2014 IEEE International Conference on Systems, Man, and Cybernetics
October 5-8, 2014, San Diego, CA, USA

Transfer Learning and Active Transfer Learning for Reducing Calibration Data in Single-Trial Classification of Visually-Evoked Potentials

Dongrui Wu[1], *Senior Member, IEEE*, Brent Lance[2], *Member, IEEE*, Vernon Lawhern[2,3,4]
[1] Machine Learning Laboratory, GE Global Research, Niskayuna, NY USA
[2] Translational Neuroscience Branch, U.S. Army Research Laboratory, USA
[3] Department of Computer Science, University of Texas at San Antonio
[4] Knowledge Services Branch, DCS Corporation, Alexandria, VA USA
E-mail: wud@ge.com, brent.j.lance @us.army.mil, vernon.lawhern@utsa.edu

***Abstract*—Single-trial Event-Related Potential (ERP) classification is a key requirement for several types of Brain-Computer Interaction (BCI) technologies. However, strong individual differences make it challenging to develop a generic single-trial ERP classifier that performs well for all subjects. Usually some subject-specific training samples need to be collected in an initial calibration session to customize the classifier. However, if implemented into an actual BCI system, then this calibration process would decrease the utility of the system, potentially decreasing its usability. In this paper we propose a Transfer Learning approach for reducing the amount of subject-specific data in online single-trial ERP classifier calibration, and an Active Transfer Learning approach for offline calibration. By applying these approaches to data from a Visually-Evoked Potential EEG experiment, we demonstrate that they improve the classification performance, given the same number of labeled subject-specific training samples. In other words, these approaches can also attain a desired level of classification accuracy with less labeling effort when compared to a randomly selected training set.**

***Keywords*—Single-trial classification, ERP, VEP, EEG, transfer learning, active learning, active transfer learning**

I. Introduction

Brain-computer interfaces (BCIs) have attracted explosive research interest in the last decade [9], [13], [26], thanks to recent advances in neurosciences, wearable/mobile biosensors, and analytics. Some have left their laboratory settings and begun to seek real-life applications [15], [24], [26], such as gaming. Many of these real-world BCI applications require single-trial classification of Event-Related Potentials (ERPs). However, people demonstrate strong individual differences in neural response to tasks or stimuli, which make it challenging, if not impossible, to develop a generic single-trial ERP classifier whose parameters fit all subjects. Usually, some subject-specific training samples need to be collected in an initial calibration session to customize the classifier. A typical calibration session can take anywhere from five to 20 minutes [24]. When implemented into a BCI system this calibration session would decrease the utility of these systems, potentially slowing their rate of acceptance. So, there is a critical need to reduce the number of subject-specific training samples required for calibration.

Take an EEG-based BCI systems for labeling large numbers of images using single-trial ERP classification as an example [4], [20]. The EEG correlates of target detection can be stable across sessions within a single subject, but can vary widely across different subjects. As a result, building a reliable single-trial ERP classification model requires a significant amount of calibration data for each individual. However, although EEG responses from different subjects are different, they still share some similarity in the underlying ERP. So, the amount of subject-specific data in online calibration could be reduced by making use of information contained in other subjects' data. This is the idea of transfer learning (TL) [17], which, along with adaptive approaches [25], started to find applications in the BCI domain [1], [10], [11], [21] and will be further explored in this paper.

Additionally, in some application domains we have large amounts of offline unlabeled data and the calibration session is focused on labeling this data. For example, a user interested in quantified self-tracking [16] may use a wearable EEG-based BCI system to record his/her EEG everyday, and also an integrated camera to automatically take pictures of interesting things he/she sees, identified from ERP responses. Undoubtedly the BCI system can make mistakes and take uninteresting pictures, which the user may want to delete at the end of the day. So, from the large amount of pictures taken each day, the user needs to label some as interesting and some as uninteresting so that the ERP classification algorithm can filter existing pictures and improve its accuracy the next day. Instead of randomly selecting the pictures for labelling, the most informative pictures can be selected so that the system can provide improved performance given the same number of labeled pictures. In other words, a desired level of performance can be obtained with less labeling effort. This is the idea of active learning (AL) [22], which will also be explored in this paper.

Interestingly, TL and AL are complementary to each other, so they can be integrated to further reduce the number of subject-specific training samples in offline BCI calibration. The idea of integrating TL and AL, or Active Transfer Learning (ATL), was proposed recently [23] and there has not been much literature on it [6], [7], [18], [29]. Furthermore, all previous works are outside of the EEG analysis domain. A closely related work is [27], in which we focused on integrating TL and Active Class Selection [14], a specific variant of AL, in order to speed up an online calibration for detecting cognitive

978-1-4799-3840-7/14/$31.00 ©2014 IEEE

state from EEG and other physiological signals.

