A Comparison of Artificial Neural Networks, Logistic Regressions, and Classification Trees for Modeling Mental Workload in Real-Time

Allan Fong[1], Ciara Sibley[1], Anna Cole[2], Carryl Baldwin[3], and Joseph Coyne[1]

[1]Naval Research Laboratory, Washington D.C., [2]Strategic Analysis, Inc., Arlington, VA, [3]George Mason University, Fairfax, VA

Acknowledgements

-ARCH lab in the Department of Psychology at George Mason University
-Office of Naval Research's Human Performance and Education Program

Abstract

The use of eye metrics to predict the state of an individual's mental workload involves reliable and accurate modeling techniques. This study assessed the workload classification accuracy of three data mining techniques: artificial neural networks (ANNs), logistic regressions, and classification trees. The results showed that the selection of model technique and the interaction between model type and time segmentation have significant effects on the ability to predict an individual's mental workload during a recall task. The ANN and classification tree both performed much better than logistic regression with 1-s incremented data. The classification tree also performed much better with data averaged over the full recall task. In addition, the transparency of the classification tree showed that pupil diameter and divergence are significantly more important predictors than fixation when modeling 1-s incremented data.

Introduction

Motivation

By monitoring an operator's workload, it may be possible to keep the operator in his/her "sweet spot" thereby increasing his/her overall effectiveness.

Literature

-Physiological data, such as heart rate, electroencephalogram and eye metrics have been used to model and predict an individual's mental workload. *(Van Orden, et. al., 2001; Wilson and Russell, 2003)*

-Studies used ANNs to develop these workload models. *(Marshall, 2007; Wilson and Russell, 2003, 2007)*

Problem

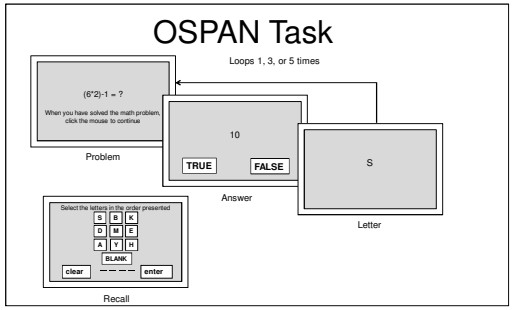
-ANNs lack transparency and are difficult to interpret.
-Understanding the interactions of predictor variables and how they influence a model's classification ability can lead to the development of more accurate models.

Purpose

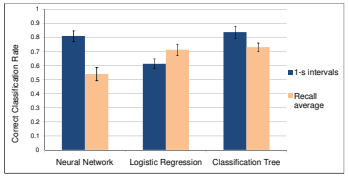
Compare the advantages and disadvantages of the ANN with two other modeling techniques: logistic regression and classification tree.

Method

-Subject performs OSPAN task
-Eye metrics and performance data are collected
Pupil diameter [Left & Right], Divergence, Fixation, Movement
-Classification models are generated using eye metrics and evaluated



Results



ANNs and classification trees performed much better than logistic regressions with 1-s incremented data. Classification trees also performed much better with data averaged over the full recall task.

	Correct Classification Rate
Time Segmentation	$F(1,8) = 2.276$ $p = 0.206$
Model Type	$F(2,8) = 13.180$ $p = 0.003$
Time x Model	$F(2,8) = 11.385$ $p = 0.005$

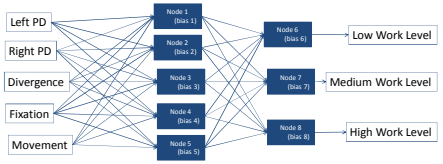
The selection of model technique and the interaction between model type and time segmentation have significant effects on the ability to predict an individual's mental workload during a recall task.

The ANN and classification tree for a subject shows similar classification rates but the logic rules in the classification tree are more transparent.

ANN	Target Class: Low	Med	High	
Output Class: Low	7	2	1	70%
Med	1	15	0	93.8%
High	0	1	11	91.7%
	87.5%	83.3%	91.7%	86.8%

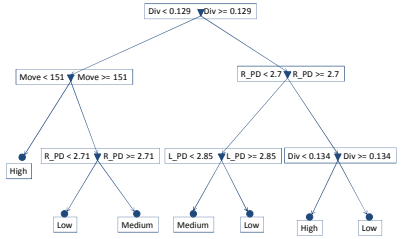
Class. Tree	Target Class: Low	Med	High	
Output Class: Low	14	0	0	100%
Med	6	31	3	75.6%
High	2	2	18	81.8%
	63.6%	93.9%	85.7%	82.9%

