33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

An Assessment of Non-stationarity in Physiological Cognitive State Assessment Using Artificial Neural Networks

Justin R. Estepp, *Member, IEEE*, Samantha L. Klosterman, James C. Christensen

Abstract—**With increased attention toward physiological cognitive state assessment as a component in the larger field of applied neuroscience, the need to develop methods for robust, stable assessment of cognitive state has been expressed as critical to designing effective augmented human-machine systems. The technique of cognitive state assessment, as well as its benefits, has been demonstrated by many research groups. In an effort to move closer toward a realized system, efforts must now be focused on critical issues that remain unsolved, namely instability of pattern classifiers over the course of hours and days. This work, as part of the Cognitive State Assessment Competition 2011, seeks to explore methods for 'learning' non-stationarity as a mitigation for more generalized patterns that are stable over time courses that are not widely discussed in the literature.**

I. Introduction

WHILE computational neuroscience has enjoyed immense growth as a research field over the last decade, the domains to which techniques stemming from basic research can be applied have become ever more apparent. As a framework for describing the state-of-the-art in neuroscience as applied to non-traditional areas, the term 'neuroergonomics' [1] has been coined to describe the overarching idea and potential applications. While previous terminology such as 'operational neuroscience' or 'applied neuroscience' do not necessarily differ in their meaning from that of neuroergonomics, the formulation of the word in itself demonstrates how far-reaching the benefits may be.

One specific area of research, cognitive state assessment, is perhaps one of the areas under the umbrella of neuroergonomics closest to seeing implementations in every-day life. As recently described in [2] as 'passive brain-computer interface (BCI)', this term also emphasizes the general methodology behind cognitive state assessment. It should be noted, however, that while the techniques may be similar, the application areas of cognitive state assessment and BCI are unique and different.

Differing terminology aside, physiological cognitive state assessment seeks to use neural, peripheral and behavioral measures to infer the 'state' of an individual in relation to a specific context. While not limited to the various 'states' listed here, examples in the literature can be found for states such as workload [3]-[7], vigilance [8], fatigue [9], and emotion [10].

As research can easily be cornered by perfectly controlled conditions in a laboratory environment, it is often the case that lessons learned and knowledge gained needs to be 'pushed' into real-world environments. As is the case with cognitive state assessment, which has shown great promise for augmenting human performance [11], the time has come to concern ourselves with extending the methodology such that it is robust enough to be considered for use in human-machine systems.

In the context of most research studies, the course of time over which cognitive state assessment has been vetted is on the order of hours. Even in real-time implementations of cognitive state assessment, when the encompassed time is on the order of hours, accuracy in assessing state can be very high [7]. What has not been studied, to a large extent, is the ability of the machine learning and pattern recognition techniques to withstand the 'test of time'; that is, how can models of state withstand the inherent non-stationarity in neurophysiological data [12]?

While BCI researchers are actively pursuing the answer to this question, it has largely been unaddressed with respect to cognitive state. This work, as well as the overall goal of the Cognitive State Assessment Competition (2011), is to begin investigating looming questions related to the stability and ability to generalize (resist over-fitting) of machine learning techniques using physiological data.

II. Methods

The dataset provided for the Cognitive State Assessment Competiton was collected while study participants completed the Multi-Attribute Task Battery (MATB; [13]). There were 8 participants in total, and each participant completed 3 trials on 5 separate days. In each trial, segments of task difficulty intended to produce low, medium and high workload were presented in a random order, with 'transition' time between workload segments. This transition time was 60 seconds between low/high and high/low segments and 30 seconds between low/medium, medium/low, medium/high, and high/medium segments. Each segment was 5 minutes in length.

The five days of data collection for each participant were not sequential, but spread out over the course of one month. The data collection days were randomly distributed such that each study participant had data collection days that were one day, one week (two instances) and two weeks apart.

Manuscript received 15 April, 2001. This work was supported in by the 711th Human Performance Wing Chief Scientist's Seedling Grant Program.

