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

QUASAR's QStates Cognitive Gauge Performance in the Cognitive State Assessment Competition 2011

Neil J. McDonald and Walid Soussou

Abstract—**The Cognitive State Assessment Competition 2011 was organized by the U.S. Air Force Research Laboratory (AFRL) to compare the performance of real-time cognitive state classification software. This paper presents results for QUASAR's data classification module, QStates, which is a software package for real-time (and off-line) analysis of physiologic data collected during cognitive-specific tasks. The classifier's methodology can be generalized to any particular cognitive state; QStates identifies the most salient features extracted from EEG signals recorded during different cognitive states or loads.**

I. Introduction

At present there are no direct measures of a subject's cognitive state. However, to infer the cognitive state it is possible to use psycho-physiological techniques, in which changes in physiological signals that are affected by the cognitive state are measured and then processed using mathematical algorithms. Some parametric algorithms are based upon theoretical descriptions of the relationship between the cognitive state and the relevant physiological signals. Alternatively, some non-parametric algorithms use predictive models that define the cognitive state based on the statistical features of the physiologic data without inferences from *a-priori* psycho-physiological knowledge.

Early work suggested that psycho-physiological measures could provide continuous monitoring of operator's mental workload, and elucidate which cognitive modalities were most engaged. These measures could complement behavioral measures in controlling states of automation, especially during periods of low operator activity [1]. In the mid-1990s, the first biocybernetic system to estimate engagement in real-time based on scalp EEG power band ratios (alpha, beta, theta …) was developed [2]. Power band ratios were investigated and found to be better predictors of wakefulness and vigilance than any single power bands alone, with beta/(alpha+theta) being the most sensitive.

QUASAR's cognitive state classification gauge, QStates, has been developed alongside QUASAR's dry EEG sensor technology [3] and wireless EEG system [4]. QStates is a software package for real-time (and off-line) non-parametric classification of cognitive state or mental load based on physiologic data collected during cognitive-specific tasks. QStates handles multiple sources of physiologic data (EEG, EOG, ECG and EMG) as input, and preprocesses each signal appropriately in order to extract an array of features that are computed for every 2 second epoch of the task.

The features are then combined with ground truths as inputs to either train a Partial Least Squares (PLS) model, or are used to classify a cognitive state using an existing model. PLS is useful in situations such as EEG analysis, where the number of explanatory variables (features) exceeds the number of observations and/or a high level of multi-collinearity among those variables is assumed. In contrast to principal components analysis (PCA), PLS creates components by modeling the relationship between input and output variables while maintaining most of the input variables' information. Efficient pre-processing code and the PLS core enables rapid off-line training of models and real-time classification of physiological data streamed via a software socket.

QUASAR has used QStates to classify cognitive engagement and workload during a First Person Shooting (FPS) game and during simulated Unmanned Aerial Vehicle (UAV) control missions, on AugCog's Warship Commander simulation, as well as on X-ray screening simulation tasks. Classifications accuracies averaged across more than 30 subjects performing these varied tasks have consistently produced >90% accuracies on two-state classification for cognitive engagement, workload and fatigue [5], [6]. In addition, QStates has a linear gauge whose output reliably produces values that correlate with variable task difficulty. However, for the binary classification task in this paper, this output is not used nor discussed.

II. Cognitive State Assessment Competition 2011

The Cognitive State Assessment Competition 2011 (CSAC 2011) aims to compare the efficiency of various cognitive state gauges at classifying cognitive workload. Data provided to the participants by AFRL were collected while subjects completed the Multi-Attribute Task Battery (MATB) [7]. There were five days of data collection for each subject (number of subjects = 8), spread out over the course of approximately 1 month.

On each day, the subject performed 3 sessions of the MATB, where the MATB task difficulty was modulated during each session to produce Low, Medium and High workload segments of 300 seconds duration. For each trial, 19 channels of EEG and 2 channels of EOG were recorded at 256 Hz, and with signal bandwidth of 0.05-100Hz.

For the competition, Low and High task data for a single

Manuscript received April 16, 2011.

