Assessing Workload in Human-Machine Teams from Psychophysiological Data with Sparse Ground Truth

David Dearing*, Aaron Novstrup, and Terrance Goan

Stottler Henke Associates, Inc.
1107 NE 45th Street, Suite 310
Seattle, WA, USA
{ddearing, anovstrup, goan}@stottlerhenke.com

Abstract. Data-driven approaches to human workload assessment generally attempt to induce models from a collection of available data and a corresponding ground truth comprising self-reported measures of actual workload. However, it is often not feasible to elicit self-assessed workload ratings with great frequency. As part of an ongoing effort to improve the effectiveness of human-machine teams through real-time human workload monitoring, we explore the utility of transfer learning in situations where there is sparse subject-specific ground truth from which to develop accurate predictive models of workload. Our approach induces a workload model from the psychophysiological data collected from subjects operating a remotely piloted aircraft simulation program. Psychophysiological measures were collected from wearable sensors, and workload was self-assessed using the NASA Task Load Index. Our results provide evidence that models learned from psychophysiological data collected from other subjects outperform models trained on a limited amount of data for a given subject.

Keywords: Workload Assessment, Transfer Learning, Human-Machine Teams, Psychophysiological Sensors, Human-Automation Interaction, Machine Learning

1 Introduction

Effective human workload assessment techniques have long been sought after by researchers in hopes of preventing fatigue, stress, and other negative influences on performance. One particular application of such techniques is to diagnose performance successes and failures in human-machine teams to help identify effective training and design interventions. A wide range of research suggests that problems in such teams are greatly exacerbated by the harmful effects that high cognitive demands can have on human operator performance [1,2,3].

Traditional approaches to producing an index of workload have typically been theory-driven, following a top-down approach that begins with a hypothesis based on existing knowledge and then moves towards the measurement and quantification of the factors believed to influence workload [4,5,6]. Recently, however, there has been an increased focus on data-driven approaches [7,8,9,10]. Unlike theory-driven

approaches, these data-driven approaches can induce a workload model bottom-up from data acquired through subjective self-report measures and other measurable factors. In particular, there has been a rise in the use of psychophysiological data to generate these data-driven models, in part because such measures can be captured unobtrusively with wearable sensors and, thus, fully integrated into real-world work environments [11,12,13,14]. However, these data-driven approaches still rely on self-assessed workload ratings to serve as the labeled ground truth for training a machine learning classifier. Because these self-reports often require the full cognitive attention of the user, it is often infeasible to elicit these self-assessed ratings with great frequency (i.e., while performing attention-demanding tasks such as driving or flying).

In such situations, where there is sparse subject-specific ground truth data from which to develop accurate predictive models of workload, a more effective alternative might be to utilize models produced from the labeled data collected from other subjects. This technique, called *transfer learning*, has been used in other applications where it is expensive or impossible to collect the needed training data and rebuild the models [15]. Although the psychophysiological data from other subjects may not be completely consistent with a new subject's profile, it may still contain useful information, as people may exhibit similar responses to the same task [16].

Our research focuses on improving the effectiveness of human-machine teams through real-time human workload monitoring captured by wearable sensors. In this paper, we report on our initial efforts to evaluate the utility of transfer learning in situations where there is sparse ground truth for a given subject (i.e., labeled psychophysiological data) from which to develop accurate predictive models of workload. More specifically, we describe our progress in the context of an ongoing effort to develop an extensible modeling framework and software system for real-time human state assessment in human-machine teaming environments.

2 Design and Methodology

Our research centers on the hypothesis that in situations where there is sparse data for a particular human operator, we might learn better predictive models by utilizing not only the psychophysiological data for that operator, but also including the measures collected from a whole community of operators. In particular, we are interested in the value of transfer learning when there is a limited amount of data from which to train a predictive model for a given operator (e.g., when a new operator joins a human-machine team). We tested our hypothesis with a comparative evaluation using the approach described in the following.

2.1 Dataset

We utilized an existing dataset consisting of psychophysiological measures collected as part of a formal study conducted by the Air Force Research Laboratory (AFRL) Human Universal Measurement and Assessment Network (HUMAN)

Laboratory. In this study, a total of 13 participants were instrumented with wearable sensors to collect physiological data consisting of respiration measures, electrocardiography (ECG) measures, and electroencephalography (EEG) measures (see [17] for details). Each participant was monitored while operating a remotely piloted aircraft simulation program and workload was self-assessed using the NASA Task Load Index (NASA-TLX) [18]. Over the course of ten days, each participant experienced two training sessions and eight data collection sessions. Each session included two seven-minute trials of a target tracking task that played out along distinct scripted timelines during which the subject manually tracked either one or two targets (among other independent variables to vary the task demand). When each trial of the tracking task ended, participants were asked to fill out a new NASA-TLX questionnaire. The NASA-TLX scale, which is built upon six factors and their individual weights, has been widely used by human factors researchers over the last four decades. Each 7-minute trial of a subject's session is labeled with a single, constant composite NASA-TLX measure of perceived workload. The physiological measures were collected constantly throughout each trial.