ANN

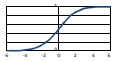


Node	Weight 1	Weight 2	Weight 3	Weight 4	Weight 5	Bias
Node 1	-0.98615	-0.67276	1.625733	0.780627	-2.4806	1.51237
Node 2	2.302286	2.387594	-3.42605	0.7072	-0.59151	0.031449
Node 3	2.413652	-2.69007	4.791826	0.252226	0.626162	1.734574
Node 4	-0.16142	-0.56592	-2.42011	2.071603	-0.15759	2.675662
Node 5	0.278498	0.051553	0.899813	0.623945	0.652556	2.648848
Node 6	1.426582	1.535971	2.362639	-2.06413	0.457082	-2.13526
Node 7	-2.18468	-1.03778	1.247581	1.770219	0.163335	-0.47508
Node 8	3.042216	0.434494	-2.86314	-0.07399	-0.1925	-2.19975

Classification Tree



Classification Methods

Method	Model Description	Example	Strengthens	Limitations
Artificial Neural Network	Input and output layers connected by hidden layer(s) with trained weights and biases		-Good predictive performance -Handles complex relationships well -High tolerance to noisy data	-"Black box" approach -Relies heavily on having sufficient training data -Slow run-time
Ordinal Logistic Regression	Linear regression concepts applied to dependent variables that are categorical using the logistic function		-Does not assume linear relationships between dependent and independent variables -Incorporates ordinal information	-Assumes linear relationships between the independent and the log odds of the dependent variables
Classification Tree	Binary tree structure partitioned to reduce the amount of impurities after each split		-Good for variable selection -Robust to outliers and at handling missing data	-Sensitive to small changes in data -Can overlook relationships between predictors

Conclusion

-ANNs and classification trees exhibit similar classification rates.
-Classification trees provide more transparency.
-This research can assist in the development of a model for real-time training systems capable of adapting to an individual's mental workload.

HFES 54th Annual Meeting, San Francisco, CA
September 30, 2010

Contact Information
Allan Fong (allan.fong@navy.mil)
Ciara Sibley (ciara.sibley@nrl.navy.mil)

A Comparison of Artificial Neural Networks, Logistic Regressions, and Classification Trees for Modeling Mental Workload in Real-Time

The use of eye metrics to predict the state of one's mental workload involves reliable and accurate modeling techniques. This study assessed the workload classification accuracy of three data mining techniques; artificial neural network (ANN), logistic regression, and classification tree. The results showed that the selection of model technique and the interaction between model type and time segmentation have significant effects on the ability to predict an individual's mental workload during a recall task. The ANN and classification tree both performed much better than logistic regression with 1-s incremented data. The classification tree also performed much better with data averaged over the full recall task. In addition, the transparency of the classification tree showed that pupil diameter and divergence are significantly more important predictors than fixation when modeling 1-s incremented data.

INTRODUCTION

The ability to apply cognitive workload models and theories in real-time is becoming more practical as physiological sensors and processing algorithms advance and the computational power to process these data become more ubiquitous. To apply indices of mental workload to real-time training and operational scenarios, reliable and predictive models have to be developed to correctly interpret these data. Several papers have used physiological data, such as heart rate, electroencephalogram (EEG) and eye metrics to model and predict an individual's mental workload (Van Orden, Limbert, Makeig, & Jung, 2001; Wilson and Russell, 2003). Incorporating information about an operator's mental state into training scenarios can provide the operator with more effective training scenarios that are tailored to the individual's workload and skill level at any given time. Real-time information about an operator's mental state can also allow an intelligent system to better complement the operator by adjusting its level of automation and supervisory control.

Several studies used ANNs to develop these workload models (Marshall, 2007; Wilson and Russell, 2003, 2007). Although ANNs are robust and have good predictive powers, they lack transparency and are difficult to interpret. Understanding the interactions of predictor variables and how they influence a model's ability to classify workload is especially important when developing models to be generalizable across individuals and different tasks. This paper compares the advantages and disadvantages of the ANN with two other modeling techniques; logistic regression and classification tree.

