + \epsilon_i^{(m)},$$

where

- $S_i$ ($S_j$) is the intra classification score of target (source) subject $i$ ($j$), obtained via cross-validation with training and testing datasets coming from the same subject. Note that since each model $L_i^{(m)}$ is estimated for one fixed target subject $i$, $S_i$ is a constant in (10) and acts as a scaling for the intercept; thus, it is not considered as an independent variable in the statistical analysis.
- Factor $\eta_{ij}^{(m)}$ is the MMD between datasets $S$ and $T$ after the operations of method $m$, defined as [4]

$$MMD(S,T) = \frac{1}{K_S^2} \sum_{i,j} k(C_i^S, C_j^S) + \frac{1}{K_T^2} \sum_{i,j} k(C_i^T, C_j^T) - \frac{2}{K_S K_T} \sum_{i,j} k(C_i^S, C_j^T),$$

where

$$k(P,Q) = \exp \left( -\frac{\delta_R^2(P,Q)}{2\sigma^2} \right),$$

for $P, Q \in P(n)$ and $\sigma$ is taken as the median value of all pairwise distances of elements in $S$ and $T$. $K_S$ and $K_T$ are defined in (1).

- The variable $\epsilon_i^{(m)}$ stands for all residual factors that are not explained by the linear regression model.

Once the linear models are all estimated, we perform a set of hypothesis tests for each target subject $i$. The goal is to assess the statistical significance of the coefficients of each model. The first kind of test is a $F$-test for the omnibus null hypothesis:

$$H_0 : \beta_{2,i}^{(m)} = \beta_{3,i}^{(m)} = 0,$$

$$H_1 : \beta_{k,i}^{(m)} \neq 0 \text{ for at least one } k \in \{2,3\}.$$
This is a standard test used for inspecting whether the set of independent variables of a linear regression model, $S_j$ and $\delta_{ij}$ in (10), is statistically significant for explaining at least part of the variability of the dependent variable, $S_{ij}^{(m)}$ in (10). When the null hypothesis is rejected, we say that there is enough statistical evidence for considering that the slope of at least one of the independent variables is different than zero. In this case, we perform t-tests for checking which explanatory variable in $L_i^{(m)}$ is statistically significant. We have:

$$\mathcal{H}_0 : \beta_{\ell,i}^{(m)} = 0,$$

$$\mathcal{H}_1 : \beta_{\ell,i}^{(m)} \neq 0,$$

for $\ell \in \{2,3\}$. When the null hypothesis of (14) is rejected for $\beta_{\ell,i}^{(m)}$, we say that there is statistical evidence for considering it different than zero and so the independent variable related to it contributes for explaining the dependent variable $S_{ij}^{(m)}$.

The statistical procedure described above yields two sets of $p$-values for each method $m \in \{DCT, RCT, RPA\}$. The first set contains the $p$-values for each $F$-test on each target subject $i$, whereas the second set gathers the $p$-values of the t-tests. The results presented in the next section are based on the analysis of these sets of $p$-values and how they are distributed along different source subjects for each method.

**Dataset:** We carried out our analysis on a publicly available dataset [10] which we will refer as Cho2017 from now on. The dataset contains recordings of subjects performing BCI trials following a Motor Imagery (MI) paradigm with 64 EEG electrodes (sampling frequency 512 Hz) from 52 subjects, each one performing 200 trials (100 of each class). We filtered the EEG signals in the 8-30 Hz band and each trial was considered as a segment from 0.5 to 2.5 seconds after the trial onset. We estimated the spatial covariance matrices using (2). Not all subjects in Cho2017 display data which can be well discriminated, so we kept only those with intra-score in terms of AUC (Area Under the ROC-curve) above chance level; this criterion retains 40 subjects out of the 52 in total.

**RESULTS AND DISCUSSION**

In this section, we present the results of our analysis of TL via RPA on the Cho2017 dataset. We begin with a qualitative analysis of the output of the seriation procedure applied to the cross-subject scores. Then, we study the correlation of each factor defined in (10) with the cross-subject scores. Finally, we analyse the results of the statistical hypothesis tests for the linear models $L_i$ and discuss the patterns observed for the whole dataset (from now on, we will indicate the superscript specifying the method $m$ only when necessary).