J.R. Estepp and J.C. Christensen are with the 711th Human Performance Wing, Air Force Research Laboratory at Wright-Patterson Air Force Base, 45433-7022 USA (email: {justin.estepp, james.christensen}@wpafb.af.mil).

S.L. Klosterman is with Wright State University, Department of Biomedical, Industrial and Human Factors Engineering, Dayton, OH 45435 USA (email: crossen.3@wright.edu).

Deviations from this paradigm were minimized to the extent possible, though some accommodations were made due to participant availability and scheduling conflicts, thus resulting in minor deflections from this ideal schedule for some participants. Table I depicts this data collection schedule for two of the eight participants.

TABLE I
EXAMPLE DATA COLLECTION SCHEDULES

SUN	MON	TUE	WED	THU	FRI	SAT
	Day 1		Day 1	Day 2		
				Day 3		
	Day 2					
	Day 3			Day 4		
	Day 4	Day 5		Day 5		

For each trial 19 channels of EEG (according to the International 10-20 System were collected, as well as peripheral measures such as ECG, VEOG, HEOG and respiration. For the competition dataset, most of the peripheral measures were omitted so that participants could concentrate their efforts on creating features from only the EEG data. Both VEOG and HEOG were included in the dataset in the event any method required them for artifact correction. In total, there were 21 channels of data (19 channels of EEG from the 10-20 System, VEOG, and HEOG) available.

All 21 of these data channels were collected using the MICROAMPS system from SAM Technologies, Inc. (San Francisco, CA, USA). MICROAMPS has default high-pass and low-pass filters at 0.05 [Hz] and 100 Hz, respectively, and a sampling rate of 256 [Hz]. Aside from these filters, no other processing was performed on the dataset. All values are in [μV]. The 19 EEG channels were referenced to a single (left) mastoid. VEOG was a bipolar channel with electrodes placed above and below the left eye. HEOG was also a bipolar channel with electrodes placed outside the outer canthus of each eye. All electrodes were tin cup electrodes (9 [mm]). Impedances for the EEG channels were all below [5 kΩ], and impedances for the VEOG and HEOG channels were all below 15 [kΩ].

The competition dataset was structured such that it closely mirrored a real-time paradigm in which the training sets used as input features to the classifiers were collected *a priori* of the test sets used to evaluate the robustness of the learning algorithm to changes in workload. In order to simulate a range of magnitudes of time between data in the training and test sets, each workload level from a session (3 per day) was split into sequential halves. These halves, denoted as 'a' and 'b', represent the first 50% and last 50% of data collected from each workload level (only differing from the real-time condition in that 'b' from the first-encountered workload level came prior to 'a' in the second-encountered workload level). For each participant and each of the first four days, training sets were created by combining all of the first session and the first half (the 'a' half) from each workload state in the second session. Test sets were then created from the remaining half of the second session (the 'b' half), all of the third session, and all three sessions from the next sequential day of data collection (thus necessitating the need to not create a training set from the fifth day, as there was no additional data for which to use as a test set). Training sets were provided with labeled truth class, but not identified to any particular participant or day. Test sets were blind with respect to truth class, participant and day; only the raw time-series data themselves were provided (with an associated label to match them to a particular training set). While this created a somewhat artificial environment in which to operate (as, in a real-world context, you would certainly know the identification of the participant and relation to any previous or planned data collections), it does facilitate a 'baseline' by which learning techniques are forced to rely only on the information provided about the workload state, thus establishing an accuracy floor for investigating methods that use the larger context of the dataset to improve accuracy.

