2017 2017 IEEE 17th International Conference on Bioinformatics and Bioengineering

Multi-Task Learning for Commercial Brain Computer Interfaces

George Panagopoulos
Computational Physiology Lab
University of Houston
Houston, TX 77004
gpanagopoulos@uh.edu

***Abstract*—In the field of Brain Computer Interfaces (BCIs), one of the most crucial hindrance towards everyday applicability is the problem of subject-to-subject generalization. This adheres to the fact that neural signals vary significantly across subjects, rendering a subject calibration process necessary for the pattern recognition mechanisms of a BCI to achieve a notable performance. In the present work, we explore this phenomenon on two open datasets from mental monitoring experiments which utilized a commercial BCI device (Neurosky). This passive BCI setting with economical hardware is one of the must promising in terms of commercial appeal and hence it has more potential to be employed by multiple subjects-users. We visualize the inter subject variability problem and apply machine learning methods commonly used in the BCI literature. Subsequently we employ multi-task learning algorithms, setting each subject specific classification as a separate task. The experiments reveal that multi-task approaches achieve better accuracy with increasing number of subjects in contrast to subject-invariant approaches, while providing insights that are consistent among subjects and agree with the relevant literature.**

***Keywords*-Brain Computer Interfaces, Multi-task learning, EEG**

I. Introduction

A Brain Computer Interface, or BCI, is a device used to form a communication pathway between a human's or animal's brain and a computer. The device combines medical imaging techniques, ranging from EEG to fMRI on the one end, with machine learning and signal processing techniques on the other, in order to extract patterns of a subject's brain activity and map them to specific actions or meanings. One of the most persistent problems in the field lies in the machine learning phase, where the substantial variance between recordings of different subjects calls for subject specific calibration, to enhance the accuracy of the classifier. This enduring obstacle emanates from the fact that neural signals are highly subject specific, which in term stems from the mapping of cognitive functions to brain regions being highly volatile between human beings. Because of this, patterns inherent throughout all subjects are not exploited as common ground. In contrast the parameters of the model are tuned from scratch before the subject starts using it, which has several drawbacks. In terms of usability, it is time consuming and inconvenient, becoming a burden for the commercial appeal of BCIs, especially for devices with moist electrodes that already need rigorous preparation during the electrode placement. From a scientific perspective, a BCI is a potential source of meaningful insights on how the brain works, where if generalization can not take place, universal conclusions can not be derived. Although the need for intensive and lengthy training has been alleviated [1] and the BCI placement becomes more and more effortless with commercial hardware [2], the problem of subject-to-subject generalization can still be considered detrimental for both, the commercial and the scientific usage of BCI.

We approach the problem from a multi-task learning perspective, showing that sharing knowledge between subjects can prove beneficial towards subject-to-subject generalization and reveal subject invariant patterns that agree with the literature. We focus on passive BCI [3] where the subjects observe a continuous stimulus, while the device keeps track of their neural signals, using a commercial and low-cost hardware. This combination has potentially sizable commercial impact and a broad range of applications[4]. The datasets employed are open and come from experiments run in Berkeley[1] and Carnegie Mellon[2] using Neurosky device. This device is one of the most portable BCIs in the market with a cost around $100. On the other hand, it is one of the weakest BCIs since it consists of only one dry sensor, placed in the forehead, rendering capture patterns from other brain areas nearly impossible and being very prone to noise. This, combined with the difficulty of the mental monitoring task, renders the problem at hand so challenging that baseline approaches hardly surpass randomness [5].

Overall, the goal of the paper is twofold; the first is to evaluate machine learning algorithms utilized in other approaches with expensive hardware, such as active (ex. motor imagery) or reactive BCI (ex. P300 spellers) to a passive BCI with economical hardware; the second is to highlight the benefits of exploiting inter subject knowledge, using multi-task learning. The paper is organized as follows. Section 2 provides a review on transfer learning applications in Neuroimaging. Section 3 describes the multi-task learning algorithms we employed and implementation details. In

[1] https://www.kaggle.com/wanghaohan/eeg-brain-wave-for-confusion

[2] https://www.kaggle.com/berkeley-biosense/synchronized-brainwave-dataset

2471-7819/17/31.00 ©2017 IEEE
DOI 10.1109/BIBE.2017.00022

IEEE computer society

section 4 the datasets are delineated, displaying the "curse" of inter subject variability. Section 5 outlines the experimental design and discusses the results. Finally, the paper is concluded in Section 6, with future directions.

II. Related Work

The first systematic approach to multi-task learning was proposed through hierarchical bayesian inference [6] and revolves around the idea that the parameter estimation of a machine learning model is a task in itself, while the estimation of the right bias term, which is shared throughout all tasks and is crucial for generalization, can be approximated using bayesian inference. The bias term estimation represents the objective prior distribution for all relevant tasks. A similar idea is employed to enhance the inter-subject generalization capability of a BCI in [7], which is one of the algorithms we employ in the experiments. Each task corresponds to a subject and all tasks' parameters stem from the same prior Gaussian distribution. These parameters can then further adapt to the subject itself. The same group examined thoroughly transfer learning and its application in BCI [8]. A feature decomposition process was added to the aforementioned algorithm to take advantage the BCI's data structure. Results verify the subject-to-subject knowledge transfer is indeed beneficial. A joint prior connecting subjects combined with knowledge on language models was utilized in [9] to create a p300 based speller BCI that can work accurately without lengthy calibration.

