A SUBJECT-TO-SUBJECT TRANSFER LEARNING FRAMEWORK BASED ON JENSEN-SHANNON DIVERGENCE FOR IMPROVING BRAIN-COMPUTER INTERFACE

Joshua Giles[1,2]*, Kai Keng Ang*[2,3]*, Lyudmila S. Mihaylova*[1]*, Mahnaz Arvaneh*[1]

[1]Automatic control and System Engineering Department, University of Sheffield, UK
Emails: jgiles1,l.s.mihaylova, m.arvaneh@sheffield.ac.uk
[2]Institute for Infocomm Research, ASTAR, Singapore
[3]School of Computer Science and Engineering, Nanyang Technological University

ABSTRACT

One of the major limitations of current electroencephalogram (EEG)-based brain-computer interfaces (BCIs) is the long calibration time. Due to a high level of noise and non-stationarity inherent in EEG signals, a calibration model trained using limited number of train data may not yield an accurate BCI model. To address this problem, this paper proposes a novel subject-to-subject transfer learning framework that improves the classification accuracy using limited training data. The proposed framework consists of two steps: The first step identifies if the target subject will benefit from transfer learning using cross-validation on the few available subject-specific training data. If transfer learning is required a novel algorithm for measuring similarity, called the Jensen-Shannon ratio (JSR) compares the data of the target subject with the data sets from previous subjects. Subsequently, the previously calibrated BCI subject model with the highest similarity to the target subject is used as the BCI target model. Our experimental results using the proposed framework obtained an average accuracy of 77% using 40 subject-specific trials, outperforming the subject-specific BCI model by 3%.

1. INTRODUCTION

Electroencephalogram (EEG)-based brain computer interfaces (BCI) are systems which allow for direct communication between a person and a machine using only the brain waves produced by the user [1]. This form of communication can allow many people who are unable to communicate otherwise, due to damage in their neural pathways [2], to obtain some control of their environment. Recently BCI has started to be implemented to assist with rehabilitation of stroke patients [3].

Despite several recent advances, there are still a number of major issues with the current BCI which need to be addressed. One of these issues is the fact that currently 20% to 25% of the users are unable to achieve the classification accuracy of 70% or more while using the BCI [4]. The other major issue is that even for the users who can obtain high levels of accuracy, they typically require a 20 to 30 minutes calibration period at the beginning of each session [5]. During this calibration period a large number of labelled training data are recorded for adjusting the feature extractor and the classifier. This is necessary to adapt the BCI to the target user and deal with the variations in the EEG signals, both from subject to subject and from session to session. These long calibration periods cause fatigue and stress for the users, limiting the time available to use the BCI as intended. As such improving the accuracy that a BCI can achieve while reducing the training trials required is an important area of research.

In order to improve the accuracy, research has been carried out to improve the components within the BCI [6]. These range from feature extractors, such as the filter-bank common spatial patterns algorithm [7], to classifiers, such as the adaptive linear discriminant analysis (aLDA) which updates the classifier parameters when new trials are available [8]. A range of other adaptation methods have also been explored to further improve the classification accuracy possible. An example being data space adaptation which reduces the difference between the training data and test data through a linear transform [9] [10]. Despite these techniques improving accuracy, they often require a large number of calibration trials to provide a significant improvement.

To reduce the need for the long calibration time, transfer learning between sessions and subjects has been investigated. Transfer learning often refers to a procedure of using a data set from a different task to improve the accuracy of a related task [11]. When used for BCI, the data sets are often from the same task but different users. One form of transfer learning is through identifying features which are stationary across multiple subjects, known as domain adaptation. This area has been explored by Lotte and Guan [12] and Kang et

al [13] [14] with some success. The other form of transfer learning is called rule adaptation which attempts to find the framework of classification rules. The rule adaptation-based transfer learning attempts to select the most appropriate feature extraction and classification rules from a pool of available components [11]. This area of transfer learning has not been explored much within BCI although it has been explored by He and Wu recently [15].

In this paper first a new measurement of similarity is proposed named the Jensen Shannon ratio (JSR). This measure is used to compare calibration trials with existing data sets for transfer learning. Then a framework is proposed which identifies whether the target user will benefit from using rule adaptation transfer learning. If so, the data set with the highest similarity to the trials of the target user is selected, from previously recorded data, for training a BCI model for the target user.

The proposed framework will be evaluated using the publicly available BCI Competition IV data set 2a [16]. The algorithm will then be compared to utilizing only the Kullback Leibler (KL) divergence for data set selection and a framework previously proposed by Lotte using other subjects data to alter the co-variance and mean of the training data set [17].