This paper proposes a novel implementation of TL for online calibration[1] of a single-trial ERP classifier, and ATL for offline calibration. We show that, at equal training set sizes, TL and ATL can outperform a training set selected purely at random. We also show, for a small subset of individuals, that ATL can either match or improve performance over 5-fold cross-validation while using only a fraction of the overall number of trials. This initial demonstration indicates that TL and ATL may be effective tools for building robust neural signal classifiers in online and offline calibrations.

The rest of the paper is organized as follows: Section II introduces the details of several algorithms: two baselines, TL, and ATL. Section III describes experimental results, performance comparison of the algorithms, and possible improvements to TL and ATL. Section IV draws conclusions.

II. ALGORITHMS

This section introduces the details of our algorithms, and two baseline approaches for comparison purpose. In this paper, we focus on 2-class classification problems, and use a Support Vector Machine (SVM) classifier with a Radial Basis Function (RBF) kernel. We consider the problem of classifying EEG data, but the algorithms should be generalizable to calibration problems.

A. Baseline 1 (BL1)

BL1 assumes we know all labels of the subject-specific samples, and uses 5-fold cross-validation and SVM to find the classification accuracy. This usually represents an upper bound of the classification performance we can get from other algorithms, although not always the case.

B. Baseline 2 (BL2)

BL2 is a simple iterative procedure for online calibration: in each iteration we randomly select a few unlabeled subject-specific training samples, ask the subject to label them, add them to the labeled training dataset, and then train an optimal SVM by cross-validation. We iterate until the maximum number of iterations is reached, or the cross-validation performance is satisfactory.

C. Transfer Learning (TL)

Assume we have already collected lots of labeled EEG epochs from other subjects, and now we are customizing a single-trial ERP classifier online for a new subject. Although EEG epochs from other subjects may not be completely consistent with those from the new subject, they usually still contain useful information, due to the similarity of the underlying ERP. As a result, the amount of online calibration data may be reduced if these auxiliary EEG epochs are used properly. TL [17], [28] is a framework for addressing this type of problems.

Definition 1: **(Transfer Learning)** [17] Given a source domain $\mathcal{D}_S$ with learning task $\mathcal{T}_S$, and a target domain $\mathcal{D}_T$ with learning task $\mathcal{T}_T$, TL aims to help improve the learning of the target predictive function $f_T(\cdot)$ in $\mathcal{D}_T$ using the knowledge in $\mathcal{D}_S$ and $\mathcal{T}_S$, where $\mathcal{D}_S \neq \mathcal{D}_T$, and/or $\mathcal{T}_S \neq \mathcal{T}_T$.

In the above definition, a domain is a pair $\mathcal{D} = \{\mathcal{X}, P(X)\}$, where $\mathcal{X}$ is a feature space and $P(X)$ is a marginal probability distribution, in which $X = \{x_1, ..., x_n\} \in \mathcal{X}$. $\mathcal{D}_S \neq \mathcal{D}_T$ means that $\mathcal{X}_s \neq \mathcal{X}_T$, and/or $P(X_S) \neq P(X_T)$, i.e., the features in the source domain and the target domain are different, and/or their marginal probability distributions are different. Similarly, a task is a pair $\mathcal{T} = \{\mathcal{Y}, P(Y|X)\}$, where $\mathcal{Y}$ is a label space and $P(Y|X)$ is a conditional probability distribution. $\mathcal{T}_S \neq \mathcal{T}_T$ means that $\mathcal{Y}_S \neq \mathcal{Y}_T$, and/or $P(Y_S|X_S) \neq P(Y_T|X_T)$, i.e., the label spaces between the source and target domains are different, and/or the conditional probability distributions between the source and target domains are different.

For example, in the domain of classifying target and non-target stimuli from time-locked EEG responses, labeled EEG epochs from a new subject would be the *primary data* in the target domain, while labeled EEG epochs from other subjects would be the *auxiliary data* from the source domain. A single data sample would consist of the feature vector for a single EEG epoch from one subject, collected as a response to a specific stimulus. Though the features in this primary data and auxiliary data are computed in the same way, generally their marginal distributions are different, i.e., $P(X_S) \neq P(X_T)$, due to the fact that the baseline EEG levels for different subjects are likely to differ. Moreover, the conditional probabilities are also different, i.e., $P(Y_S|X_S) \neq P(Y_T|X_T)$, due to the significant individual differences in EEG response to stimuli. As a result, the auxiliary data from the source domain cannot represent the primary data in the target domain accurately, and must be integrated with some labeled primary data in the target domain to induce the target predictive function.