Multi output artificial neural networks can also be considered a multi task learning algorithm. If each output unit is a different task, then the internal representations of the neural network e.g. the transformations of the data with the weights in the hidden layers, are shared knowledge between all tasks and hence contribute to perform better in each task [10]. Such settings have been used for prediction using clinical recordings like in [11] with pneumonia risk assessment and in [12], with patient's length of in hospitalization, mortality etc. A combination of neural networks and hierarchical bayesian inference led to multi-task bayesian neural networks with multiple levels of sharing between tasks, like sharing parameter's prior distributions or sharing the parameters themselves [13], [14]. Such a network is tested in [15] in the context of BCI to classify multiple cognitive tasks and motor imagery assignment, surpassing the single task approaches. In addition, Gaussian Processes (GP) were utilized for multi task learning [16] and were critically acclaimed, with extended theoretical examinations [17], [18] and a broad range of applications, in the analysis of fMRI data It is used in [19] to facilitate functional connectivity estimation while classifying resting state. The multi task approach yielded better generalization through sessions and uncovered consistent brain connectivity graphs that could be tallied to known cognitive networks. In [20] each task adheres to classifying whether the subject is aroused by a stimulus, given her/his fMRI recordings, using knowledge transfer between subjects and explaining away variation that is not shared. The methodology relies on having a primary task which borrows knowledge from secondary ones, while explaining away variation that is not shared. In order to achieve generalization and subjectwise consistent pattern extraction based on fMRI data in [21], the GP covariance function is decomposed to intra-task and inter-task component, restricting the latter to be close to the respective average for all tasks. The results were evaluated based on both, generalization in new subjects and reproducibility of weight vectors through tasks in multiple runs compared with single task classifiers.

One of the must popular approaches to multi-task learning is based on regularized linear models, the traditional form being a linear model with a regularization term consisting of the parameters of all tasks [22]. The multi-task L_{12} norm regularization was applied to simulated and real magneto-encephalographic (MEG) recordings [23], achieving improved generalization accuracy and uncovering meaningful hidden patterns. The same group used multi task learning for MEG [24] with elastic net regularization, each task corresponding to a spatial unit e.g. a sensor. An adaptive mixed norm regularization was used for P300 speller BCI [25] to enhance classification accuracy and perform sensor selection. In [26] a temporal multi-task model is used to forecast MMSE and ADAS-Cog based on fMRI data from ADNI dataset. Each time point regression served as a different task and shared regularizations ensured common feature selection, smoothness in consecutive time points and time specific feature selection. Finally, joint feature selection for fMRI data is performed via regularization for group-wise and subject-wise sparsity in [27], surpassing group Lasso.

Sharing knowledge between subjects is not limited to learning algorithms though. It can be achieved in preprocessing steps, such as sharing filters in Common Spatial Patterns (CSP) [28] to reduce calibration time for motor imagery. The effectiveness of this approach led to the examination of several types of novel CSP regularizations and a theoretical framework to rely on [29]. In a similar fashion, non-stationary directions of the data are transfered trough subjects for motor imagery BCI in [30]. Transfer knowledge for EEG based visual-spatial attention tasks is facilitated in [31], creating a common dictionary for all subjects and transferring resting state activity, which improved the decoding efficiency and produced meaningful brain activations. The approaches using CSP have a solid mathematical background and are very promising, however we can not apply CSP in current work since our datasets consist only of one sensor and its frequency components.

Finally, training weak classifiers on specific subjects and classify a new one based on their ensemble, is also a form of multi-task learning [32]. A combination of CSP filters and Linear Discriminant Analysis (LDA) parameters was computed for each subject during training and a sparse ensemble of them was utilized to achieve zero calibration time for motor imagery assignment on a new subject in [33]. In addition, a subject based ensemble is used for fMRI classification in [34] with subject based kernel classifiers and the total classification was formed by their weighted combination.

Our approach diversifies from the ones addressed in the aforementioned literature in several aspects. First of all, it is based on a commercial and economic BCI. Data from this type of hardware suffer numerous limitations and that is why, to the best of our knowledge, multi-task learning has not been applied yet on such data. Neurosky, has only one channel. This effectively means traditional EEG workhorses like ICA [35] and CSP [29] are out of scope. Since mining for spatial patterns becomes meaningless, we have to focus on temporal patterns in a device with substantially inferior sampling rate and signal to noise ratio of clinical BCIs. Another difference is that we apply this methodology in a passive BCI. Passive BCIs exhibit high potential of applicability in real life [4], but are overlooked due to their challenging nature. More specifically, subjects do not intentionally change their brain activity, which renders the signal's behavior obscure and hard to decode. Finally, a substantial contrast between our work and the literature is the purely mental nature of our classes, such as confusion and mental burden. Classification of mental notions has proven significantly harder, calling for longer periods of training and fully subject adaptive classifiers [36], while it is significantly more ambiguous and less explored than motor imagery [37] or P300 spelling [9]. Thus the need for novel algorithms and methodologies is prevalent for mental classes.