2. METHODOLOGY

2.1. The Proposed Jensen Shannon Ratio

The proposed JSR measures the difference of the average EEG signals from the same class between users and the opposing classes using the Jensen Shannon divergence. The JSR is then used to select an appropriate signal for training, where the same classes are similar and opposing classes are far apart. The Jensen Shannon divergence is based on the Kullback Leibler (KL) divergence with some useful differences.

$$\begin{aligned}\mathrm{KL}[N_j \parallel D_j] = \frac{1}{2}[(\overline{\boldsymbol{\mu}}_j - \boldsymbol{\mu}_j)^T \overline{\boldsymbol{\Sigma}}_j^{-1}(\overline{\boldsymbol{\mu}}_j - \boldsymbol{\mu}_j) \\ +\mathrm{tr}(\overline{\boldsymbol{\Sigma}}_j^{-1}\boldsymbol{\Sigma}_j) - \ln(\frac{det(\boldsymbol{\Sigma}_j)}{det(\overline{\boldsymbol{\Sigma}}_j)}) - k],\end{aligned} \tag{1}$$

The band pass filtered EEG signals can be modelled as Gaussian distributions. The similarity between two Gaussian distributions can be measured through the KL divergence, as shown in (1). For this equation $N_j(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $D_j(\overline{\boldsymbol{\mu}}, \overline{\boldsymbol{\Sigma}})$ are used to represent the distributions of class j from the target subject N and training subject D. $\overline{\boldsymbol{\mu}}$ and $\boldsymbol{\mu}$ represent the means of the distribution, and $\overline{\boldsymbol{\Sigma}}$ and $\boldsymbol{\Sigma}$ denote covariances.

Jensen Shannon divergence is an extension of the KL divergence. This extension provides a symmetric and finite value for the similarity by measuring to a middle point providing, as shown in (2). The middle point $M_{ji}(\boldsymbol{\mu}_{ji}, \boldsymbol{\Sigma}_{\mathbf{ji}})$ is calculated from the average of the two distributions being compared, with $\boldsymbol{\mu}_{ji} = 0.5(\boldsymbol{\mu}_j + \overline{\boldsymbol{\mu}}_i)$ and $\boldsymbol{\Sigma}_{ji} = 0.5(\boldsymbol{\Sigma}_j + \overline{\boldsymbol{\Sigma}}_i)$. This Jensen Shannon divergence is then used to calculate the JSR and select the best data sets for the test data.

$$\mathrm{JS}[N_j \parallel D_i] = \frac{1}{2}(\mathrm{KL}[N_j \parallel M_{ji}] + \mathrm{KL}[D_i \parallel M_{ji}]) \tag{2}$$

Through knowing the differences between the EEG signals for each subject and class the JSR can be calculated. This aims to select a data set which has similar distributions for the same class while ensuring that the opposing classes are not similar. This is done through equation (3), with C representing the number of classes. The JSR aims to minimize the dissimilarity between the classes of two data sets while simultaneously maximizing the dissimilarity between different classes.

$$JSR = \frac{\sum_{j=1}^{C} \mathrm{JS}[N_j \parallel D_j]}{\sum_{i=1 i\neq j}^{C}(\mathrm{JS}[N_j \parallel D_i])} \tag{3}$$

When using the JSR for BCI subject to subject transfer learning the band-pass filtered EEG signals are used. As such $D_i(0, \overline{\boldsymbol{\Sigma}}_i)$ can be used to represent the distribution of one of the training data sets. While $N_j(0, \boldsymbol{\Sigma}_j)$ represents the distribution of the few subject specific trials we have from the user for each j class. In each of these distributions the normalized co-variance is estimated through the signal values x as shown in (4), with N number of trials.

$$\boldsymbol{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathbf{x_i}\mathbf{x_i}^T}{tr(\mathbf{x_i}\mathbf{x_i}^T)} \tag{4}$$

As the band pass filtered EEG has a zero mean, equation (3) can be simplified to equation (5). Once the JSR has been calculated between the distribution of the subject specific trials and each of the possible training data distributions, the training data with the lowest JSR value is then selected. This data set is used to train the CSP and LDA of the BCI.