There are many different TL algorithms [17]. The basic idea used in this paper is illustrated in Fig. 1. For each new subject, we combine his/her labeled samples with labeled samples from each auxiliary subject in building a classifier, where the contribution of labeled samples from an auxiliary subject is determined by the response similarity of the two subjects. The detailed implementation is given in Algorithm 1, which can be viewed as an instance transfer approach [17], as specific instances of data are being transferred from one subject to another in order to improve model performance. First, a classifier C_0 is trained based on labeled primary training samples from the new subject only. Then, a classifier C_i for the ith subject in the auxiliary data is trained by combining his/her data with labeled primary training samples from the new subject. The final classification is a weighted voting from all these classifiers, where C_0 has unit weight, and C_i's weight is its cross-validation accuracy.

D. Active Learning (AL)

As mentioned in Introduction, in some applications there are large amounts of offline unlabeled training samples, and calibration consists of selecting some of them for labeling. The selection strategy may have a significant impact on the calibration performance. AL [22] tries to select the most informative samples to label so that a given learning performance can be achieved with less labeling effort.

[1] Actually TL can be used in both online and offline calibrations; however, as will be demonstrated later in this paper, ATL has better performance in offline calibration. So, we will only use TL in online calibration.

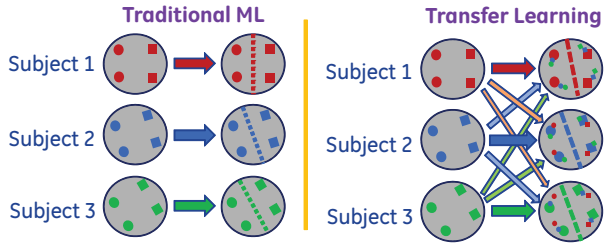

Fig. 1. An illustration of TL. The circles and squares are labeled training samples from two classes. The size of a circle or square indicates its weight. The dotted and dashed lines are classification boundaries.

Algorithm 1: The TL algorithm.

Input: N^l labeled primary training samples;
N^u unlabeled primary training samples;
M, the number of subjects in auxiliary data;
N_m $(m = 1, 2, ..., M)$, the number of labeled training samples for the mth subject;
Output: Labels for the N^u unlabeled samples.
Cross-validation using the N^l samples to find the best SVM parameters, $\mathbf{P}$;
Train a SVM on the N^l samples using $\mathbf{P}$;
Classify the N^u unlabeled samples and denote their labels as $\{L_i^0\}_{i=1,2,...,N^u}$;
for m in $[1,\ M]$ **do**
  Combine the N^l primary samples and N_m auxiliary samples;
  Cross-validation using these $N^l + N_m$ samples to find the best SVM parameters, $\mathbf{P_m}$;
  Record the best cross-validation accuracy, a_m;
  Train a SVM on the $N^l + N_m$ samples using $\mathbf{P_m}$;
  Classify the N^u unlabeled primary samples and denote their labels as $\{L_i^m\}_{i=1,2,...,N^u}$;
end
Compute weighted sum of the $1 + M$ SVM outputs: $L_i = L_i^0 + \sum_{m=1}^{M} a_m \cdot L_i^m$, $i = 1, 2, ..., N^u$;
Return $sign(L_i)$, $i = 1, 2, ..., N^u$;

A popular AL idea is illustrated in Fig. 2. Suppose we have two classes, A and B, and they are separated by the green dashed circle in the left part of the figure. In practice we do not know the true distributions of these two classes, hence the resulting linear classification boundary. We only have a few labeled samples from each class, and many more unlabeled samples from both classes, as shown in the middle part of the figure. Our task is to select a few more unlabeled samples (say two) to label such that a better classifier can be trained. The easiest way is to do random selection. Then we may end up with two new labeled samples shown as the stars in the top right part of the figure, which actually do not provide new information as they have no impact on the classification boundary. In AL, the goal is to select the two most informative samples for labeling, shown as the stars in the bottom right part of the figure. The classification boundary will be changed significantly once these two new samples are added to the training dataset, and it is a better approximate of the true classification boundary. As we iterate through this process, the AL classification boundary should rapidly approach the true one.

The key problem in using AL is estimating which of the data samples are the most informative. There are many different heuristics for this purpose. In our implementation a committee is created by training multiple classifiers on different subsets of the data. The data samples selected as the most informative are those with the greatest amount of uncertainty, defined as those points with the most disagreement between the classifiers in the committee. More specifically, assume m_1 classifiers classify a sample as positive and m_2 as negative, then the smaller $|m_1 - m_2|$ is, the most disagreement there is.

