Pattern Recognition Letters 94 (2017) 96–104

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Deep long short-term memory structures model temporal dependencies improving cognitive workload estimation

Ryan G. Hefron[a,*], Brett J. Borghetti[a], James C. Christensen[b], Christine M. Schubert Kabban[c]

[a] Deptpartment of Electrical & Computer Engineering, Air Force Institute of Technology, WPAFB, OH 45433, USA
[b] Air Force Research Laboratory, WPAFB, OH 45433, USA
[c] Deptpartment of Mathematics & Statistics, Air Force Institute of Technology, WPAFB, OH 45433, USA

ARTICLE INFO

Article history:
Received 30 November 2016
Available online 20 May 2017

MSC:
41A05
41A10
65D05
65D17

Keywords:
Psychophysiological workload estimation
EEG
Electroencephalograph
LSTM
Long short-term memory
Temporal nonstationarity
Temporal dependence
Day-to-day variability
Time-series analysis
Recurrent neural network
Operator functional state assessment
Human-machine teams

ABSTRACT

Using deeply recurrent neural networks to account for temporal dependence in electroencephalograph (EEG)-based workload estimation is shown to considerably improve day-to-day feature stationarity resulting in significantly higher accuracy ($p < .0001$) than classifiers which do not consider the temporal dependence encoded within the EEG time-series signal. This improvement is demonstrated by training several deep Recurrent Neural Network (RNN) models including Long Short-Term Memory (LSTM) architectures, a feedforward Artificial Neural Network (ANN), and Support Vector Machine (SVM) models on data from six participants who each perform several Multi-Attribute Task Battery (MATB) sessions on five separate days spread out over a month-long period. Each participant-specific classifier is trained on the first four days of data and tested using the fifth's. Average classification accuracy of 93.0% is achieved using a deep LSTM architecture. These results represent a 59% decrease in error compared to the best previously published results for this dataset. This study additionally evaluates the significance of new features: all combinations of mean, variance, skewness, and kurtosis of EEG frequency-domain power distributions. Mean and variance are statistically significant features, while skewness and kurtosis are not. The overall performance of this approach is high enough to warrant evaluation for inclusion in operational systems.

1. Introduction

Teams composed of both humans and machines can potentially work together to mitigate their respective inherent weaknesses. A computer's strength is manifested in its ability to quickly and correctly compute answers, while humans exhibit superior flexibility of response to unexpected situations. Thus, Human-Machine Teams (HMTs) promise to mitigate inherent limitations on computational decision-making in all-human teams while simultaneously reducing the brittleness and inflexibility of fully-autonomous systems [11]. Team outcomes are improved when one agent (human or computer) assists another in the right way at the right time [7]. For computers to help humans in HMTs, they must know the human's cognitive state; this knowledge can be obtained through operator functional state assessment (OFSA) [45]. Several methods of OFSA exist, which can generally be broken into two classes of measures–objective and subjective. Subjective measures usually ask the operator to evaluate themselves either during or after the task, while objective measures use a physiological sensor such as electroencephalograph (EEG) or electrocardiogram (ECG) to provide inputs to an algorithm that assesses the operator's functional state. The benefit of objective measures is that they do not interrupt the operator while performing the task [41,42]. Continuous non-interrupting state assessment is an important characteristic for viable HMTs outside the laboratory.

A key subarea of research within OFSA is mental workload estimation. Enabling the machine in a human-machine team to unobtrusively and continuously ascertain the operator's mental workload is the first step in closing the machine-to-human augmentation loop. In order for augmentation to be effective, it must be driven by an accurate estimate of mental workload [7]. A common

* Corresponding author.
E-mail address: ryan.hefron@afit.edu (R.G. Hefron).

http://dx.doi.org/10.1016/j.patrec.2017.05.020

method for estimating mental workload is to first use statistical machine learning to fit a model which enables prediction of mental workload from the physiological signals, and then use that model to make mental workload estimates from newly-gathered physiological signals [44].

The utility of an OFSA system will depend on the benefits of accurate assessment and the costs of errors. This cost-benefit tradeoff will be application-specific and different for correctly identifying high and low workload states depending on the types of augmentation tied to a given state and the consequences of incorrect/inappropriate activation or lack of activation. These errors directly impact an operator's trust in the automation, in-turn affecting future utility of that automation in a closed loop-fashion [28]. Rouse et al.[35] indicated that a 95% accuracy rate for workload estimation may be required for a system to be acceptable. Parasuraman et al. [31] went further and suggested that if the system does not approach 100% accuracy then the costs of inaccuracy and lack of trust may lead to the system being unacceptable, especially in safety-critical environments.