W. Soussou is Vice President of Research of QUASAR, San Diego, CA 92121 USA (e-mail: wsoussou@quasarusa.com).

N.J. McDonald is with QUASAR, San Diego, CA 92121 USA (e-mail: neil@quasarusa.com).

subject from one day and the next day (e.g. Day 1 & Day 2, or Day 2 & Day 3) were combined to form a single Test Set. This corresponded to a total of 32 sets. The data for each workload level were split into two 150 second segments, resulting in 24 files for each set. Of these, 6 were reserved for training and 18 for validation. The medium workload levels were excluded so that participants were required only to provide a binary state classification (High/Low).

The Low ground truth for each test set was constructed from the Low task data from Session 1 on the first day, and the first half of the Low task data from Session 2 on the first day (and similarly for the High ground truth). The resulting 450 second training files are not contiguous, which can present difficulties during signal processing at the boundary between the two workload segments.

III. QSTATES PROCESSING

A. Signal Preprocessing

EEG and EOG data are filtered EOG artifact is removed via an adaptive algorithm using both EOG channels. The filtered data is then divided into 2 second epochs.

EEG features are derived from power spectral density (PSD) estimations calculated for each bipolar EEG channel. A feature of QStates is that it calculates total of 112 features per EEG channel, including PSD values at individual frequencies, EEG power in the α, β, θ and γ bands, and algebraic combinations of the EEG bands. For 19 channels of EEG, over 20,000 EEG features are extracted.

EOG features include the power in the VEOG and HEOG channels and a Blink index derived from the VEOG channel. The EEG and EOG features are then combined to generate a classifier feature vector.

QStates also includes the capability to extract features from EMG and ECG data. However, the CSAC 2011 data did not include these channels, so they will not be discussed further.

B. Signal Quality

A Quality of Data (QoD) metric is calculated for each epoch, which can be used for vetting epochs prior to classification. The QoD is partly based upon the observed amplitude and also on the integrity of data within an epoch.

C. Training Cognitive Models

The classification method in QStates relies on a core algorithm of regularized PLS, following the general methods described by Abdi [8]. Classification models are trained using PLS regression on the preprocessed data and their associated Ground Truths.

Normalization of each feature follows a smoothing step to reduce the impact of outliers and improve the stability of the features in the Training set. The normalization maximizes the separation of a feature between the two states based on the statistical properties of the feature in each state.

In the process of training its models, QStates identifies the most salient features between the High and Low states, and eliminates those others that individually make no significant contribution to the classification, but *en masse* may affect the final result.

D. Cognitive Classification

QStates calculates epoch feature vectors with an update rate of 2 seconds. The features are normalized according to the parameters determined by the training data for the model and a weights matrix is then used to transform the input feature vector for each epoch into an m-dimensional latent vector. In practice, we have found that m=2 is sufficient to achieve 90% classification accuracy [5], [6]. The latent vectors for consecutive epochs can be averaged with a sliding window to provide a more stable, smoothed result. The results presented in this paper use an averaging of N=5.

Classification of an epoch is an estimate of the likelihood that a given 2 second epoch (after averaging) belongs to the High state. Specifically, this is derived from the probabilities estimated using the multivariate normal probability density functions (MVNPDF) for the High and Low training data. In QStates, the classification output has been normalized to have a value between 0 (Low) and 1 (High). The division between two states is not necessarily at a point equidistant from the centers of the High and Low states in latent vector space because the distributions for the two states may be significantly different.

IV. RESULTS AND DISCUSSION

A. Classification Accuracy vs. Model Training Time

The training data were split into 60 second segments and paired High/Low segments were used singly or combined with other paired segments to train set-specific models. This approach enabled training models using segments unaffected by any boundary between two workload segments. The models generated for this analysis were: M1 (2^{nd} segment), M2 (2^{nd}, 4^{th} & 7^{th} segments), M3 (1^{st}, 2^{nd}, 4^{th} & 7^{th} segments), M4 (1^{st}, 2^{nd}, 4^{th}, 5^{th} & 7^{th} segments), M5 (1^{st}, 2^{nd}, 4^{th}, 5^{th}, 6^{th} & 7^{th} segments) and M6 (contiguous training set).

Workload classifications (High/Low) of each model were supplied to AFRL, who revealed the blind and provided accuracy scores based on the Ground Truths for each validation dataset. Files a & b are from Session 2 on Day 1, files c-f are from Session 3 on Day 1, and the remainder are from Day 2. The organizers did not reveal which sessions on Day 2 corresponded to individual files.