III. Methods

Subject-invariant algorithms handle the data in a pooled manner. This means that each input $X \in \mathbb{R}^F$ and output $Y \in \mathbb{R}$ is treated by the classifier in a subject agnostic manner, meaning without knowing which subject it belongs to. In contrast, when training a multi-task algorithm, the X and Y of each subject correspond to the input and output of a specific task, as are depicted in Figure 3 for two subjects. The evaluation phase is the same in both cases, testing in unseen subjects. In a subject adaptive algorithm, each task would be trained independently and evaluated only in the same subject, thus we could think of multi-task learning as a middle ground solution between pooled and fully adaptive. We utilize two multi-task learning algorithms, one discriminative and one generative.

A. Discriminative Model

The discriminative algorithm is a logistic regression with fused regularization [38]

$$\min_{w,i} \sum_{t=1}^{T}\sum_{i=1}^{N_t} log(1+e^{(-Y_{t,i}(X_{t,i}w_t+c_i))})+p_1|W|_{2,1}+p_{L_2}|W|_F^2 \tag{1}$$

Where T is the number of tasks, in our case subjects, N_t, $X_{t,i} \in R^{N_t x F}$ and $Y_{t,i} \in R_t^N$ are the number of samples, the input samples and the output labels for subject t. F is the number of features per sample, in our case 9. $W \in R^{TxF}$ is the matrix including the feature coefficients of all tasks and W_t is the row of W that corresponds to the coefficients for task t. Regularization coefficient p_1 is set to 1 after cross validation and p_{L2} is left to default.

The term $|W|_{2,1}$ induces the group sparsity regularization, which means the vectors of coefficients of all tasks are constrained to share the same sparsity. In other words, the feature selection for one task, must compromise with the feature selection for all tasks. In this manner, features that might achieve better accuracy for one subject but do not fit well with the rest, are overlooked. Instead, features that perform sufficiently well for all subjects are encouraged and this is how knowledge is shared between subjects. The second regularization term is a traditional L_2-norm penalty for sparsity withing each task. The optimization of the function is achieved through accelerating gradient descent [39]. In order to assess the generalization to new subjects, we take the average of all subjects' coefficient vectors to serve as the coefficients for an one. The MALSAR [40] implementation of the method was used to perform the experiments in MATLAB.

B. Generative Model

The generative model is based on a Bayesian approach where the prior distribution of the coefficients is shared between subjects [7]. More specifically, the coefficient vectors of all subjects $w_t \in R^F$ are assumed to stem from the same distribution $p(W) \sim N(\mu, \Sigma)$. Σ and μ are inferred by maximizing the posterior probability given the data of all tasks t

$$p(W; X, Y, \sigma^2) \sim \prod_{t \in T} p(y_t, X_t; w_t)p(w_t) \tag{2}$$

which is the same as minimizing the negative log-posterior

$$\min_{W,\mu,\Sigma} \frac{1}{\sigma^2}\sum_{t \in T}||X_t w_t - y_t||^2 + \frac{1}{2}\sum_{t \in T}(w_t-\mu)^T\Sigma^{-1}(w_t-\mu) + \frac{T}{2}logdet(\Sigma) \tag{3}$$

The optimization algorithm employed is minimized alternatively with respect to W and (μ, Σ). It is imperative to underline the second term of equation (3), which signifies that the precision matrix Σ^{-1} acts as a regulator for feature coefficients. This means that, since learning Σ requires training data from all tasks, the feature selection process will adhere to the patterns of all tasks, much like the algorithm in the discriminative case. Since this is regression, we use it for classification by taking the sign of the regressed value.

Once the parameters of the prior distribution are derived, the model generalizes to a new subject by deriving a new coefficient vector from that prior distribution to act as a starting point and then it adapts to the new subject's data using simple regression. For the current work, the model is adapted in 10% of the new subject's data, and evaluate in the rest 90%, because we want to evaluate the generalization capability through subjects. The code[3] that accompanies [7] was employed for the implementation of this algorithm.

The choice of the classifiers used with the pooled approach and their hyper parameters was based on extended surveys on machine learning for BCI [41][42]. The R statistical language and the caret package [43] were utilized for these algorithms and the ensemble model. The code pipeline to reproduce the analysis is open on github[4].

IV. Datasets

Neurosky was used as a recording device in both experiments. This BCI provides recordings of raw signal in 512Hz and magnitudes of frequency bands in 8Hz. The bands are Delta (0.5 - 2.75Hz), Theta (3.5 - 6.75Hz), Low-alpha or Alpha1 (7.5 - 9.25Hz), High-alpha or Alpha2 (10 - 11.75Hz), Low-beta or Beta1 (13 - 16.75Hz), High-beta or Beta2 (18 - 29.75Hz), Low-gamma or Gamma1 (31 - 39.75Hz), and mid-gamma or Gamma2 (41 - 49.75Hz). Measurements of attention and meditation are also provided, but in the current work we refrain from employing them, as we want to use information that can be found in other types of BCIs, without leaning on metrics extracted from the BCI's software.