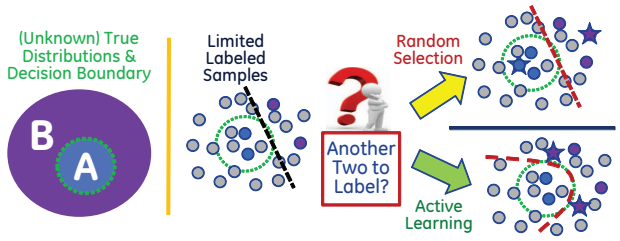

Fig. 2. An illustration of AL. The blue or purple circles are initial labeled training samples, where different colors denote different classes. The stars are the newly labeled samples. The black line is the decision boundary based on initial labeled samples only, and the red line and curve are the decision boundaries after the newly labeled samples have been added.

E. Active Transfer Learning (ATL)

Because AL considers how to optimally label offline subject-specific data and TL (which can be used both online and offline) considers how to make use of training data from other subjects, they are complementary. So, we conjecture that integrating AL with TL will further improve the performance of TL in offline calibration. The fundamental concept is to use TL to select the optimal classifier parameters for the new subject based on available data obtained from the new subjects and other subjects, and then use AL to select the most informative unlabeled samples for labeling for the new subject, until the desired cross-validation accuracy is obtained, or the maximum number of iterations is reached, as illustrated in Fig. 3.

In our implementation of ATL, the TL part is the same as Algorithm 1, and the AL part uses the idea introduced in the previous subsection. Recall that the label for the ith unlabeled training sample is determined as $sign(L_i)$ in TL (Algorithm 1). Because a smaller $|L_i|$ means a larger disagreement among the $M+1$ classifiers, in AL we simply pick those samples corresponding to the smallest $|L_i|$ to label.

III. Experiments and Discussions

Experimental results are presented in this section to compare the algorithms proposed in the previous section. Potential improvements to the algorithms are also discussed.

A. Experiment Setup

We used data from a standard Visually Evoked Potential (VEP) oddball task [19]. In this task, image stimuli were presented to subjects at a rate of 0.5 Hz (one image every

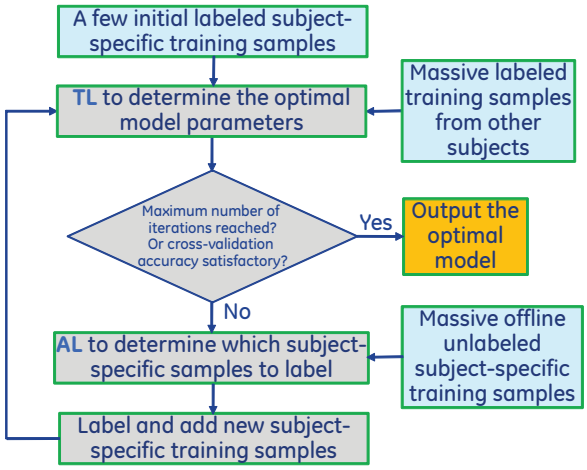

Fig. 3. Flowchart of ATL.

two seconds). The images presented were either an enemy combatant [target; an example is shown in Fig. 4(a)] or a U.S. Soldier [non-target; an example is shown in Fig. 4(b)]. The subjects were instructed to identify each image as being target or non-target with a unique button press as quickly, but as accurately, as possible. There were a total of about 270 images presented to each subject, of which the number of targets ranged from 30 to 55. The experiments were approved by U.S. Army Research Laboratory (ARL) Institutional Review Board.

(a) (b)

Fig. 4. (a) A target image; (b) A non-target image.

16 subjects participated the experiments, which lasted on average 15 minutes. Data from two subjects were not used due to data corruption, so we only used data from 14 subjects in this analysis. EEG signals were recorded using a 64-channel BioSemi ActiveTwo system with 4 additional EOG channels to record eye movement activity. The EEG data was sampled at 512Hz.

B. Preprocessing and Feature Extraction

We used EEGLAB [8] for EEG signal preprocessing and feature extraction. We compared several different types of features, e.g., raw magnitudes, power spectral features, and time-frequency features. Raw magnitudes achieved robust performance, and also were the easiest to extract. Since the goal of this paper is to demonstrate how advanced machine learning algorithms can improve single-trial ERP classification performance based on existing data and features, we used raw magnitude features in our algorithms.

Of the 64 BioSemi EEG channels, we only used 21 channels (Cz, Fz, P1, P3, P5, P7, P9, PO7, PO3, O1, Oz, POz, Pz, P2, P4, P6, P8, P10, PO8, PO4, O2) mainly in the parietal and occipital areas, as research has shown that they demonstrate strong visual ERPs [12]. We first downsampled the 512 Hz EEG signals to 64 Hz, and then epoched the EEG signals to the $[0, 0.7]$ second interval timelocked to stimulus onset. We removed mean baseline from each channel in each epoch and removed epochs with incorrect button press responses. Because the sizes of the two classes were highly imbalanced, we downsampled the non-target class to match the target class by selecting the non-target epoch that occurred immediately before each target epoch. In the rare case that there was no non-target epoch before a target epoch; i.e., a target image was presented first in the sequence, we selected the non-target epoch immediately following that target epoch. After preprocessing, on average each subject had 54 epochs, half target and half non-target.