Eye Metrics

Different eye metrics have been assessed in several research papers. Amongst the most popular are pupil dilation, blinks, divergence, fixation, and saccades (Sweller, 2006; Tokuda, Palmer, Merkle, & Chaparro, 2009; Van Orden et al., 2001). Pupil diameter and divergences have been shown to correlate well with workload (Marshall, 2007; Moresi et al., 2008). Fixation and dwell time measurements are more task dependent but have been shown to correlate with mental workload as well (Marshall, 2007). This present study used pupil diameter, divergence, and fixation metrics because these measurements are more reliable and accurate compared to blink and saccade data, given the use of the Tobii X120 eye tracker setup. Because the subject's head is not fixed in location and orientation in this experiment, the Euclidean distance between pupil centers is used for the divergence metric. This accounts for head movement and tilt. Fixation is an incremental count that simply increases if two consecutive gazes are within ten pixels of each other. Pilot data analysis suggested that ten pixels offered good resolution and separation of the data but a more rigorous approach needs to be taken to determine the impact of different fixation cut-off intervals. EEG data was also collected which could be used to identify blink frequencies and will be incorporated into future analysis.

Classification Models

There are several classification models that are useful for data mining and developing predictive models (Shmueli, Patel, Bruce, 2006). These methods have been widely used in several disciplines, such as medicine and finance (Detsky et al., 2007; Shmueli et al., 2006). However, selecting amongst models requires understanding their different strengths and weaknesses as well as assumptions concerning the data. ANN, ordinal logistic regression, and classification tree techniques are commonly used to develop predictive nonlinear models for categorical responses which are suitable for eye metrics. Table 1 offers a short summary comparing these methods as adopted from Shmueli, Patel, and Bruce (2006).

This study compared the usefulness of different modeling techniques to predict mental workload. ANN, ordinal logistic regression, and classification tree models are all capable of capturing non-linear relationships between predictor and response variables. These three models were assessed on their ability to correctly classify mental workload state for two time segmentation conditions.

METHOD

Participants

Data was collected from 12 university student volunteers (7 males and 5 females) ranging from 18 to 30 years of age. All had normal or corrected-to-normal vision. Data from three subjects were not used because they were not sufficient for data mining purposes. Data from additional subjects will be incorporated in later analysis to increase the power and significance of this study.

Table 1: A short comparison of ANN, classification tree, and ordinal logistic regression models

Method	Data Description	Strengthens	Weaknesses
Artificial Neural Networks	-Data can be nonlinear and nonparametric	-Good predictive performance -Handles complex relationships well -High tolerance to noisy data	-"Black box" approach -Rely heavily on having sufficient training data -Slower run-time
Ordinal Logistic Regression	-Continuous input variables -Data can be nonlinear and nonparametric	-Does not assume linear relationship between dependent and independent variables -Incorporates ordinal information	-Assumes linear relationships between the independents and the log odds of the dependents
Classification Tree	-Data can be nonlinear and nonparametric -Performance better with categorical predictors	-Good for variable selection -Robust to outliers and in handling missing data	-Sensitive to small changes in data -Can overlook relationships between predictors

OSPAN Task

Participants completed a modified version of the Automated Operation Span (OSPAN) Task that has been previously used to measure an individual's working memory capacity (Turner and Engle, 1989; Unsworth, Heitz, Schrock, & Engle, 2005). Processing information in working memory is closely associated with, and could arguably be considered synonymous with, mental workload (Parasuraman and Caggiano, 2005). As working memory processing requirements increase, mental workload increases. For the OSPAN task, subjects were first presented with a series of basic arithmetic questions in varying set sizes and were given a limited amount of time to provide an answer. After the subject submits an answer to an arithmetic problem, a memory stimulus (letter) is displayed. Following presentation of all items in the series set, participants are asked to recall the memory items from that trial in the correct order. The modified version in this study used three levels of OSPAN difficulty (low, medium, and high). The levels varied based on the number of letters the subjects were required to remember (2 letters, 4 letters, and 6 letters respectively). Figure 1 provides a graphical illustration of the successive visual screens presented in the modified OSPAN task.