The VEOG and HEOG channels were used to remove vertical and horizontal eye movement artifact from the EEG through a linear regression technique provided in the accompanying analysis software with MICROAMPS. From the filtered data in each of the training and test sets, input features to the classifier were generated using 5-second, non-overlapping windows of mean frequency band power (using 1-s, non-overlapping Hanning windows) at each of the 19 electrode sites, where the frequency bands used were delta (1:3 Hz), theta (4:7 Hz), alpha (8:12 Hz), beta (13-30 Hz) low gamma (31-58 Hz) and high gamma (62:100 Hz). In addition to these features, a waveform length measure called the string [14, 15] was calculated for each window using the average of 1-second, non-overlapping windows contained within (analogous to the 1-second, non-overlapping Hanning window used to compute the frequency band features). In total, 133 features were created and used as inputs to 3-layer artificial neural network (ANN), with backpropagation training, where the hidden layer contained the same number of nodes as the input layer (133), and the ouput layer contained two nodes (one for each cognitive state, or workload, class).

For each training set, estimation of the unbiased accuracy was calculated by randomly separating the data into 50% training, 25% validation and 25% test. Normalization parameters (N(0,1)) were derived from the training set and applied to the validation and within-training test sets (as well as the independent post-hoc training sets). Because of the relatively small amount of data in the within-training test set (approximately 45 samples in validation, 45 samples in the within-training test set and 90 samples in the training set), a 10-fold randomization of the training/validation/within-training test sets was used in lieu of a full 10-fold cross-validation. Trained classifiers (weights and biases for each of the 10 randomized folds) were then applied in the feed-forward direction to calculate classification accuracies on the post-hoc test sets.

With the accuracies of the within-training test set and the post-hoc set, it's possible to begin to assemble accuracies of the learning algorithm as they evolve with distance in time from the training set. To establish orders of magnitude related to time, let the within-training test set be referred to as being 'seconds' away from the training set (as each 5-s, non-overlapping, independent window, given random sampling, is likely to be on the order of seconds away from a feature vector that was used for training and/or validation). Similarly, the 'b' half of the second session can be viewed as minutes away from the training set, the third session as being an hour away from the training set, and the next sequential day as being day(s) away from the training set (at minimum, one day). For the purpose of analysis, this convention will be used to described the elapsed time between the training set and the accuracies reported for the associated test sets.

In order to test the effects of the number of days between the training set and post-hoc days test set, the results for each of the competition sets was averaged across 1-, 7- and 14-day lags between data collection sessions (per Table I). This should help to better define the non-stationarity of the physiological feature data at a finer resolution on the order of days.

III. Results

Results of the 10-fold randomization, with respect to the second/minutes/hours/days convention, are shown below in Table II. In order to avoid confusion with the less biased accuracies from the within-training test sets, accuracies for the validation sets are not reported (they were slightly higher, on average, than the within-training test set accuracies).

TABLE II
ACCURACIES FROM COMPETITION DATASET

Seconds	Minutes	Hours	Days
86.9%	73.8%	60.0%	55.8%

Re-ordering of the results in Table II to categorize each dataset by the number of days worth of lag between the training and post-hoc days test set resulted in the accuracies in Table III.

TABLE III
ACCURACIES FROM COMPETITION DATASET, AVERGED BY TIME LAG BETWEEN TRAINING SET AND POST-HOC DAY(S) TEST SET

Δ Days	Seconds	Minutes	Hours	Days
1	85.4%	69.6%	70.3%	51.9%
7	86.7%	67.9%	61.2%	59.1%
14	87.8%	78.9%	54.2%	56.2%

IV. Discussion

From the results in Table II, it is evident that there is non-stationarity in the ANN as time passes between the training set and the post-hoc test sets. While the post-hoc 'seconds' test sets were comparable to the validation accuracies (although the validation accuracies were slightly higher), there are precipitous declines in accuracy of the learning algorithm from seconds to minutes, minutes to hours, and hours to days. This non-stationarity in the physiological feature data suggest that robust results cannot be obtained using techniques largely derived from single-day laboratory research experiments.