Each [0, 0.7] second epoch contains 44 raw EEG magnitude samples (64×0.7). The feature vector obtained by concatenating the features from all 21 EEG channels would be excessively large. To reduce the dimensionality, we performed a simple principal component analysis for each channel and took only the scores for the first five principal components. As a result, each epoch had $5 \times 21 = 105$ features. We then normalized each feature dimension separately to $[0, 1]$ for each subject.

C. Experimental Results

Although we know the labels of all EEG epochs for all 14 subjects in the experiment, we simulate a different scenario: we have labeled EEG epochs for 13 subjects, but only a small number of epochs for the 14th subject are labeled. Our goal is to iteratively label epochs for the 14th subject so that the remaining unlabeled epochs can be reliably classified. We repeat this procedure 14 times so that each subject has a chance to be the "14th" subject.

We compared the performance of BL1, BL2, TL and ATL introduced in the previous section. BL2, TL and ATL started with the same four randomly selected labeled samples from the "14th" subject, two in target class and two in non-target class. In each iteration, two new EEG epochs were labeled and added to the training dataset. For BL2 and TL, these two were selected randomly from unlabeled samples, simulating online calibration. For ATL, these two were selected by AL, simulating offline calibration. Testing classification accuracy was computed from the remaining unlabeled subject-specific samples. We used libSVM [5] with RBF kernel as the base classifier in all algorithms. Due to the large variability in classification performance, each algorithm was repeated 100 times so that statistically meaningful results could be obtained.