2.2 Workload Model Training

As part of an ongoing effort to improve the effectiveness of human-machine teams through real-time human workload monitoring, Stottler Henke is developing an extensible modeling framework and software system for real-time human state assessment in human-machine teaming environments. The system, called *Illuminate*, employs machine learning techniques to induce a workload model from psychophysiological data and can—in real-time—update its internal model as new labeled data is made available. This enables us to evaluate *adaptive* models that incrementally incorporate subject-specific data as it is collected (e.g., after each session), thereby addressing complications typically associated with the analysis of biometric data, including: the non-stationarity of psychophysiological data [19] and the initial sparsity (or lack) of subject-specific data.

Data Preparation. Prior to model training, we first normalize the measures on a per-subject basis and designate each trial as either a high or low workload based on the normalized composite NASA-TLX measure (high workload being greater than the 50^{th} percentile). Psychophysiological measures are aligned by their individual timestamps and down-sampled to a frequency of 2 Hz with a five-second rolling average, so as to synchronize the measurements collected by the various sensors.

Model Training and Evaluation. Our system uses machine learning classification techniques to induce a workload model for each individual subject and uses that model to assess a subject's workload at a given moment in time. For these experiments, we utilize the Weka implementation of a multinomial logistic regression model with a ridge estimator [20,21]. For the purposes of our comparative evaluation, we trained models for each subject using three configurations:

- *Community model*: We use a leave-one-out approach in which one subject at a time is taken to be the "current" subject and the data for the other subjects provides the basis for training the other-subject community model.
- *Adaptive single-subject model*: For each "current" subject, a single-subject model is trained on the data for all *previous* sessions. That is, when evaluating data for the *n*th session, the corresponding model has been trained on all data for the current subject's first *n-1* sessions.
- *Adaptive transfer model*: As with the adaptive single-subject model, the model for each "current" subject is updated between each session so that it has been trained on data for all *previous* sessions. It uses model stacking to combine the corresponding single-subject model with the community model by including the output of the community model as an additional input to the single-subject model.

These configurations are meant to simulate a human-machine team scenario in which the system initially has a sparse amount of human operator-specific data from which to develop predictive models of workload (i.e., for a new human operator). As the operator completes additional sessions and provides workload feedback (e.g., NASA-TLX measure of perceived workload), the system updates its internal model for subsequent workload assessment. Note that each model is evaluated only on data for sessions 2 – 8—there is no single-subject model for the first session.

3 Results and Discussion

Here we describe the results of our comparative evaluation of the three configurations for model training and evaluation that are described in Section 2.3. Because the underlying models are logistic regression models with probabilistic output (as opposed to binary output), when evaluating the models on a given trial we average the workload assessment scores over all instances for that trial. To measure the performance of each model, we calculate the sensitivity and specificity metrics. Sensitivity provides an estimate of how good the model is at predicting a high level of workload, whereas specificity estimates how good the model is at predicting a low level of workload. We plot these measures on a Receiver Operating Characteristic (ROC) curve, a graphical representation wherein the points of the curve are obtained by moving the classification threshold from favoring a correct assessment of low workload to favoring a correct assessment of high workload. Each chart also indicates the point with the "optimal" threshold, maximizing Youden's J statistic [22].

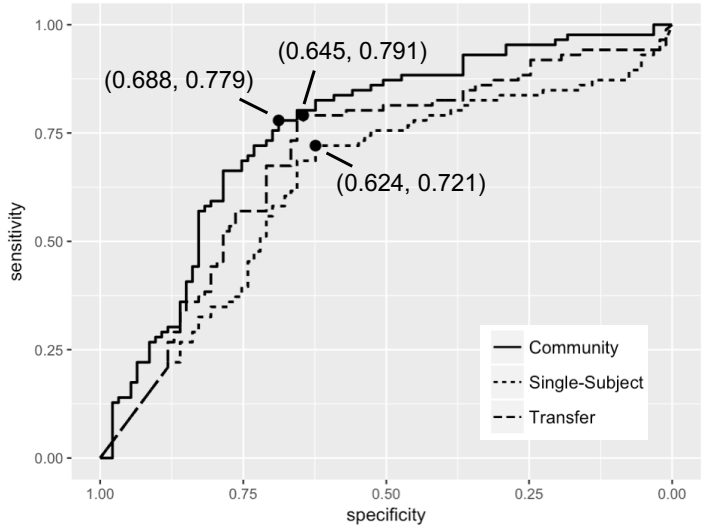

Figure 1. ROC curve comparing each model configurations, averaged across all 13 subjects.