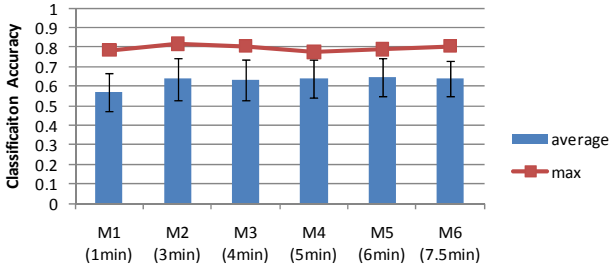

Fig. 1. Average classification accuracy for each model, averaged across all validation datasets.

Each model was found to have a maximum accuracy (averaged across the 18 validation files of each single set) of approximately 80%. The average classification accuracy across all models, averaged across all validation datasets (18 files per set, 32 sets) was 62.8% (Fig. 1). The M1 model possessed the lowest classification accuracy. However, accuracy did not improve beyond a training time of 3 minutes (i.e. models M2 through M6). Therefore the discussion that follows only considers models M1 and M6.

For comparison, as noted in the Introduction, earlier cognitive workload studies conducted using QStates reported a classification accuracy in excess of 90% for subject-specific models, averaged across all subjects.

B. *Variance in Classification Accuracy*

Removing the blind revealed that the validation files had been provided in a Low-High alternating sequence. Fig. 2 and Fig. 3 present classification accuracies for all 18 files in validation sets 17 and 25, respectively. The performances of the M1 and M6 models are reversed between these sets, with M1 and M6 having accuracies (averaged across all files) of 53.6% and 80.5%, respectively, for set 17, and 78.4% and 56.9% for set 25. However, models classifying at close to chance do not classify epochs randomly. Rather, the models appear to preferentially classify all of the validation files as a single state.

For Set 17, model M1 preferentially classifies the validation data as High. For Set 25, model M6 preferentially classifies the validation data as Low. (The exception for Set 25 is file 'b', which is the 2nd half of the High task data from Session 2 on Day 1.) It will be shown in Section C that Set 17 is one for which the training data is not stationary, and M1 predominantly classifies the training file as High.

It is interesting to note that for the poorly performing models shown in Fig. 2 and Fig. 3, the most heavily weighted features for M1 (set 17) and M6 (set 25) are γ power related, which can be strongly influenced by EMG artifact from jaw clenching or tensioning of the back and neck muscles, both often associated with increased workload. In contrast, the top features for M6 (set 17) and M1 (set 25) were terms involving α, β, and θ, and therefore less influenced by EMG.

Changes in muscle tension between testing or training conditions could potentially explain our observed classification differences (e.g., subject sporadically chewing or jaw clenching, changing posture or relaxing over course of sessions). Our experience indicates that training across these conditions has been shown to eliminate such unstable features from use in the models.

Classification accuracies for individual validation sets could be improved or degraded by as much as 18% by removing gamma features from the M6 models. However, this did not improve overall classification efficiency, when averaged across all validation sets (Fig. 4).

This analysis, however, also revealed that there is a low

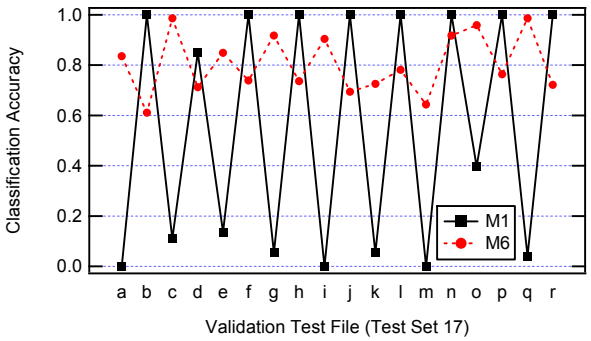

Fig. 2. Classification accuracies for each file in validation set 17.