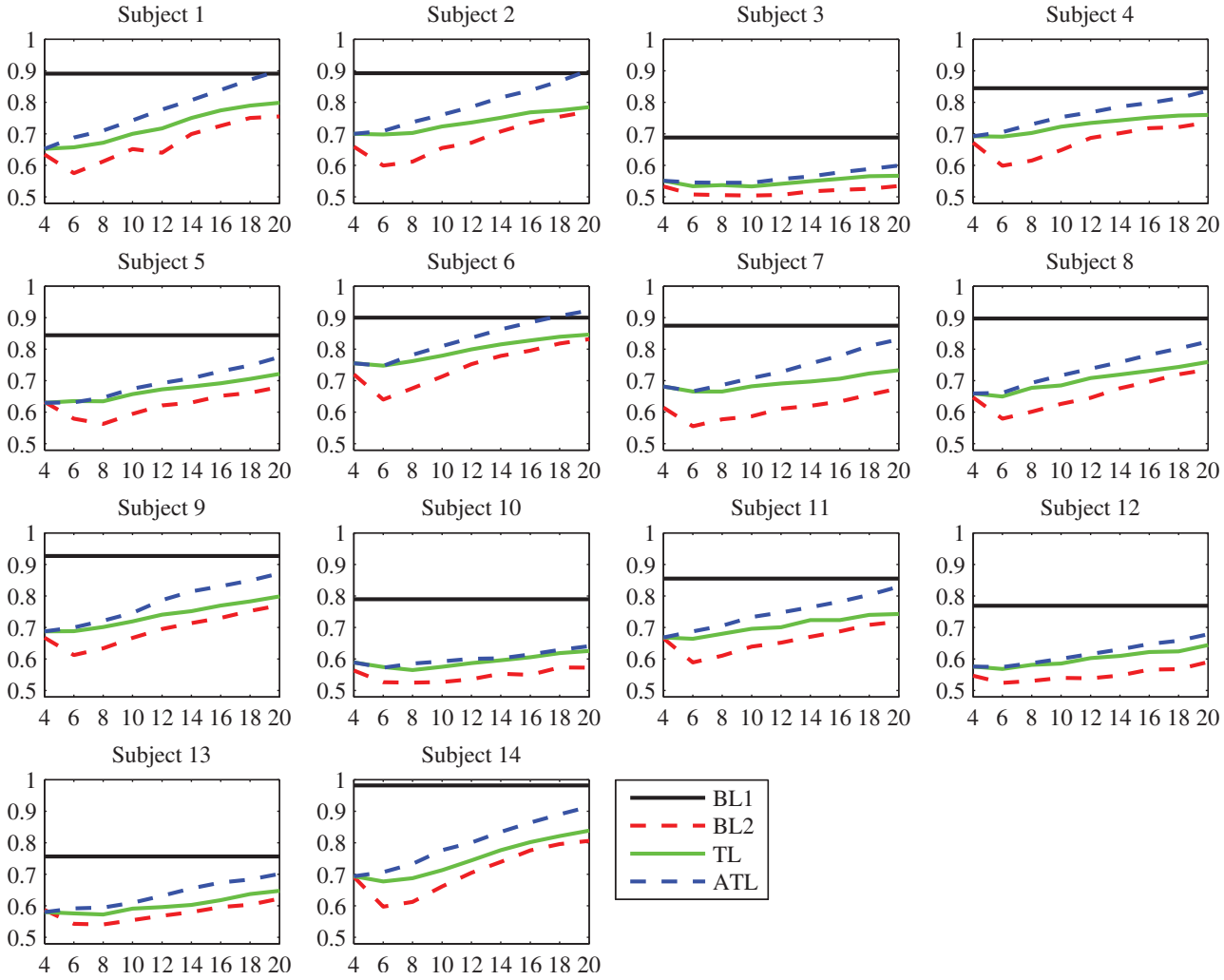

Fig. 5. Performance of the four algorithms for each individual subject, averaged over 100 runs. Horizontal axis: N^l, number of labeled subject-specific training samples; vertical axis: testing classification accuracy.

The performances of the four algorithms, which are averaged across the 100 runs for each subject, are shown in Fig. 5, where each subfigure represents a different "14th" subject. The average performance of the four algorithms across the 14 subjects is shown in Fig. 6. Observe that:

1) Generally the performance of BL2 increases as more subject-specific training samples are labeled and added; however, it drops when the first two new labeled samples are added, i.e., when the number of labeled samples increases from four to six. This is because the random sampling approach used in BL2 and TL may result in significant class imbalance, when the number of labeled samples is small. We have ensured that of the four initial labeled samples, two from the target class and two from the non-target class; however, in the next iteration the two new labeled samples may be from the same class. For example, if the two new samples are both from the target class, then of the six samples after the first iteration, four are from target class and two from non-target class, so the two classes are highly unbalanced. BL2 may simply classify all samples as target, resulting in a training classification accuracy of 67% but testing accuracy of about 50%. We will improve our method in future research to overcome this problem, e.g. by ensuring more balanced sampling, or by using F-score instead of classification accuracy to determine the best SVM parameters.
2) TL almost always outperforms BL2, which coincides with our conjecture. Furthermore, the performance drop of TL when the number of labeled samples increases from four to six is much smaller than that of BL2, which means that by considering auxiliary data from other subjects, the primary data class imbalance problem can be significantly alleviated.
3) ATL almost always outperforms TL in Fig. 5, and the average performance improvement is quite evident, as shown in Fig. 6. This verifies our conjecture that TL and AL are complementary, and hence integrating AL with TL can further improve the offline calibration performance. Furthermore, with the help of AL, the

performance drop of ATL when the number of labeled primary samples increases from four to six is also smaller than that of TL. Surprisingly, Fig. 5 also shows that sometimes ATL can even outperform BL1, which suggests that by utilizing data from other subjects, we may be able to achieve classification accuracy that is unreachable by using only subject-specific data, even though there may be a lot of such data.

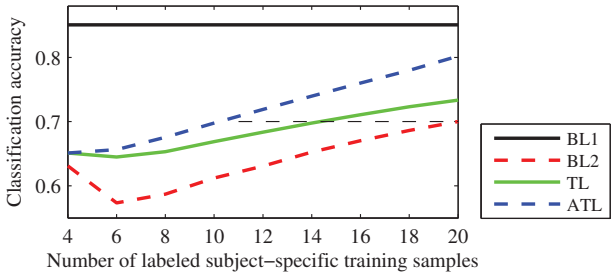

Fig. 6. Average performance of the four algorithms across the 14 subjects.

To show that the performance differences among BL2, TL and ATL are statistically significant, we performed a 3-way mixed-effects analysis of variance (ANOVA), considering random effects on the subjects and the number of labeled subject-specific samples. The p values for the three factors (subjects, number of labeled subject-specific samples, and algorithms) are 0.0001, 0.0249, and 0.0458, respectively, as shown in Table I. A post-hoc multiple comparison procedure using the Tukey-Kramer test showed that the performance improvement of TL over BL2 is statistically significant, and the performance improvements of ATL over BL2 and TL are also statistically significant.

TABLE I. RESULTS OF 3-WAY ANOVA.