A. CMU Dataset

The dataset from Carnegie Mellon consists of a study with 10 college students that watched Massively Open Online Course videos while being recorded. They watched two types of videos, one rudimentary and one more challenging and after each session they rated their level of confusion. 10 different 2-minute videos where shown to each student and a self-assessed confusion level is assigned by the subjects themselves. The recordings are resampled by the authors in 2 Hz. Figure 1 contains boxplots of each subject's features distribution, distinguishing the confused and non-confused cases side by side. Subject 6 was removed from the plot and the experiments due to massive differences with the rest of the subjects, which might be caused by hardware malfunction. We have scaled the features of the whole dataset to [0,1] to allow for inter-subject comparisons and removed outliers that spanned outside of 0.025 and 0.975 to produce a cleaner plot. We did not scale each subject separately because the boxplots would then be relative and the differences between subjects would not be clear. Examining it one can easily identify a pattern prevalent amongst most cases, especially in the frequency domain features. A subject's distributions display consistently higher values during confused cases. Although a subject specific classifier can take advantage of this, separating between cases of different subjects can be impossible. For example, subject 9 confused Gamma2 could resemble more of subject's 7 non-confused rather then the confused. Thus a subject-agnostic classifier could have severe trouble distinguishing between these two. Numerous underlying subject-relative patterns may exist but are not visible through simple visualization or statistical hypothesis tests. A naive Bayes non subject adaptive classification achieved 51% accuracy in this dataset [44].

B. Berkeley Dataset

The dataset from the experiment in Berkeley shares a lot of similarities with the one from CMU, in that the hardware is the same, the subjects were students, and each subject was presented videos with certain stimuli to identify and decode changes in the brain signals. That said, there are also certain differences addressing the type of stimuli and the structure of the experiment. The experiment is cross sectional, in that although the subjects are divided in two groups, all of them undergo all types of similar stimuli. Each experimental session, which lasts roughly five minutes, includes different types of stimulus, including asking to blink, doing math, listening to music, watching a video, thinking of a certain type of items and memorizing color blocks always in that order. The tasks that do not require visual aid are done with eyes closed. The stimuli's between the two groups are not always the same, but are of the same type. Having a more complex set of stimuli, and since each one is brief and just seconds away from each other, this dataset might be more challenging the the former one. Each instance in the dataset corresponds roughly to 1 second, although it is not constant. A list of 512 raw signal values is given in every instance, which is averaged to refrain from having multiple rows with only one column different. Furthermore we filter out Neurosky values with signal quality indicator different than perfect. The binary classification task we will examine is to distinguish whether the subject performs an active or passive action e.g. watching a video, listening to music, relaxing versus solving math, memorizing and coming up with items. This task can simulate the notion of mental

[3] http://brain-computer-interfaces.net/

[4] https://github.com/GiorgosPanagopoulos/Multi-task-Learning-for-Commercial-Brain-Computer-Interfaces

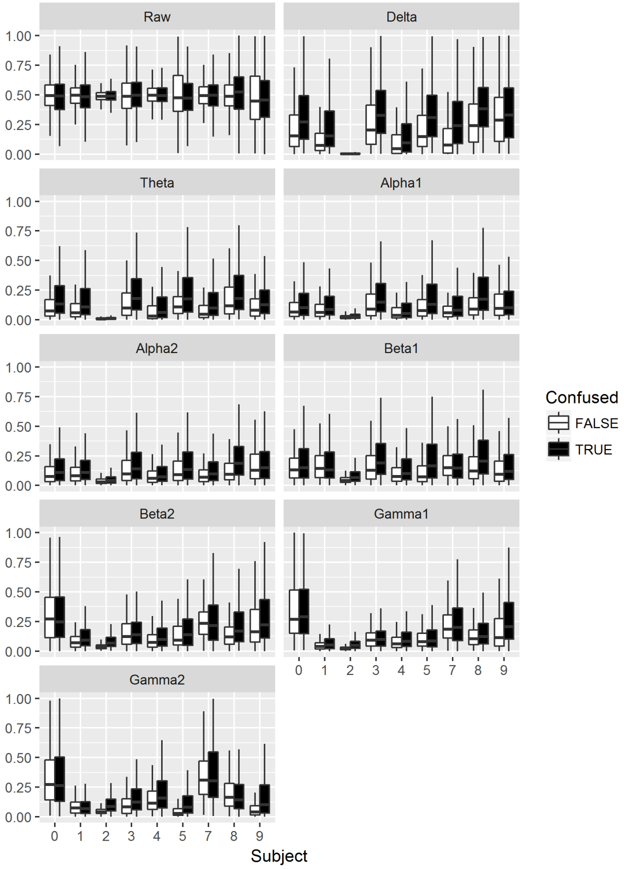

Figure 1: Boxplots of the distribution of each feature, for each subject, distinguishing between confused and non confused cases (CMU Dataset).

burden and binary classification that is also present in the CMU dataset. In addition, in this manner, we alleviate the factor of having eyes closed or open, which effects neurosky values, since both classes contain activities with eyes closed and open. Similar to CMU case, we removed 5% of the data that corresponded to outliers and scaled it. Figure 2 gives an overview of the separate distributions for all 30 subjects, distinguishing between subjects that saw video 1 and video 2 and scaled to [0,1]. In contrast with the CMU case, there is no obvious universal pattern to distinguish between classes for each subject. Still though, the inter-subject variability problem is much more prevalent here, meaning that confusing non active values of a subject with active of another subject is much easier for all features.