For each of the three evaluation configurations, the chart in Figure 1 shows the ROC curve averaged across all subjects and sessions. Figure 1 shows that the community model outperforms not only the adaptive single-subject model, but also the adaptive transfer model. This provides evidence toward confirming our hypothesis that psychophysiological data collected from other subjects can be used to train a predictive workload model. However, the fact that the community model outperformed the transfer model that combines psychophysiological data for the current subject with the results of the community model was surprising.

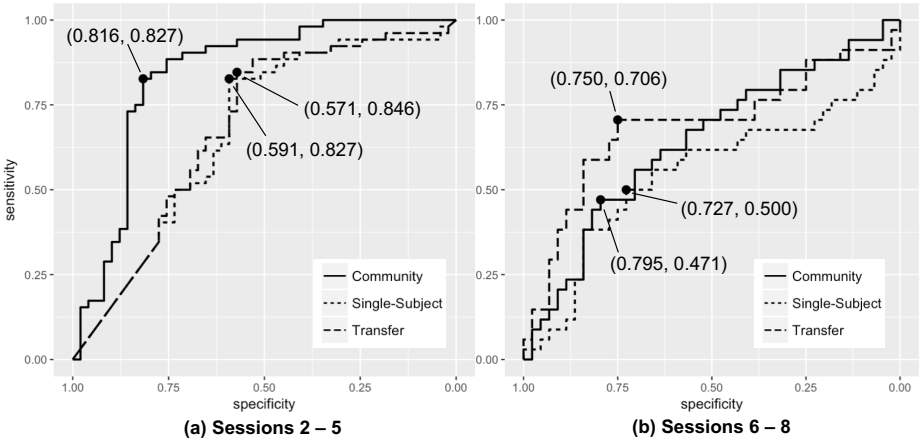

Figure 2. ROC curves comparing each of the three model configurations, averaged across all 13 subjects for (a) the first four sessions and (b) the final three sessions.

To better understand how the adaptive single-subject and transfer models performed over time (e.g., with additional training data from later sessions), we partitioned the evaluation data to compare the results of the first four sessions (sessions 2 – 5) with the results of the final three sessions (sessions 6 – 8), which resulted in partitions with roughly balanced class labels. As the ROC curves in Figure 2 illustrate, the adaptive transfer model actually outperforms the community model for sessions 6 – 8 (i.e., when there are four or more sessions of labeled training data available for each subject). Comparing the optimal points on each curve, we can see that the community model's performance drops during the later sessions (from a J of 0.643 to 0.266), whereas the performance of the transfer model improves in the later sessions (from a J of 0.418 to 0.456).

4 Conclusion and Future Work

In this paper, we have described a comparative evaluation to test our hypothesis that transfer learning is useful in situations where there is insufficient subject-specific data to develop accurate predictive models of human workload. Our results provide evidence that models learned from psychophysiological data collected from other subjects outperform models trained on a limited amount of data for a given subject. More specifically, with little or no data, a community model trained on all other-subject data performed best. Once a sparse amount of subject-specific data was available, a model induced from the output of the community model and the subject-specific psychophysiological measures generally outperformed the community model alone.

There remain several items to be answered by future work as well as by our own ongoing research. First, this evaluation does not answer the question as to at what point a single-subject model (i.e., induced only from a subject's own data) outperforms the community model or combined transfer model. Future work is needed to inspect which features contribute most to the performance of each model and how those features change across the three configurations. We are also left questioning why the community model performs worse for the later sessions. Does the community model have trouble due to the non-stationarity of psychophysiological data (whereas the other models adapt as new subject-specific data is collected)? Or, alternatively, is this an artifact of this particular dataset? Additional studies that collect data over more trials would be necessary to answer these questions.

Another topic for future research would be to vary the dividing line between what constitutes a high and low workload. Our evaluation uses the median of the normalized composite NASA-TLX measure as a simple and straightforward dividing line between high and low workload. Depending on the task and application domain, it may be more appropriate to raise (or lower) that threshold. Alternatively, to account for potential learning effects across sessions, a per-subject adaptive threshold may yield better results. The results of our comparative evaluation highlight the threshold for each model that produces the best balance of sensitivity and specificity.

However, depending on the target application, it may be more appropriate to favor specificity or sensitivity so as to more accurately predict a high or low workload, respectively.

Lastly, a more accurate model might be produced by first identifying a relevant subset of other subjects within the community data. In particular, if individual differences are high (as is often the case with psychophysiological data), a more accurate model might be induced based only on data collected from people who appear to have similar physiological responses. This is something we plan to explore in our ongoing work to improve the effectiveness of human-machine teams through real-time human workload monitoring.