**Cross-subject classification accuracy:** Figure 2 shows the output of the seriation procedure on the cross-subject TL scores for the Cho2017 dataset on the three classification methods: DCT, RCT, and RPA. We observe that with RCT and RPA there are more pairs of subjects with high values of cross-subject classification than with DCT. In particular, we note that for RCT and even more for RPA there are many target subjects for which the classification accuracy is high for almost all possible source subjects. To investigate the possible explanations for this behavior, we compute the Spearman correlation between the average cross-subject score for each target (i.e., the average value along the rows of matrix $S$) and the intra-subject accuracy of the corresponding target subject. For the RPA method, we obtain a correlation of 0.58 ($p < 10^{-3}$), for RCT it is 0.44 ($p < 10^{-2}$) and for DCT is 0.45 ($p < 10^{-2}$). We interpret these results as: subjects that are “good” for classifying their own data can better receive information from other source subjects. We also provide a quantitative analysis of the results. Figure 3 portrays the histograms of all cross-subject Transfer Learning scores $S_{ij}$ (rows and columns confounded) for each method. Their means are reported in Table 1. These results show that the transformations over the source and target datasets do improve the cross-subject classification scores on the average for RCT and greatly do so for RPA.

<table>
<thead>
<tr>
<th>Method</th>
<th>$(\eta_{ij})_{avg}$</th>
<th>$(S_{ij})_{avg}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCT</td>
<td>0.63</td>
<td>0.53</td>
</tr>
<tr>
<td>RCT</td>
<td>0.01</td>
<td>0.58</td>
</tr>
<tr>
<td>RPA</td>
<td>0.01</td>
<td>0.76</td>
</tr>
</tbody>
</table>

**Changes in MMD after each RPA step:** We evaluate how the MMD between each pair of source–target subjects changes after the re-centering step and the full RPA procedure. Table 1 gives the average values of the MMD distances for each method and shows that there is a clear decrease after each transformation. This result is not surprising, since each step of the RPA procedure was conceived exactly to make the distributions of $\mathcal{S}$ and $\mathcal{T}$ closer in some sense and the MMD allows for a quantitative assessment of it.

**Study of the linear models $L_i$:** After exploring the grand averages of the cross-subject TL scores and how they relate to a few explanatory factors, we analyse the linear models $L_i$ defined in (10) and estimated on each target subject $i$ for the three methods of interest: DCT, RCT, and RPA.

We first plot the $p$-values of the $F$-test for each model sorted in ascending order. Notice that, under the omnibus null hypothesis for all target subjects, the $p$-values follow an uniform distribution, thus, when sorted they will lie on a straight line. The leftmost plot in Figure 4 shows that for almost all subjects the variability of the cross-subject performance is well explained by the linear model estimated for the DCT method and, in a lesser extent, for the RCT method. For RPA the $p$-values lie all very close to...
Figure 2: Accuracies of the cross-subject classification for three different Transfer Learning procedures on the Cho2017 database. The rows and columns of each subplot were reordered using the *seriation* procedure explained in the text. The grayscale varies from white (accuracy 0.5) to black (accuracy 1.0).

Figure 3: Normalized histograms of the cross-subject Transfer Learning scores for the three methods described in the text. The vertical dashed line indicates chance level.

Figure 4: $p$-values of different statistical tests over the linear models $L_i$ (each one associated to a target subject $i$). Each circle represents the $p$-value of a given test on a given target subject and the $x$-axis has been rearranged so that all the $p$-values are in increasing order. The leftmost plot represents the results of the $F$-test of the full linear model $L_i$, whereas the center plot illustrates the $p$-values for the $t$-test of the coefficient $\beta_{2,i}$ in $L_i$ (related to the intra-score of the source subject), and the rightmost plot displays the $p$-values for the $t$-test on the coefficient $\beta_{3,i}$ in $L_i$ (related to the MMD between the source and target datasets).
is why we have calculated the Spearman source Learning scores in the case when all different factors on the variability of cross-subject Transfer Learning procedure should take into account other aspects of the mismatch between datasets besides the MMD between them.

CONCLUSION

In this paper, we have investigated the influence of different factors on the variability of cross-subject Transfer Learning scores in the case when all source and target labels are known. Our goal has been to assess whether some basic explanatory variables, such as the intra-score of the source and target subjects, play any role for determining the scores obtained in the cross-subject classification. A simple, and yet important, application of this study is being able to predict beforehand (i.e., before doing all transformations and then classifying the trials) which source subject would be the most appropriate for doing classification on a given target subject.

We have observed that the discriminability of the trials of the target subjects plays a fundamental role in determining how the cross-subject Transfer Learning will perform. On the other hand, the influence of the intra-scores for the source subjects have proven to be rather limited, as seen by the lack of statistical significance for the values of $\beta_{2,i}$ in most cases. We have also observed that the influence of the MMD between $S$ and $T$ is not statistically significant after using RPA to match the two datasets. What we can conclude from this is that the RPA procedure is capable of factoring out most of the influence of the discrepancy between statistical distributions of the source and target datasets.

It is our opinion that in order to devise new and more powerful strategies for Transfer Learning it is imperative to investigate which factors determine its success. The present study is a little step in this direction. Future work may include the search for richer models for describing the variability of cross-subject TL scores. One approach would be to consider non-linear relations between the explanatory variables as well as adding new factors related to other features of the source and target subjects.

REFERENCES