V. Experiment

The performance of pooled and multi-task learning methods is evaluated in both datasets using classification accuracy. Since we want to estimate the generalization through subjects, we apply 10-fold cross validation, where each fold corresponds to specific subjects' recordings, so that there is no overlap between the train and the test set subjects.

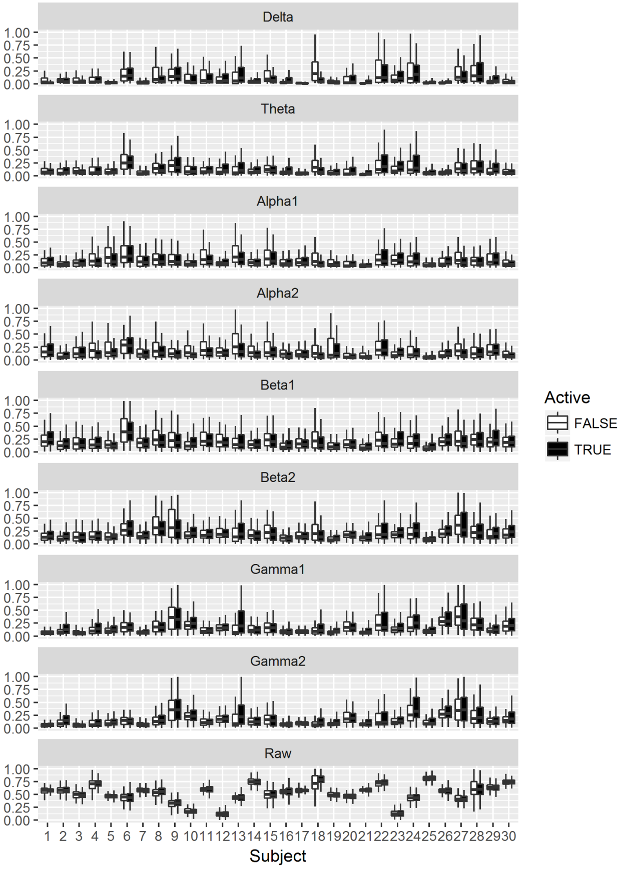

Figure 2: Boxplots of the distribution of each feature, for each subject, distinguishing between Active and Passive stimuli (Berkeley Dataset).

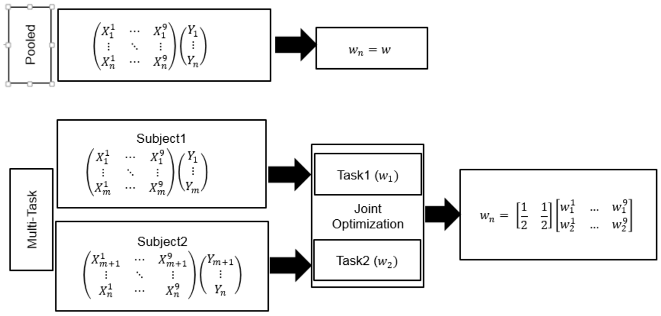
Pooled
$\begin{pmatrix} X_1^1 & \cdots & X_1^9 \\ \vdots & \ddots & \vdots \\ X_n^1 & \cdots & X_n^9 \end{pmatrix}\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$
$w_n = w$
Multi-Task
Subject1
$\begin{pmatrix} X_1^1 & \cdots & X_1^9 \\ \vdots & \ddots & \vdots \\ X_m^1 & \cdots & X_m^9 \end{pmatrix}\begin{pmatrix} Y_1 \\ \vdots \\ Y_m \end{pmatrix}$
Subject2
$\begin{pmatrix} X_{m+1}^1 & \cdots & X_{m+1}^9 \\ \vdots & \ddots & \vdots \\ X_n^1 & \cdots & X_n^9 \end{pmatrix}\begin{pmatrix} Y_{m+1} \\ \vdots \\ Y_n \end{pmatrix}$
Task1 (w_1)
Joint Optimization
Task2 (w_2)
$w_n = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \end{bmatrix}\begin{bmatrix} w_1^1 & \cdots & w_1^9 \\ w_2^1 & \cdots & w_2^9 \end{bmatrix}$

Figure 3: Difference between training using a pooled approach and a multi-task learning algorithm.

Furthermore, we use subsets of the Berkeley dataset in order to evaluate how the performance of the algorithms change with the addition of more subjects. To do this, we calculate the 10-fold cross validation accuracy in three subsets of the dataset, with 10 subjects each, and then take the average.