$$\mathrm{JSR} = \sum_{j=1}^{C}\sum_{i\neq c}\frac{(\mathrm{tr}(\mathbf{M})^{-1}\overline{\boldsymbol{\Sigma}}_j + (\mathbf{M})^{-1}\boldsymbol{\Sigma}_j) - \ln(\frac{\det(\overline{\boldsymbol{\Sigma}}_j)}{\det(\boldsymbol{\Sigma}_j)})}{C(\mathrm{tr}(\mathbf{M})^{-1}\overline{\boldsymbol{\Sigma}}_i + (\mathbf{M})^{-1}\boldsymbol{\Sigma}_j) - \ln(\frac{\det(\overline{\boldsymbol{\Sigma}}_i)}{\det(\boldsymbol{\Sigma}_j)})} \tag{5}$$

2.2. Proposed BCI Subject to Subject Transfer Learning Framework

Users who encounter BCI deficiency can benefit substantially from the application of transfer learning. While for other users, who easily obtain high classification accuracy, the transfer learning can be detrimental. To counter this the proposed framework identifies the users who can benefit from subject to subject transfer learning then selects the best previously recorded data set for these users to train their BCI models. To identify the subjects requiring transfer learning

the leave-one-out validation (LOOV) accuracy is applied on the few subject-specific target trials. If the average accuracy for those subject-specific trials is below 70% they are identified as BCI deficient. For user who are found to encounter BCI deficiency the proposed JSR was then used to select an appropriate data set for training the BCI.

2.3. Selection Comparison

To evaluate the effectiveness of the proposed JSR transfer learning, its results are compared with the accuracies obtained using a KL based similarity measure. Moreover, the proposed framework is compared to the algorithm previously suggested by Lotte utilizing other subjects training data [17]. These were also compared to training the BCI with the available subject specific trials provided to highlight the improvement in accuracy achieved by providing the additional training trials.

2.3.1. Kullback Libeler Divergence

KL divergence is a long established method of calculating the difference between two Gaussian distributions. As such it is used for comparison against the JSR as a mean of transfer learning in the data domain. Equation (1) displays the calculations required to calculate the KL divergence. In this the data set which has the lowest summation of KL divergence between the test and target subjects classes is used for the BCI training.

2.3.2. BCI utilizing other subjects data

Lotte and Guan previously developed an algorithm for BCI which used other subjects data to reduce the need for calibration trials [17]. This evaluates the training data by training a BCI using each of the training data sets available. The subject specific trails are then used to evaluate the data sets. The selected data sets are then weighted, with λ, then used to estimate a new co-variance and mean in the feature domain as shown (7) and (6). For these equations μ and $\mathbf{\Sigma}$ are the mean feature vector and co-variance of the target subject while $\overline{\boldsymbol{\mu}}$ and $\overline{\mathbf{\Sigma}}$ are the mean feature vector and co-variance of the training subset. s is the number of selected training data sets.

$$\mathbf{\Sigma} = (1-\lambda)\mathbf{\Sigma} + \lambda\frac{1}{s}\sum_{i=1}^{s}\overline{\mathbf{\Sigma}}_{\mathbf{s}} \tag{6}$$

$$\boldsymbol{\mu} = (1-\lambda)\boldsymbol{\mu} + \lambda\frac{1}{s}\sum_{i=1}^{s}\overline{\boldsymbol{\mu}}_{s} \tag{7}$$

$$\lambda = \frac{DatasetAccuracy - SubjectSpecificAccuracy}{100 - ChanceAccuracy} \tag{8}$$

The weighting of λ is calculated through comparing the leave-one-out validation (LOOV) accuracy that is achieved by the subject specific trials and the accuracy achieved when the other data sets are used for training. If the leave one out validation outperforms the other data sets it is used for training the BCI, while if it is less than chance the trials are not used at all. If the LOOV accuracy is between the chance level and the accuracy achieved by the other data sets then they are weighted as shown in (8).

3. RESULTS AND DISCUSSION

3.1. Improvement for BCI Deficient Users

Initially the proposed JSR is compared to the other transfer learning algorithms. The JSR allows BCI deficient users to achieve higher accuracy then any of the other algorithms. This can be seen in figure 1 which shows the accuracy achieved by each of the algorithms when 8 subject specific trials are available. For the users who achieved less than 70% accuracy with their subject specific trials the average improvement was 8% with JSR. In comparison the algorithm proposed by Lotte improved the accuracy for these subjects by just 3% and the KL divergence caused a fall in accuracy. Subjects 1 and 5 in particular had a large increase in classification accuracy when the JSR was applied. While the average accuracy across all the subjects is not improved by JSR, compared to the standard BCI. This could be improved with a larger data base with more subjects to select from.