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
Subjects	193.816	13	14.9089	7.68	0.0001
SampleSize	61.129	8	7.6411	3.08	0.0249
Algorithms	42.615	2	21.3074	3.60	0.0458

In summary, we have demonstrated that given the same number of labeled subject-specific training samples, TL can improve online calibration performance, and ATL can improve offline calibration performance. For Subjects 1, 2, 4, 6 and 11, we see ATL either matched or exceeded the classification performance of BL1 with only 20 labeled trials. For these subjects this represents a significant decrease in the amount of labeled data needed when compared to the sample size used to calculate BL1 ($\sim$40). While the ATL performance in all other subjects did not outperform BL1, we do see an increasing trend as the number of labeled samples increases. In other words, given a desired classification accuracy, TL and ATL can reduce the number of labeled subject-specific training samples. For example, in Fig. 6, the average classification accuracy of BL2 is 70%, given 20 labeled subject-specific training samples. However, to achieve or exceed that performance, TL only needs 16 samples, and ATL only needs 12, which are corresponding to 20% and 40% savings in labeling effort, respectively.

D. Future Improvements

The performance of TL and ATL can be further improved in several different ways. As mentioned in the previous subsection, we may be able to remedy the performance drop in the first few iterations by ensuring roughly balanced sampling, or using F-score instead of accuracy in determining the best classifier parameters. Additionally, the following improvements will be investigated in our future research:

1) Using only carefully selected auxiliary subjects may be better than using all subjects in both online and offline calibrations. In our approach all 13 auxiliary subjects were used; however, some subjects' responses may be so different from the new subject that actually it is more beneficial to not include them. This will also reduce the computational cost of TL and ATL. A systematic procedure needs to be designed to automatically select the most useful auxiliary subjects, based on the labeled and unlabeled samples from the new subject.
2) More sophisticated features may result in more robust transfer in both online and offline calibrations. As TL benefits from the similarity among subjects, and that similarity is expressed by EEG features, we conjecture that more robust features would improve the performance of TL and hence ATL. In addition, a more robust feature space may be more montage-independent, allowing us to make use of the data from the other EEG headsets. We will investigate deep learning, or representation learning [3], for this purpose.
3) More sophisticated TL and AL algorithms can be used in ATL for offline calibration. In this paper we only used a very basic instance transfer approach in TL. There are many other approaches [17] that can be investigated, e.g., feature representation transfer, parameter transfer, and relational knowledge transfer. For AL there are also many other approaches [22] that can be investigated, e.g., expected model change, expected error reduction, and variance reduction.
4) We could also make better use of information in the unlabeled subject-specific samples in offline calibration. In the current implementation, we are only using the unlabeled subject-specific samples for AL, but they could also be used in TL part, in a semi-supervised learning setting, e.g., manifold regularization [2].

Finally, it will be important to implement TL and ATL into an actual BCI paradigm and evaluate them in both online and offline conditions.

IV. CONCLUSIONS

In this paper we have proposed a Transfer Learning approach for online BCI calibration, which uses data from other subjects to reduce the amount of calibration data required to perform accurate online single-trial classification of ERPs, and an Active Transfer Learning approach for offline BCI calibration, which integrates data from other subjects while simultaneously selecting the most informative data from the current subject in order to minimize the offline calibration

effort. TL and ATL can indeed improve the classification performance, given the same number of labeled subject-specific training samples; or, equivalently, they can reduce the number of labeled subject-specific training samples, given a desired classification accuracy. This suggests that TL and ATL may be useful techniques for online and offline training of robust neural classifiers.

TL and ATL have many potential applications in EEG classification, where large between-individual variation can cause difficulty in developing classifier models. For example, they may have relevance to BCI technologies that rely on single-trial ERP classification, such as image analysis BCI systems [20]. In the future we intend to improve both TL and ATL, and to demonstrate their generality by applying them to several distinct EEG classification domains.

ACKNOWLEDGEMENT

The authors would like to thank Scott Kerick, Jean Vettel, Anthony Ries, and David Hairston at the US Army Research Laboratory (ARL) for designing the experiment and collecting the data.

Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-10-2-0022. The views and the conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S Government. The U.S Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

REFERENCES

[1] M. Ahn, H. Cho, and S. C. Jun, "Calibration time reduction through source imaging in brain computer interface (BCI)," *Communications in Computer and Information Science*, vol. 174, pp. 269–273, 2011.

[2] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *Journal of Machine Learning Research*, vol. 7, pp. 2399–2434, 2006.

[3] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[4] N. Bigdely-Shamlo, A. Vankov, R. Ramirez, and S. Makeig, "Brain activity-based image classification from rapid serial visual presentation," *IEEE Trans. on Neural Systems and Rehabilitation Engineering*, vol. 16, no. 5, pp. 432–441, 2008.

[5] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," 2009, accessed January 15, 2013. [Online]. Available: http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

[6] R. Chattopadhyay, W. Fan, I. Davidson, S. Panchanathan, and J. Ye, "Joint transfer and batch-mode active learning," in *Proc. 30th Int'l. Conf. on Machine Learning (ICML)*, Atlanta, GA, June 2013.