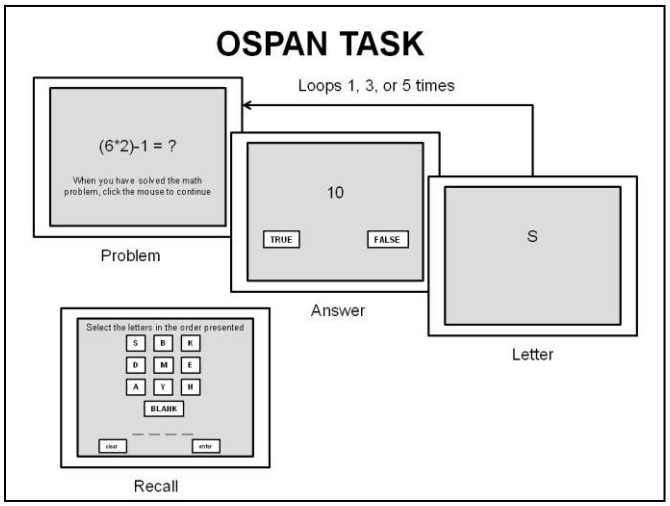


Figure 1: OSPAN Task

Apparatus and Metrics

The Tobii X120 Eye Tracker was used in this experiment to collect pupil diameters and gaze coordinates of both left and right eyes. As discussed earlier, pupil diameters, divergence, and fixation, were input predictors for all three models. This analysis focused on the recall portion of OSPAN, which is when the participants can be expected to experience the most mental workload across the levels. Furthermore, the data were analyzed using two different time segmentations; one second averages and recall averages. The 1-s interval was calculated from a running average over a 10-s window, a method used in previous studies (Marshall, 2007; Van Orden et al., 2001). This time segmentation manipulation was used to provide a better understanding of how modeling will behave in real-time or near real-time scenarios. The recall averaged data were segmented and averaged for each trial which provided a much coarser description of the data.

Analysis

This research used ANN, logistic regression, and classification tree models to predict an individual's mental workload state based on eye metrics. This paper looked at the performance differences in correctly classifying mental workload conditions between the three data mining techniques and temporal segmentation previously discussed. A 2x3 ANOVA was performed on the classification rate data. There are two time segmentation predictor levels (1-s intervals and recall average) and three modeling techniques (a feed-forward neural network trained with backpropagation, a logistic regression modeled with maximum likelihood estimation, and a recursively partitioned classification tree). Lastly, the order of predictor importance was analyzed from the models generated using the Kruskal-Wallis non-parametric test.

RESULTS

After meeting the basic assumptions for homoscedasticity, the 2x3 ANOVA showed a significant interaction effect between the time segmentation groups and model type on classification rate (Table 2).

Table 2: Results from the two statistical analysis

	Correct Classification Rate
Time Segmentation	$F(1,8) = 2.276$ $p = 0.206$
Model Type	$F(2,8) = 13.180$ $p = 0.003$
Time x Model	$F(2,8) = 11.385$ $p = 0.005$

Model type also has a significant effect on classification rate ($p = 0.003$). This significance of model type was driven primarily by the performance of the classification tree as shown in Figure 2. The interaction effect is also apparent in Figure 2, particularly in the decrease of classification performance for the neural network with recall averaged data.

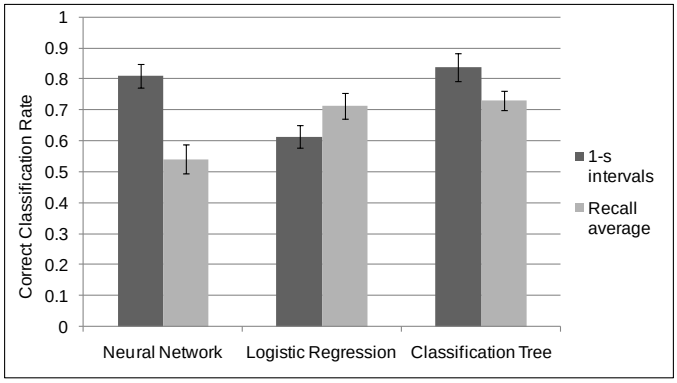


Figure 2: Significant model and interaction effects on correct classification rates (error bars are standard error of the means)

This analysis also showed the importance of pupil diameter and divergence as predictor variables. The classification tree models were used for this analysis because they provided the best overall modeling performance. Furthermore, one of the advantages of classification tree models is that the most important predictors will end up at the top of the tree (Shmueli et al., 2006). The predictors are ranked in order of importance with three being the most important and one being the least important. Figure 3 shows the sum of the ranks for the different predictors. Because of the data's ranked nature, non-parametric Kruskal-Wallis tests were used to assess significance. Predictor ranks were significantly different for the 1-s data ($p < 0.001$) which was driven primarily by the diameter and divergence scores. Predictor ranks were marginally significant for recall averaged data ($p = 0.089$) suggesting that fixation data is relatively more important in recall averaged models than in the 1-s models.