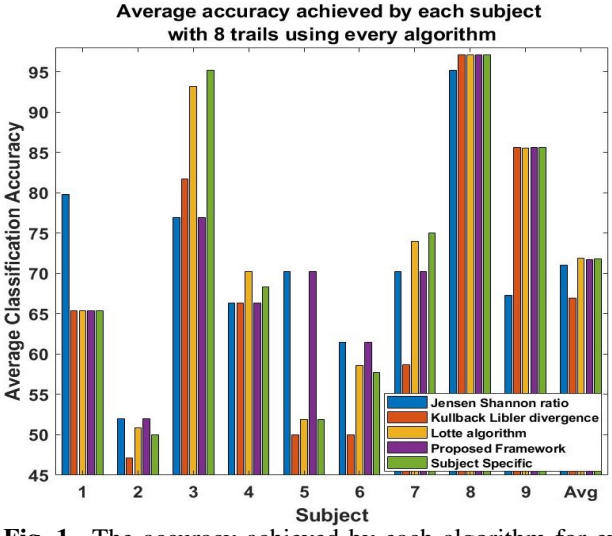

Fig. 1. The accuracy achieved by each algorithm for every subject in the data set when only 8 trials are available for either training or calibration.

When examining the average classification accuracy across all the subjects, when 8 subject specific trials are available, the algorithm proposed by Lotte outperforms the JSR 0.9%. This may be due to Lottes algorithm only using

others data for users encountering BCI deficiency, who require assistance, while JSR was used for all subjects. The subjects able to achieve high levels of accuracy with only a few subject specific trials lose accuracy with any of the other data sets available in the database. As such it is important to differentiate between the subjects who will achieve high accuracy and the subjects who will encounter BCI deficiency.

To identify these BCI deficient subjects the proposed framework incorporates the LOOV accuracy as a quick way to estimate the users competency in controlling EEG based BCI. Through this the users are classified as either BCI deficient or sufficient. The users will then either use the JSR to select the best training set for them or use the subject specific trials as the training set for the BCI.

Using the framework to select the subjects requiring transfer learning, before applying the JSR, improves the average accuracy across all the subjects. This is shown in figure 2 where the proposed framework is able to achieve 77% accuracy when 40 subject specific trials are available. When only the subject specific trials are used for training the average accuracy is only 74.5%. The proposed framework consistently outperforms the standard BCI although it does not perform optimally initially and experiences a small decrease in accuracy when 28 trials are available. The drop in accuracy which occurs when there are 28 subject specific trials is due to subjects 2 and 8 both experiencing a fall in accuracy. These subjects are both correctly identified as BCI deficient and sufficient respectively however still lose accuracy due to a few inconsistent trials. These trials causes the JSR to select a bad data set for subject 2 and leading to a fall in accuracy of 2%. This shows that the framework could benefit from an algorithm to evaluate and remove trials that are outliers.

3.2. Average Improvement from Proposed Framework

As mentioned the framework is not able to improve the average accuracy when only 8 subject specific trials are available. The LOOV misidentifies subjects 1 and 3 lowering the average accuracy by 0.1% compared to the standard BCI. The initially low accuracy of the proposed framework highlights one of the main problems which is its ineffectiveness in noticing BCI deficient users quickly. Using the LOOV accuracy is able to produce a fairly accurate prediction of the users capabilities when enough trials are available however a few outlying trials can affect the results. These outliers are not necessarily just trials that produce low levels of accuracy but can also produce uncharacteristically high levels of accuracy which lead to the subjects being miss-classified by the LOOV and the framework under performing. The LOOV does perform well when there are enough trials provided to the validation and the framework does still improve on the standard BCI when 10 or more trials are available.

Table 1 highlights this failing of the LOOV accuracy as a measurement of BCI competency. The proposed framework improves upon standard BCI however there is still a lot of room for improvement. If a better selection method was available this could further improve the accuracy of the framework. This increase in accuracy could be up to 3% if the correct subjects are selected to utilize the JSR. The current framework is able to improve the accuracy for BCI deficient subjects by over 5.5% when only 8 trials are available and by up to 6.5% when 40 trials are available.