To examine the effects of absolute number of days from training set to post-hoc test set, Table II separated the results in Table I by the delta number of days separating the two sets of interest into 1-, 7- and 14-day categories (per Table I). In general, the same pattern of non-stationarity observed in Table II can also be seen in Table III, thus suggesting that the worst-sense change in stationarity occurs at some interval between hours and days, regardless of the absolute number of days. It is also worth noting that, at accuracies only slightly better than 50% in the 'days' post-hoc test sets, the classifier is barely performing above chance (50% for the binary class case).

As an analog, these same types of non-stationary are also observed in other areas using physiological data and machine learning, such as BCI. In [16], reported waveform shape instability in raw voltage recordings using microelectrode arrays in rhesus monkeys manifests over the course of hours and days, although, as the authors note, some of this variability is likely attributable to shifts in positioning of the microelectrode array over time. [12] observes the same phenomenon in distributions of training and test features and also discusses unsupervised methods for reducing the detrimental effects of nonstationarity. To the authors' knowledge, no such attempt to mitigate nonstationarity in feature distributions in the context of cognitive state has been attempted at the time of publication, although these methods warrant a high degree of merit and are, in fact, the largest motivation for the work presented as part of the Cognitive State Assessment Competition 2011 session.

An unfortunate consequence of the competition dataset is that the non-stationarity cannot be entirely attributed to the physiological data with complete confidence. As the training sets in this implementation were very small (in comparison with the 133 input features), at least some portion of the observed non-stationarity is likely due to over-fitting of the training data (even though both validation and 'seconds' class post-hoc sets were nearly identical in accuracy at just slightly above 85%). Feature selection/reduction could have potentially improved these results, but was not implemented at the time of draft manuscript submission.

As briefly discussed in [17], the full-labeled dataset was released to competition participants after their initial analysis on the competition dataset was performed. This opens up the possibility of improving on the results shown here. One possible approach which could both a) allow the ANN to 'learn' what non-stationarity looks like with respect to the binary workload class label, and b) increase the amount of data available for training (absent feature reduction to reduce the effects of overfitting) would be to begin looking at combinations of larger time periods of data for inclusion in the training set (more sessions across more days). Post-hoc analysis performed by [18] on this same dataset revealed an increase in accuracy for the 'days' post-hoc set to nearly that seen in the validation and 'seconds' class post-hoc set (83.4% in post-hoc accuracy of 'days' class compared to the 86.9% accuracy observed for the 'seconds' class in Table II) by using combinations of four days worth of data in the training set with the independent test set being the remaining day. This suggests that by increasing the amount of 'non-stationary' data in the training set, the ANN can learn patterns related to changing distributions of the physiological data over time. What cannot be separated from this analysis, however, is the contribution of multiple days worth of data from the contribution of a larger corpus of training data (possibly on the order of a shorter lag in time between training and test) as accuracy on the remaining day increased. Simulation of this condition in a between-subjects experimental design could help to more fully resolve that ambiguity.

Even so, practitioners in human-machine systems design should begin to look toward these results as encouraging, as an accuracy of 83.4% can be of tremendous use in augmented performance paradigms where the user's (successfully) assessed cognitive state is mitigated to improve performance. As in [10], successful assessment of cognitive workload as a state, paired with mitigation implemented during periods of high cognitive demand, results in a 50% performance increase in a simulated unmanned aerial vehicle (UAV) control task. Given these results, it is reasonable to expect that robust machine learning approaches can be designed to combat non-stationarity in physiological data, thus allowing practitioners to effectively implement methods for assuring optimal cognitive state in augmented human-machine systems.

REFERENCES

[1] Parasuraman, R. and Wilson, G.F., "Putting the brain the work: neuroergonomics past, present and future", *Human Factors,* vol. 50(3), pp. 468-474, Fall 2008.

[2] Zander, T.O. and Kothe, C., "Towards passive brain-computer interfaces : applying brain-computer interface technology to human-machine systems in general,", *J. Neural Eng.*, vol. 8, pp. 1-5, March 2011.