The results of the experiments can be seen in Table I. For the CMU dataset, which had a baseline of 51%[5], the must successful classifiers are the ones that perform well

	CMU (9 subjects)	Berkeley (10 subjects)	Berkeley (30 subjects)
Linear Discriminant Analysis [42]	63.92	63.61	57.73
Shrinkage Linear Discriminant Analysis [42]	63.92	63.61	57.73
Linear Support Vector Machine [42]	64.16	64.15	56.21
Multi Layer Perceptron [45]	58.43	55.63	45.90
Radial Basis Neural Network [46]	35.84	35.85	48.71
Learning Vector Quantization [47]	59.94	61.48	55.67
K Nearest Neighbors [48]	55.21	54.48	53.03
Decision Tree	**64.28**	**64.27**	54.23
Random Forest	63.52	63.66	57.83
Extreme Boosting	63.14	61.43	55.27
Ensemble [49]	62.69	59.91	54.34
Logistic Regression l_{21} [38]	51.29	64.15	**64.16**
Bayesian Shared Prior [7]	52.80	63.58	63.42

Table I: Accuracy percentage for each algorithm and dataset. The most accurate in each dataset are highlighted. The algorithms in bold are multi-task learning algorithms.

with traditional BCI data, namely LDA, SVM and Decision Tree, but the multi-task learning approaches perform near to random. The same algorithms are successful with a subset of 10 subjects of the Berkeley dataset. However, the multi-task learning algorithms seem to perform equally well in this case. This might stem from the fact that the CMU dataset has more subject specific information, since each subject run 10 sessions instead of one, leaving less room for inter subject information and more for subject specific. In contrast, the Berkeley dataset included less information from one subject, but more subjects. The advantage of multi-task learning becomes prevalent in the results of the whole Berkeley dataset. As the number of subjects increases, pooled approaches undergo a significant drop in accuracy, while both multi-task learning approaches keep their descent accuracy. The accuracy of logistic regression with $L_{2,1}$ norm even increases by a small percentage. This indicates the ability and robustness of these methods and verifies the fundamental multi-task learning theorem [6], which states that it becomes increasingly efficient with the number of tasks. Since our BCI setting aims for mass appeal, the number of subjects-users will be substantial, which in turn means that the model will become increasingly better as more people use it. This fact highlights the main reason why multi-task learning is suitable in this case study. Moreover, the heatmaps in figures 4 and 5 indicate the importance of features according to subject-invariant and multi-task algorithms. The darker the color, the less important was the feature for classification. As is prevalent, the multi-task algorithm has indicated roughly the same features for both datasets, in contrast with subject-invariant algorithms like Random Forest and LVQ. The highlighted Delta and Theta frequency features, verify previous work on mental activity analysis based on EEG [44] [50].

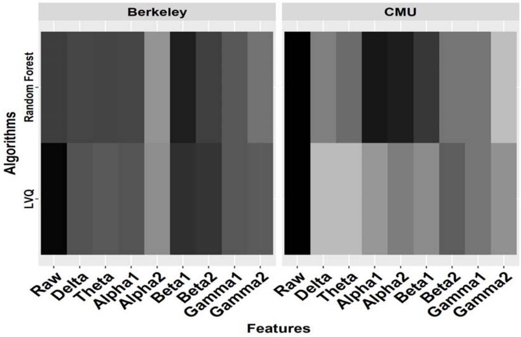

Figure 4: Feature selection of subject-invariant algorithms

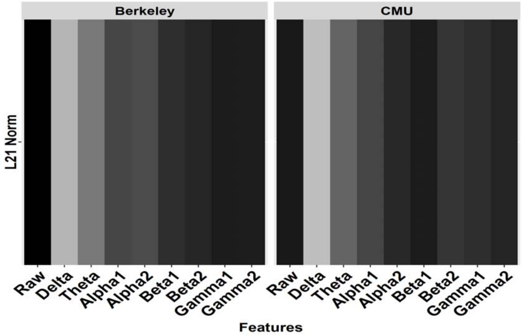

Figure 5: Weights Proportions of multi-task learning algorithm

VI. Conclusion and Future Work

In this work, we examined the efficiency of multi-task learning approaches for subject-to-subject generalization in mental monitoring using commercial EEG hardware. Algorithms used with expensive EEG hardware for active or reactive BCIs, were applied for the first time towards economic, passive BCI and were compared with multi-task learning approaches. The experiments revealed that multi-task algorithms can perform equally well with the state of the art for limited number subjects and demonstrate considerable robustness, or even enhancement, when the number of subjects increases, where the pooled approaches suffer from considerable drop in accuracy. In addition, they extract consistently features that agree with the domain literature.

Given that our BCI framework targets commercial settings with massive usage, the multi-task methodology seems ideal to increase the generalization capability of the BCI between users and alleviate the subject specific calibration problem. For future work, a first step is to evaluate the accuracy of the multi-task learning models to new data coming from subjects in the training set, instead of new unseen subjects and compare it with fully adaptive approaches. Moreover, dynamic algorithms that take advantage of the sequential nature of the recordings, such as Hidden Markov Models and Recurrent Neural Networks, need to be examined, together with their multi-task counterparts. Furthermore, the addition of datasets with elevated number of subjects is essential, to explore how much the multi-task methods improve with more tasks. Finally, experiments with datasets from different commercial BCI devices are also a meaningful extension, to argue about the robustness of the methodology throughout different hardware.