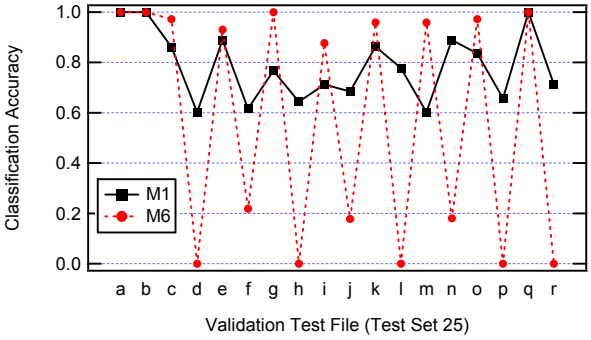

Fig. 3. Classification accuracies for each file in validation set 25.

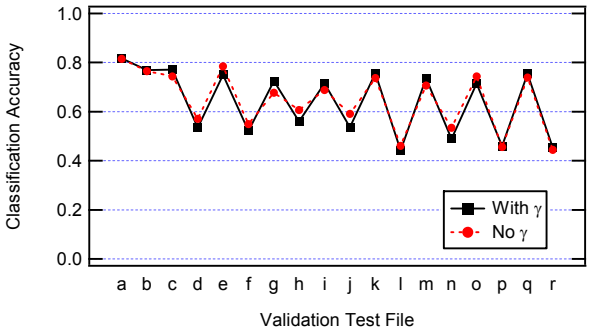

Fig. 4. Average classification accuracies for M6 models with and without γ features, averaged across all validation sets. Error bars not included for clarity.

bias in the classification of all High files from sessions not included in the training data files. Specifically, files d and f (from Session 3 on Day 1) and files h, j, k, m, o, p, and r (from Day 2) classify at near chance, compared to b (derived from a same session used for the training data), which classifies at 77.0% on average across all 18 validation sets.

C. *Stability of Ground Truths*

It was noted that for nearly half of the training datasets the M1 model had poor classification accuracies for some validation sets for one or both of the training files, characterized by a change in state (Fig. 5). In some instances, transitions at 300 seconds are consistent with the boundary between workload segments used to create the training files (Set 3 and Set 7). In other examples (Set 6 and Set 17) the transition appears to occur at the end of training data for M1 (120 seconds). Nominally identical workload tasks for sessions on the same day are notably distinct,

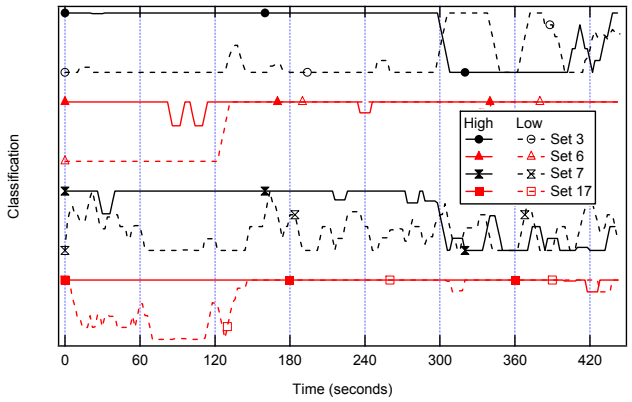

Fig. 5. Classification of training data for 4 sets for which the High and/or Low states are incorrectly classified. The models were generated using the 2nd minute of the training data set. The vertical range for each trace is 0 to 1.

suggesting a change in the subject's state not reflected by the metrics used to determine the workload ground truth.

Aside from signal integrity issues arising from boundary conditions, these transitions in the classification reflect changes in the underlying signal properties of the EEG. The PSDs plotted in Fig. 6 and Fig. 7 are for segments extracted from the Training data for Sets 3 and 6. The bipolar channels displayed were selected because they were the most salient features selected by QStates.

For Set 3, the PSD for M1 training data for the High task is characterized by greater broadband power above 5 Hz, compared with the Low training data. This is consistent for $t \leq 300$ seconds in the training dataset. However, for the data after 300 seconds this trend is reversed, and the Low and High tasks are incorrectly classified for $t > 300$ seconds.

For Set 6, the PSDs for the M1 training data (both High and Low tasks) show broadband power above 15Hz that is consistently higher than the PSDs for data $t > 120$ seconds. The High state possessed the lower PSD values above 15 Hz, and therefore both the High and Low tasks were classified as High for times > 120 seconds.