[7] M. Chen, K. Weinberger, and J. Blitzer, "Co-training for domain adaptation," in *Proc. 25th conf. on Neural Information Processing Systems (NIPS)*, Granada, Spain, December 2011.

[8] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *Journal of Neuroscience Methods*, vol. 134, pp. 9–21, 2004.

[9] B. Hamadicharef, "Brain-computer interface (BCI) literature - a bibliometric study," in *Proc. 10th Int'l. Conf. on Information Sciences Signal Processing and their Applications*, Kuala Lumpur, May 2010, pp. 626–629.

[10] P.-J. Kindermans and B. Schrauwen, "Dynamic stopping in a calibration-less P300 speller," in *Proc. 5th Int'l. Brain-Computer Interface Meeting*, Pacific Grove, CA, June 2013.

[11] P.-J. Kindermans, H. Verschore, D. Verstraeten, and B. Schrauwen, "A P300 bci for the masses: Prior information enables instant unsupervised spelling," in *Proc. Neural Information Processing Systems (NIPS)*, Lake Tahoe, NV, December 2012.

[12] D. J. Krusienski, E. W. Sellers, D. J. McFarland, T. M. Vaughan, and J. R. Wolpaw, "oward enhanced P300 speller performance," *J. Neurosci Methods*, vol. 167, no. 15, p. 21, 2008.

[13] B. J. Lance, S. E. Kerick, A. J. Ries, K. S. Oie, and K. McDowell, "Brain-computer interface technologies in the coming decades," *Proc. of the IEEE*, vol. 100, no. 3, pp. 1585–1599, 2012.

[14] R. Lomasky, C. E. Brodley, M. Aernecke, D. Walt, and M. Friedl, "Active class selection," in *Proc. 18th European Conference on Machine Learning*, Warsaw, Poland, September 2007, pp. 640–647.

[15] K. McDowell, C.-T. Lin, K. Oie, T.-P. Jung, S. Gordon, K. Whitaker, S.-Y. Li, S.-W. Lu, and W. Hairston, "Real-world neuroimaging technologies," *IEEE Access*, vol. 1, pp. 131–149, 2013.

[16] P. McFedries, "Tracking the quantified self," *IEEE Spectrum*, vol. 50, no. 8, p. 24, 2013.

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

[18] P. Rai, A. Saha, H. Daumé, III, and S. Venkatasubramanian, "Domain adaptation meets active learning," in *Proc. NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing*, Los Angeles, CA, June 2010, pp. 27–32.

[19] A. Ries, J. Touryan, J. Vettel, K. McDowell, and W. Hairston, "A comparison of electroencephalography signals acquired from conventional and mobile systems," *Journal of Neuroscience and Neuroengineering*, vol. 3, pp. 10–20, 2014.

[20] P. Sajda, E. Pohlmeyer, J. Wang, L. Parra, C. Christoforou, J. Dmochowski, B. Hanna, C. Bahlmann, M. Singh, and S.-F. Chang, "In a blink of an eye and a switch of a transistor: Cortically coupled computer vision," *Proc. of the IEEE*, vol. 98, no. 3, pp. 462–478, 2010.

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

[22] B. Settles, "Active learning literature survey," University of Wisconsin–Madison, Computer Sciences Technical Report 1648, 2009.

[23] X. Shi, W. Fan, and J. Ren, "Actively transfer domain knowledge," in *Proc. European Conf. on Machine Learning (ECML)*, Antwerp, Belgium, September 2008, pp. 342–357.

[24] J. van Erp, F. Lotte, and M. Tangermann, "Brain-computer interfaces: Beyond medical applications," *Computer*, vol. 45, no. 4, pp. 26–34, 2012.

[25] C. Vidaurre, A. Schlogl, R. Cabeza, R. Scherer, and G. Pfurtscheller, "A fully on-line adaptive BCI," *IEEE Trans. on Biomedical Engineering*, vol. 53, no. 6, pp. 1214–1219, 2006.

[26] J. Wolpaw and E. W. Wolpaw, Eds., *Brain-Computer Interfaces: Principles and Practice*. Oxford, UK: Oxford University Press, 2012.

[27] D. Wu, B. J. Lance, and T. D. Parsons, "Collaborative filtering for brain-computer interaction using transfer learning and active class selection," *PLoS ONE*, 2013.

[28] P. Wu and T. G. Dietterich, "Improving SVM accuracy by training on auxiliary data sources," in *Proc. Int'l Conf. on Machine Learning*, Banff, Alberta, Canada, July 2004, pp. 871–878.

[29] L. Zhao, S. Pan, E. Xiang, E. Zhong, Z. Lu, and Q. Yang, "Active transfer learning for cross-system recommendation," in *Proc. AAAI Conf. on Artificial Intelligence*, Bellevue, WA, July 2013.