References

[1] D. J. McFarland, L. M. McCane, S. V. David, and J. R. Wolpaw, "Spatial filter selection for eeg-based communication," *Electroencephalography and clinical Neurophysiology*, vol. 103, no. 3, pp. 386–394, 1997.

[2] M. Duvinage, T. Castermans, M. Petieau, T. Hoellinger, G. Cheron, and T. Dutoit, "Performance of the emotiv epoc headset for p300-based applications," *Biomedical engineering online*, vol. 12, no. 1, p. 56, 2013.

[3] T. O. Zander and C. Kothe, "Towards passive brain–computer interfaces: applying brain–computer interface technology to human–machine systems in general," *Journal of neural engineering*, vol. 8, no. 2, p. 025005, 2011.

[4] L. George and A. Lécuyer, "Passive brain–computer interfaces," in *Guide to Brain-Computer Music Interfacing*. Springer, 2014, pp. 297–308.

[5] H. Wang, Y. Li, X. Hu, Y. Yang, Z. Meng, and K.-m. Chang, "Using eeg to improve massive open online courses feedback interaction." in *AIED Workshops*, 2013.

[6] J. Baxter, "A bayesian/information theoretic model of learning to learn via multiple task sampling," *Machine learning*, vol. 28, no. 1, pp. 7–39, 1997.

[7] M. Alamgir, M. Grosse-Wentrup, and Y. Altun, "Multitask learning for brain-computer interfaces." in *AISTATS*, vol. 10, 2010, p. 13th.

[8] V. Jayaram, M. Alamgir, Y. Altun, B. Scholkopf, and M. Grosse-Wentrup, "Transfer learning in brain-computer interfaces," *IEEE Computational Intelligence Magazine*, vol. 11, no. 1, pp. 20–31, 2016.

[9] P.-J. Kindermans, H. Verschore, D. Verstraeten, and B. Schrauwen, "A p300 bci for the masses: Prior information enables instant unsupervised spelling," in *Advances in Neural Information Processing Systems*, 2012, pp. 710–718.

[10] R. Caruana, "Multitask learning," in *Learning to learn*. Springer, 1998, pp. 95–133.

[11] R. Caruana, S. Baluja, T. Mitchell *et al.*, "Using the future to" sort out" the present: Rankprop and multitask learning for medical risk evaluation," *Advances in neural information processing systems*, pp. 959–965, 1996.

[12] H. Harutyunyan, H. Khachatrian, D. C. Kale, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," *arXiv preprint arXiv:1703.07771*, 2017.

[13] T. Heskes *et al.*, "Empirical bayes for learning to learn," in *ICML*, 2000, pp. 367–374.

[14] B. Bakker and T. Heskes, "Task clustering and gating for bayesian multitask learning," *Journal of Machine Learning Research*, vol. 4, no. May, pp. 83–99, 2003.

[15] K. Kovac, "Multitask learning for bayesian neural networks," Ph.D. dissertation, University of Toronto, 2005.

[16] T. P. Minka and R. W. Picard, "Learning how to learn is learning with point sets," *Unpublished manuscript. Available at http://wwwwhite. media. mit. edu/˜ tpminka/papers/learning. html*, 1997.

[17] K. Yu, V. Tresp, and A. Schwaighofer, "Learning gaussian processes from multiple tasks," in *Proceedings of the 22nd international conference on Machine learning*. ACM, 2005, pp. 1012–1019.

[18] E. V. Bonilla, K. M. A. Chai, and C. K. Williams, "Multitask gaussian process prediction." in *NIPs*, vol. 20, 2007, pp. 153–160.

[19] G. Varoquaux, A. Gramfort, J.-B. Poline, and B. Thirion, "Brain covariance selection: better individual functional connectivity models using population prior," in *Advances in Neural Information Processing Systems*, 2010, pp. 2334–2342.

[20] G. Leen, J. Peltonen, and S. Kaski, "Focused multi-task learning using gaussian processes," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2011, pp. 310–325.

[21] A. F. Marquand, M. Brammer, S. C. Williams, and O. M. Doyle, "Bayesian multi-task learning for decoding multi-subject neuroimaging data," *NeuroImage*, vol. 92, pp. 298–311, 2014.

[22] G. Obozinski, B. Taskar, and M. Jordan, "Multi-task feature selection," 2006.

[23] S. M. Kia, S. Vega-Pons, E. Olivetti, and P. Avesani, "Multi-task learning for interpretation of brain decoding models," in *International Workshop on Machine Learning and Interpretation in Neuroimaging*. Springer, 2014, pp. 3–11.

[24] S. M. Kia, F. Pedregosa, A. Blumenthal, and A. Passerini, "Group–level spatio–temporal pattern recovery in meg decoding using multi–task joint feature learning," *Journal of Neuroscience Methods*, 2017.