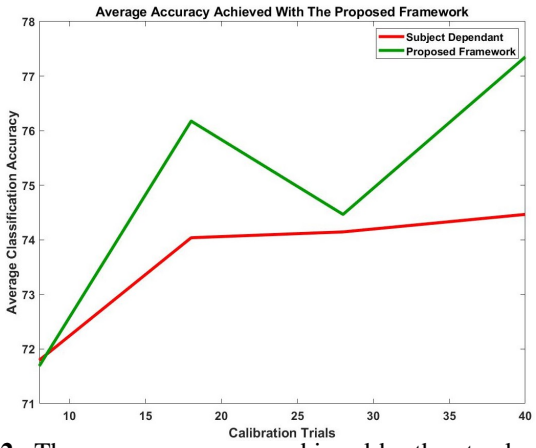

Fig. 2. The average accuracy achieved by the standard BCI and framework improves as the number of trials increase.

Table 1. Average accuracy for BCI deficient users

Trials	Subject Specific	Proposed framework	JSR
8	57	**62.5**	**62.5**
40	58.65	65.1	**66.1**

4. CONCLUSION

Overall an improvement in the classification accuracy was consistently achieved for users encountering BCI deficiency by the proposed Jensen Shannon ratio data selection. When the "leave one out" method was used to select the users who required alternative training trials the average accuracy of the system outperformed the standard BCI by 3%. It is also important to remember that this was conducted using a publicly available data set with only nine subjects, providing a relatively small amount of training data to select from. A larger data set with more subjects may be able to find more appropriate data sets for each deficient subject. A number of users who were miss-classified could have benefited from using a different users data. As such to progress this work a key area to focus on will be in selecting a better predictor of classification accuracy going forward. This could potentially improve the systems allowing it to achieve an accuracy of 77% with only 8 trials. As the framework improves it can be applied to assist stroke patients with rehabilitation.

5. REFERENCES

[1] T. W. Berger, J. K. Chapin, G. A. Gerhardt, D. J. McFarland, J. C. Principe, W. V. Soussou, D. M. Taylor, and P. A. Tresco, *Brain-Computer Interfaces*. Dordrecht: Springer Netherlands, 2008.

[2] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-computer interfaces for communication and control.," *Clinical Neurophysiology : Official Journal of the International Federation of Clinical Neurophysiology*, vol. 113, no. 6, pp. 767–91, 2002.

[3] K. K. Ang, S. Member, C. Guan, and S. Member, "EEG-Based Strategies to Detect Motor Imagery for Control and Rehabilitation," vol. 25, no. 4, pp. 392–401, 2017.

[4] C. Vidaurre, C. Sannelli, and M. Klaus-robert, "Co-adaptive calibration to improve BCI," *Journal of Neural Engineering*, 2011.

[5] M. Krauledat, M. Schröder, B. Blankertz, and K.-R. Müller, "Reducing calibration time for brain-computer interfaces: A clustering approach," *Adv. Neural Information Processing Syst.*, vol. 19, p. 753, 2007.

[6] W. Wu, X. Gao, and S. Gao, "One-versus- the-rest (OVR) algorithm: An extension of common spatial patterns (CSP) algorithm to multi-class case," *Engineering In Medicine And Biology*, vol. 3, pp. 2387–2390, 2005.

[7] 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, no. MAR, pp. 1–9, 2012.

[8] C. Vidaurre, M. Kawanabe, P. Von Bünau, B. Blankertz, and K. R. Müller, "Toward unsupervised adaptation of LDA for brain-computer interfaces," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 3 PART 1, pp. 587–597, 2011.

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

[10] J. Giles, K. K. Ang, L. Mihaylova, and M. Arvaneh, "Data Space Adaptation for Multiclass Motor Imagery-based BCI," *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2018.

[11] V. Jayaram, M. Alamgir, Y. Altun, and M. Grosse-wentrup, "Transfer Learning in Brain-Computer Interfaces," *arXiv:1512.00296*, no. February, pp. 20–31, 2016.

[12] 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.

[13] H. Kang, Y. Nam, and S. Choi, "Composite Common Spatial Patterns for Subject-to-Subject Transfer," *IEEE Signal Processing Letters*, vol. 16, no. 8, pp. 683–686, 2009.

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

[15] H. He and D. Wu, "Transfer Learning for Brain-Computer Interfaces : An Euclidean Space Data Alignment Approach," *arXiv:1808.05464*, pp. 1–10, 2018.

[16] C. Brunner, R. Leeb, G. R. Muller-Putz, A. Schlogl, and G. Pfurtscheller, "BCI Competition 2008 - Graz data set A," 2008.

[17] F. Lotte and C. Guan, "Learning from other Subjects Helps Reducing Brain-Computer Interface Calibration Time," *International Conference on Audio Speech and Signal Processing (ICASSP), Mar 2010, Dallas, United States*, 2010.