These results demonstrate that there are differences within training files that suggest unstable experimental conditions, perhaps due to changes in posture, learning effects, electrode contact, etc. For Set 3 in particular, it is difficult to correct for these differences because the relative powers of the EEG (alpha, beta, gamma) in each state are inverted between Session 1 and Session 2 on the same day. We also observed state transitions close to the beginning of test files (data not shown), suggesting that subjects' workload levels were ramping up or down on a task during test sessions.

It is worthwhile restating that 6 validation datasets were from sessions recorded on the same day as the training data and 12 were from the next day. If variability exists within training files between segments from different same-day sessions, or between segments within a single session, it is reasonable to expect that daily variability in set-up or subjects' mental or physiological condition could thus cause further inaccuracies on two thirds of the data set. The discussion of Fig. 4 had already mentioned differences

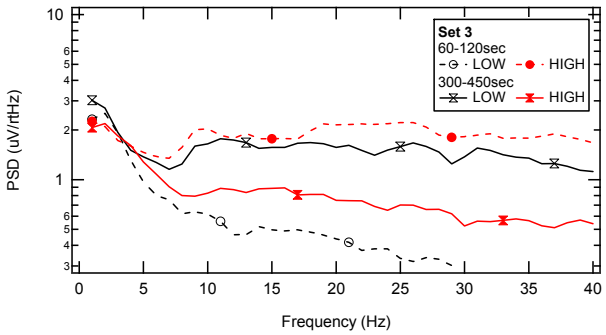

Fig. 6. Power spectral densities of Set 3 Training data (Pz-O2) in the intervals 60-120 seconds (M1 training) and 300-450 seconds. PSDs were estimated using Welch's method with 1 second window and 75% overlap. The most salient feature in the model was Pz-O2 power (4-25Hz).

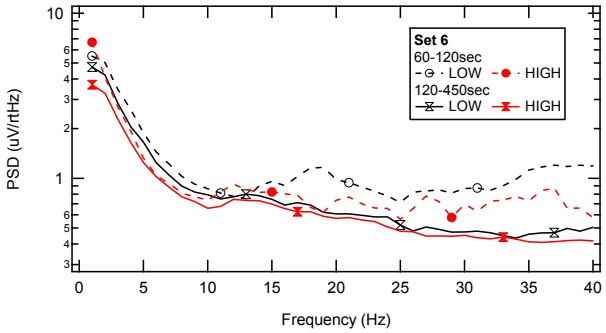

Fig. 7. Power spectral densities of Set 6 Training data (F7-T3) in the intervals 60-120 seconds (M1 training) and 120-450 seconds. PSDs were estimated using Welch's method with 1 second window and 75% overlap. The most salient feature in the model was F7-T3 gamma (26-40Hz).

between same day sessions, by showing that High state files extracted from the non-training sessions had lower classification accuracies.

Additionally, learning effects can significantly affect cognitive load. From our experience comparing expert and novice X-ray screeners performing X-ray screening tasks, we noted that novices had significantly higher cognitive load than the experts, even when every other monitored performance metric could not discriminate expertise levels. Without information about task performance, it is difficult to determine the influence of training effects to accurately classify workload and interpret the classification results.

D. *Reanalysis of Unblinded CSAC Source Data*

For this analysis, AFRL provided the competition data (with ground truths) for all 8 subjects (A through H) so that the data could be reanalyzed using QUASAR's training methodology. Specifically, QUASAR recommends using models trained using data at the beginning and end of each day to train subject-specific daily models. This is intended to remove features that vary during the course of one day's recording.

A model was created for each day using data between 120-180 seconds from the Low and High tasks in Session 1 and Session 3. Gamma features were not included. The remaining 4 minutes each for Low/High in Sessions 1 & 3, and the entirety of Session 2, were used for validation.