[25] R. Flamary, N. Jrad, R. Phlypo, M. Congedo, and A. Rakotomamonjy, “Mixed-norm regularization for brain decoding,” *Computational and mathematical methods in medicine*, vol. 2014, 2014.

[26] J. Zhou, “Multi-task learning and its applications to biomedical informatics,” Ph.D. dissertation, Arizona State University, 2014.

[27] L. Wang, X. Tang, W. Liu, Y. Peng, T. Gao, and Y. Xu, “Multi-subject brain decoding with multi-task feature selection,” *Bio-medical materials and engineering*, vol. 24, no. 6, pp. 2987–2994, 2014.

[28] F. Lotte and C. Guan, “Learning from other subjects helps reducing brain-computer interface calibration time,” in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 614–617.

[29] ——, “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.

[30] W. Samek, F. C. Meinecke, and K.-R. Müller, “Transferring subspaces between subjects in brain–computer interfacing,” *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 8, pp. 2289–2298, 2013.

[31] H. Morioka, A. Kanemura, J.-i. Hirayama, M. Shikauchi, T. Ogawa, S. Ikeda, M. Kawanabe, and S. Ishii, “Learning a common dictionary for subject-transfer decoding with resting calibration,” *NeuroImage*, vol. 111, pp. 167–178, 2015.

[32] S. Dalhoumi, G. Dray, and J. Montmain, “Knowledge transfer for reducing calibration time in brain-computer interfacing,” in *Tools with Artificial Intelligence (ICTAI), 2014 IEEE 26th International Conference on*. IEEE, 2014, pp. 634–639.

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

[34] S. Takerkart and L. Ralaivola, “Multiple subject learning for inter-subject prediction,” in *Pattern Recognition in Neuroimaging, 2014 International Workshop on*. IEEE, 2014, pp. 1–4.

[35] 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, no. 1, pp. 9–21, 2004.

[36] J. d. R. Millán, J. Mourino, F. Babiloni, F. Cincotti, M. Varsta, and J. Heikkonen, “Local neural classifier for eeg-based recognition of mental tasks,” in *Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on*, vol. 3. IEEE, 2000, pp. 632–636.

[37] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, “Multiclass brain–computer interface classification by riemannian geometry,” *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 4, pp. 920–928, 2012.

[38] A. Argyriou, T. Evgeniou, and M. Pontil, “Convex multi-task feature learning,” *Machine Learning*, vol. 73, no. 3, pp. 243–272, 2008.

[39] A. Ben-Tal and A. Nemirovski, *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*. SIAM, 2001.

[40] J. Zhou, J. Chen, and J. Ye, *MALSAR: Multi-tAsk Learning via StructurAl Regularization*, Arizona State University, 2011. [Online]. Available: http://www.public.asu.edu/ jye02/Software/MALSAR

[41] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi, “A review of classification algorithms for eeg-based brain–computer interfaces,” *Journal of neural engineering*, vol. 4, no. 2, p. R1, 2007.

[42] S. Lemm, B. Blankertz, T. Dickhaus, and K.-R. Müller, “Introduction to machine learning for brain imaging,” *Neuroimage*, vol. 56, no. 2, pp. 387–399, 2011.

[43] M. Kuhn, “Caret package,” *Journal of Statistical Software*, vol. 28, no. 5, pp. 1–26, 2008.

[44] H. Wang, Y. Li, X. Hu, Y. Yang, Z. Meng, and K.-m. Chang, “Using eeg to improve massive open online courses feedback interaction.” in *AIED Workshops*, 2013.

[45] M. Congedo, F. Lotte, and A. Lécuyer, “Classification of movement intention by spatially filtered electromagnetic inverse solutions,” *Physics in medicine and biology*, vol. 51, no. 8, p. 1971, 2006.

[46] M. Kaper, P. Meinicke, U. Grossekathoefer, T. Lingner, and H. Ritter, “Bci competition 2003-data set iib: support vector machines for the p300 speller paradigm,” *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 6, pp. 1073–1076, 2004.

[47] G. Pfurtscheller, D. Flotzinger, and J. Kalcher, “Brain-computer interfacea new communication device for handicapped persons,” *Journal of Microcomputer Applications*, vol. 16, no. 3, pp. 293–299, 1993.

[48] J. F. Borisoff, S. G. Mason, A. Bashashati, and G. E. Birch, “Brain-computer interface design for asynchronous control applications: improvements to the lf-asd asynchronous brain switch,” *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 6, pp. 985–992, 2004.

[49] J. Qin, Y. Li, and A. Cichocki, “Ica and committee machine-based algorithm for cursor control in a bci system,” *Advances in Neural Networks–ISNN 2005*, pp. 293–318, 2005.

[50] E. Kirmizi-Alsan, Z. Bayraktaroglu, H. Gurvit, Y. H. Keskin, M. Emre, and T. Demiralp, “Comparative analysis of event-related potentials during go/nogo and cpt: decomposition of electrophysiological markers of response inhibition and sustained attention,” *Brain research*, vol. 1104, no. 1, pp. 114–128, 2006.