Unfortunately, current state-of-the-art systems are not yet able to achieve the required accuracy, due in part to the challenge of temporal non-stationarity in psychophysiological signals. This challenge relates to variation over longer periods of time and dependence within shorter periods. Both can negatively impact the generalizable long term accuracy of workload assessment systems [7]. Within shorter spans of time, signals tend to exhibit hysteresis or serial dependence. This suggests that there is inherent structure in the statefulness in the brain that can be exploited with appropriate machine learning techniques. While it is difficult to attribute this dependence to any discrete set of factors, some of the likely possibilities include consistency in default mode activity [34] and hysteresis exhibited by most physiological systems.

In the context of machine learning, temporal non-stationarity can be addressed in two ways. The first is through feature generation or selection. A better set of features will exhibit less long-term non-stationarity and will lead to better model performance. In this work, we examine several feature generation techniques to determine empirically if certain feature sets are superior to others. The second way to address non-stationarity with machine learning is to use algorithms that make different assumptions about the nature of the data being processed. As it stands, most published research on operator workload estimation implicitly assumed temporal independence from one time segment to the next. This is likely a poor assumption due to both the factors discussed above as well as longer term effects such as fatigue and performance hysteresis with mental workload transitions [22]. An example from aviation illustrates this nicely. If a pilot has just completed flying an instrument approach in instrument meteorological conditions (IMC) when an unexpected emergency requires attention, pilot workload will increase differently than if the pilot had the same unexpected emergency arise following a period of autopilot-on flight at cruising altitude in visual meteorological conditions (VMC). This simple example illustrates that what has happened in the recent past temporally, matters for operator workload assessment.

Machine learning algorithms that consider past information as well as current information when fitting models should perform better. Such algorithms must be able to learn a temporal representation of the data. A common model used for modeling temporal data is the Recurrent Neural Network (RNN). RNNs are neural networks that are able to learn sequences that are not composed of independent, identically distributed observations [16]. Rather, they are able to elicit the context of observations within sequences and accurately classify sequences that have strong temporal correlations [16]. Historically, RNNs had limitations when training models with more than 10–20 time steps which led to poor performance. Incorporating longer time-series data streams would cause computational sensitivity problems that stymied RNN training.

Recent developments have resulted in RNN architectural and training advances which mitigate these computational problems and allow much longer temporal sequences to be processed. One approach is the Long Short-Term Memory (LSTM) layer. LSTM architectures extend the length of sequences that can be considered by a RNN by overcoming computational sensitivities encountered during backpropagation [21]. For these reasons, they may offer improved workload classification accuracy over other methods when using EEG data. With these improvements in machine learning, there is no longer a reason to avoid incorporating temporal context in a workload model. We capitalize on these machine learning developments in our research.

The primary contribution of this research is demonstration of significantly improved cross-day workload classification accuracy by integrating contextually relevant algorithmic architectures with improved feature generation techniques. We statistically evaluate all combinations of mean, variance, skewness, and kurtosis of frequency-domain power distributions and contrast a variety of RNN architectures, to include deeply stacked LSTMs, with baseline algorithms and features. Both linear and Radial Basis Function (RBF) Support Vector Machines (SVMs) and single-layer feedforward Artificial Neural Network (ANNs) using mean-only features are used as baseline cases. We show that by accounting for temporal dependence using deep LSTM models trained with new feature combinations, we can maximize cross-day workload estimation accuracy resulting in a 58% reduction in classification error over baseline methods and a 59% decrease in error compared to the best published results for this dataset.

2. Background and related work

Temporal non-stationarity of electroencephalograph (EEG) signals within individuals is likely caused by a large number of intrinsic and extrinsic factors. Participant motivation and mental or physical readiness are examples of some intrinsic factors; extrinsic factors include significant differences in EEG electrode placement, changes in conductance, and different motion artifacts [8,23,30]. Due to the challenge of handling these factors, cross-day non-stationarity of EEG signals has motivated a number of related studies including several using the same dataset described below.

2.1. Dataset

Data for our investigation was used in the 2011 Cognitive State Assessment Competition [13] and was recorded during a prior human research study performed by Wilson et al. [43]. Eight participants completed scenarios within the Multi-Attribute Task Battery (MATB) [10] environment across five test days spread out over a month-long period. Monitoring, communication, resource management, and tracking tasks were presented and manipulated to induce three levels of difficulty: low, medium, and high [8,43]. Resource allocation errors, monitoring task reaction times, and communication response times were recorded and used to validate that participants experienced distinct low and high difficulty levels. Participants were trained to asymptotic proficiency prior to the first test day [43].

For each participant, horizontal electrooculogram (HEOG), vertical EOG (VEOG), and 19 channels of EEG voltages (according to the International 10–20 System) were sampled at 256 Hz. On each of the five days, each participant performed three five-minute trials at low, medium, and high difficulty for a total of nine trials per day. Trials were presented in a random ordering with transition periods in between. Each participant completed a 30 s resting baseline at the start of each session prior to the MATB task. Only six of the participants were used in our study due to missing data from two of the original eight participants [19]. Similar to Christensen et al.