[3] Berka., C., Levendowski, D.J., Cvetinovic, M.M., Petrovic, M.M., Davis, G., Lumicao, M.N. et al., "Real-time analysis of EEG indexes of alertness, cognition, and memory acquired with a wireless EEG headset," *Int. J. of Human-Computer Interaction*, vol. 17, pp, 151-170, 2004.

[4] Freeman, F.G., Mikulka, P.J., Prinzel, L.J., and Scerbo, M.W., "Evaluation of an adaptive automation system using three EEG indices with a visual tracking task," *Biological Psychology*, vol. 50, pp. 61-76, May 1999.

[5] Wilson, G.F. and Fisher, F., "The use of cardiac and eye blink measures to determine flight segment in F4 crews," *Aviation, Space and Environmental Medicine*, vol. 62, pp. 959-961, October 1991.

[6] Wilson, G.F. and Russell, C.A., "Operator functional state classification using psychophysiological features in an air traffic control task," *Human Factors*, vol. 45(3), pp. 381-389, Fall 2003.

[7] Wilson, G.F. and Russell, C.A., "Real-time assessment of mental workload using psychophysiological measures and artificial neural networks," *Human Factors*, vol. 45(4), pp. 635-643, Winter 2003.

[8] Mikulka, P.J., Scerbo, M.W., and Freeman, F.G., "Effects of a biocybernetic system on vigilance performance," *Human Factors*, vol. 44(4), pp. 654-664, Winter 2002.

[9] Wilson, G.F., Caldwell, J.A. and Russell, C.A., "Performance and psychophysiology measures of fatigue effects on aviation related tasks of varying difficulty," *Int. J. of Aviation Psychology*, vol. 17(2), pp. 219-247, 2007.

[10] Lin, Y., Wang, C., Jung, T., Wu, T., Jeng, S., Duann, J. et al, "EEG-based emotion recognition in music learning," *IEEE Trans. Biomed. Eng.*, vol. 57, pp. 1798-1806, July 2010.

[11] Wilson, G.F. and Russell, C.A., "Performance enhancement in a UAV task using psychophysiologically determined adaptive aiding," *Human Factors*, vol. 49(6), pp. 1005-1019, December 2007.

[12] Krusienski, D.J., Grosse-Wentrup, M., Galan, F., Coyle, D., Miller, K.J., Forney, E. et al., "Critical issues in state-of-the-art brain-computer interface signal processing," *J. Neural Eng.*, vol. 8, pp. 1-8, March 2011.

[13] Comstock, J.R. and Arnegard, R.J., "The multi-attribute task battery for human operator workload and strategic behavior research," NASA Technical Memorandum 104174, January 1992.

[14] Shelley, J. and Backs, R.W., "Categorizing EEG waveform length in simulated driving and working memory tasks using feed-forward neural netwoks," in Foundations of Augmented Cognition, Strategic Analysis, Inc., pp. 151-161, 2006.

[15] Pleydell-Pearce, C.W., Whitecross, B.T., and Dickson, B.T., "Multivariate analysis of EEG : predicting cognition basis of frequency decomposition, inter-electrode correlation, coherence, cross phase ans cross-power," *IEEE Computer Society Proceedings of the 36th Annual Hawaii International Conference on Systems Science*, pp. 131-141, 2003.

[16] Chestek, C.A., Cunningham, J.P., Gilja, V., Nuyujukian, P., Ryu, S.I. and Shenoy, K.V., "Neural prosthetic systems: current problems and future directions," *31st Annual Int. Conf. IEEE EMBS*, pp.2269-3375, September 2009.

[17] Estepp, J.R. and Christensen, J.C., " Physiological Cognitive State Assessment: Applications for Designing Effective Human-Machine Systems," *33rd Anuual Int. Conf. IEEE EMBS*, submitted for publication.

[18] Christensen, J.C., Estepp, J.R., Wilson, G.F. and Russell, C.A., "The effects of day-to-day variability of physiological data on operator functional state classification,", *NeuroImage*, submitted for publication, April 2011.