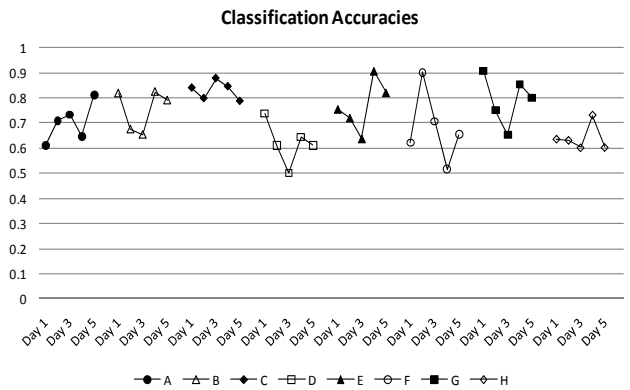

Fig. 8. Classification accuracies for subject-specific daily models for Subjects A through H.

The classification accuracy for daily models, averaged across all days and all subjects, was 72.3% (Fig. 8). For comparison, the average same-day classification accuracies of the M6 model for the CSAC dataset was 69.9%. This difference is not significant, but this result was obtained for models that used less than 1/3rd of the amount of data used to create M6 models. The principal difference is in the fact that QUASAR's methodology uses a combination of data from the beginning and end of each day for training, whereas training data for the M6 models were from the beginning of the day. In principle, QUASAR's approach provides classifiers that are less susceptible to drift in EEG features.

Using the same training data, but combining the training data from all 5 days on a subject-by-subject basis, gave a classification accuracy, averaged across all subjects, of 74.2%. This is not significantly different from the average of the results in Fig. 8. This result suggests that intersession variability is comparable to daily variation in EEG features, as indicated by the classification results presented in Fig. 4.

V. Conclusion

On this competition data set, QStates' accuracy was 62.8% when using the competition methodology. QStates was designed and optimized for rapid daily calibrations requiring 1 minute of data for each state in order to mitigate daily variations. When using this approach with the competition data set, classification accuracy went up to 72.3%. However, neither of these results are in accordance with our previous experience classifying cognitive workload, where we average >90% classification accuracy.

Classification accuracy is limited by the fidelity of ground truths to the cognitive load of the subjects. The data presented here appear to contain considerable variability in EEG features between states that are similarly labeled. These data present interesting points to examine and consider with regards to defining ground truths. Further exploration into the relationship between classification accuracy and testing day, or between accuracy and subject performance (which was not revealed when the blind was removed), may reveal interesting correlations.

Acknowledgment

The authors would like to acknowledge the assistance of Dr. Richard Shelby in modifying the QStates code to support the competition data set. We would also like to thank Andrew Agundes and Neelima Sollety for their efficient assistance in processing the CSAC dataset.

References

[1] E. A. Byrne and R. Parasuraman, "Psychophysiology and adaptive automation," *Biological Psychology*, vol. 42, pp. 249-268, 1996.

[2] A. T. Pope, E. H. Bogart, and D. Bartolome, "Biocybernetic system evaluates indices of operator engagement," *Biological Psychology*, vol. 40, pp. 187-196, 1995.

[3] R. Mathews, N. J. McDonald, H. Anumula, L. J. Trejo, "Novel Hybrid Sensors for Unobtrusive Recording of Human Biopotentials," in *Foundations of Augmented Cognition,* 2nd ed., San Ramon, 2006, pp. 91-101.

[4] R. Matthews, P. J. Turner, N. J. McDonald, K. Ermolaev, T. McManus, R. A. Shelby, and M. Steindorf, "Real time workload classification from an ambulatory wireless EEG system using hybrid EEG electrodes," in *Engineering in Medicine and Biology Society, 2008. EMBC 2008*, Proceedings of the 30th Annual International Conference of the IEEE, 20-25 August, 2008, pp5871-5875.

[5] Army Contract #W91ZLK-08-C-0011, "Neurophysiological Status Monitor," Final Report, 20th October, 2010.

[6] DHS Contract #N10PC20029, "EEG and Eye-Tracking Based Measures for Enhanced Screener Training," Final Report, 15th August, 2010.

[7] J. L. Comstock and R. J. Amegard, *The Multiattribute Task Battery for human operator workload and strategic behavior research*, Technical Report 104174. Hampton, VA, NASA Langley Research Center, 1992.

[8] H. Abdi., "Partial least squares regression (PLS-regression)," in M. Lewis-Beck, A. Bryman, T. Futing (Eds): *Encyclopedia for research methods for the social sciences*. pp. 792-795. Thousand Oaks (CA): Sage, 2003.