Acknowledgments. This material is based upon work supported by the United States Air Force under Contract No. FA8650-15-C-6669. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Air Force.

References

1. Hancock, P.A., Williams, G., & Manning, C. M.: Influence of task demand characteristics on workload and performance. The International Journal of Aviation Psychology, vol. 5, no. 1, pp. 63--86 (1995)
2. Orasanu, J.M.: Shared problem models and flight crew performance. In: Johnston, N., McDonald, N. and R. Fuller (eds.). Aviation Psychology in Practice. Aldershot, England: Ashgate Publishing Group (1994)
3. Chen, J.Y., Haas, E.C., & Barnes, M.J.: Human performance issues and user interface design for teleoperated robots. In: IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. *37*, no. 6, pp. 1231--1245 (2007)
4. Kantowitz, B.H.: Mental workload. In: Advances in psychology. Hancock, P.A. (Ed.) vol. 47, pp. 81--121. North-Holland (1987)
5. Longo, L., Barrett, S.: A computational analysis of cognitive effort. In: Asian Conference on Intelligent Information and Database Systems. pp. 65--74. Springer, Berlin, Heidelberg (2010)
6. Moray, N. (Ed.): Mental workload: Its theory and measurement. vol. 8. Springer Science & Business Media (2013)
7. Wilson, G.F., Russell, C.A.: Real-time assessment of mental workload using psychophysiological measures and artificial neural networks. *Human factors*, vol. *45*, no. 4, pp. 635--644 (2003)
8. Zhang, J., Yin, Z., Wang, R.: Recognition of mental workload levels under complex human–machine collaboration by using physiological features and adaptive support vector machines. In: IEEE Transactions on Human-Machine Systems, vol. 45, no. 2, pp. 200--214 (2015)
9. Moustafa, K., Luz, S., Longo, L.: Assessment of mental workload: a comparison of machine learning methods and subjective assessment techniques. In: International Symposium on Human Mental Workload: Models and Applications, pp. 30--50, Springer, Cham. (2017)
10. Appriou, A., Cichocki, A., Lotte, F.: Towards Robust Neuroadaptive HCI: Exploring Modern Machine Learning Methods to Estimate Mental Workload From EEG Signals. In: Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems, p. LBW615, ACM (2018)
11. Roscoe, A.H.: Assessing pilot workload. Why measure heart rate, HRV and respiration?. Biological psychology. vol. 34, no. 2-3, pp. 259—287. (1992)
12. Verwey, W.B., Veltman, H.A.: Detecting short periods of elevated workload: A comparison of nine workload assessment techniques. Journal of experimental psychology: Applied, vol. 2, no. 3, p. 270. (1996)
13. Veltman, J.A., Gaillard, A.W.K.: Physiological workload reactions to increasing levels of task difficulty. Ergonomics, vol. 41, no. 5, pp. 656--669. (1998)
14. Prinzel III, L.J., Parasuraman, R., Freeman, F.G., Scerbo, M.W., Mikulka, P.J., Pope, A.T.: Three experiments examining the use of electroencephalogram, event-related potentials, and heart-rate variability for real-time human-centered adaptive automation design. Report TP-2003-212442, NASA, Hampton, VA: Langley Research Center. (2003)
15. Pan, S. J., Yang, Q.: A survey on transfer learning. In: IEEE Transactions on knowledge and data engineering, vol. 22, no. 10, pp. 1345--1359 (2010)
16. Wu, D., Lance, B.J., Parsons, T.D.: Collaborative filtering for brain-computer interaction using transfer learning and active class selection. PloS one, vol. 8, no. 2, e56624. (2013)
17. Hoepf, M., Middendorf, M., Epling, S., Galster, S.: Physiological indicators of workload in a remotely piloted aircraft simulation. In: 18th International Symposium on Aviation Psychology, pp. 428--433, Curran, Dayton (2015)

18. Hart, S.G., & Staveland, L.E.: Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In: *Advances in psychology*, vol. 52, pp. 139--183. North-Holland (1988)
19. Christensen, J.C., Estepp, J.R., Wilson, G.F., Russell, C.A.: The effects of day-to-day variability of physiological data on operator functional state classification. In: *NeuroImage*, vol. 59, no. 1, pp. 57--63 (2012)
20. Witten, I.H., Frank, E., Hall, M.A., Pal, C.J.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann. (2016)
21. Le Cessie, S., Van Houwelingen, J.C.: Ridge estimators in logistic regression. In: *Applied statistics*, 41(1), pp. 191--201 (1992)
22. Youden, W.J.: Index for rating diagnostic tests. In: Cancer, vol 3, no. 1, pp. 32--35 (1950)