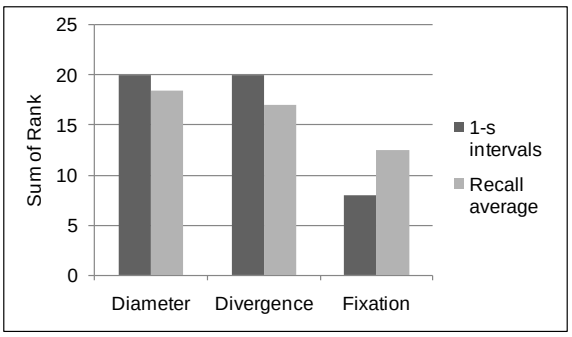


Figure 3: Sum of ranks for classification tree predictor variables

DISCUSSION

The results from this study highlight the benefits of using ANNs and classification trees for real-time applications. The ANN performs well with the 1-s interval data, however its performance worsens with recall averaged data which is most likely caused by an insufficient amount of training data. This can be mitigated with multiple model runs while randomizing the data selected for training. The classification tree generated some of the most successful models with both methods of time segmentation. Besides having high correct classification rates, classification tree models are more transparent and easier to interpret, especially for indentifying the relative importance of predictor variables as well as their critical values. Classification tree models can be decomposed into a series of logic statements. As a result, adjustments and calibrations in classification trees are much easier to understand than changes to weights and biases in neural networks. Furthermore, information about important predictor variables and their ranges will be helpful both for developing and calibrating models generalizeable for populations or for individuals performing different tasks.

This study also showed a significant interaction effect between the model type and time segmentation. Data averaged every second is more robust and useful for real-time application than data averaged through a recall task. Having 1-s incremented data produces better performing models, especially with the ANN and classification tree. However, there is a tradeoff between processing time and the fidelity of the time segments. It is important to further study how this tradeoff influences the performance of both ANNs and classification trees.

Furthermore, the results from the Kruskal-Wallis tests highlight the importance of pupil diameter and divergence in real-time models. Data fluctuations occur more in 1-s data and this fluctuation is better captured by the pupil diameter and divergence metrics. However, more generalizations were made about the variability in recall averaged data, which might explain the increased need to incorporate fixation in these models. This again emphasizes the need to understand how models perform with different time segmentations. Future studies will address these and additional questions about using real-time or near real-time data for modeling purposes.

Classification rate metrics show significantly better performance of the ANN and classification tree over logistic regression for this dataset. Additional metrics will be used to apply more rigor in the assessment of ANNs and classification trees especially for different time segmentations. Further analysis will compare Receiver Operating Characteristics curves and the effects of Type I and Type II errors for the different workload levels. Future studies will also include performance and EEG data as additional predictors and examine workload classification in a wider range of tasks.

CONCLUSION

One goal of this research is to develop a real-time training scenario that can adapt to an individual's mental workload. Adaptive training can help individuals learn more efficiently and effectively by preventing individuals from becoming severely over-stressed or bored. Having an appropriate model to predict ones workload from physiological data is a very important aspect of this initiative. Although this study shows that ANN and classification trees have comparable predictive performance with real-time data, classification tree models are more transparent and easier to interpret. This makes identifying important predictor variables, such as pupil diameter and divergence, and critical values much more intuitive. These attributes will be advantageous for developing generalizable model for a population and for different tasks.

ACKNOWLEDGEMENTS

The authors are very thankful for the help and support of the students in Dr. Carryl Baldwin's lab in the Department of Psychology at George Mason University. This research was supported by the Office of Naval Research's Human Performance and Education Program.

REFERENCES

Detsky, A. S., Naglie, G., Krahn, M. D., Redelmeir, D. A., & Naimark, D. (1997). Primer on Medical Decision Analysis. *Medical Decision Making*, 17(2), 126-135.

Marshall, S. P. (2007). Identifying Cognitive State from Eye Metrics. *Aviation, Space, and Environmental Medicine,* 78, B165-B175.

Moresi, S., Adam, J. J., Rijcken, J., Van Gerven, P. W. M., Kuipers, H., & Jolles, J. (2008). Pupil dilation in response preparation. *International Journal of Psychophysiology,* 67(2), 124-130.

Parasuraman, R., & Caggiano, D. (2005). Neural and Genetic Assays of Mental Workload. In D. McBride & D. Schmorrow (Eds.), *Quantifying Human Information Processing* (123-155). Lanham, MD: Rowman and Littlefield.

Shmueli, G., Patel, N. R., & Bruce, P. C. (2006). Data Mining for Business Intelligence. Hoboken, New Jersey: John Wiley & Sons, Inc.

Sweller, J. (2006). Discussion of Emerging Topics in Cognitive Load Research: Using Learner and Information Characteristics in the Design of Powerful Learning Environment. *Applied Cognitive Psychology*, 20, 353-357.

Tokuda, S., Palmer, E., Merkle, E., & Chaparro, A. (2009). Using Saccadic Intrusions to Quantify Mental Workload. *Human Factors and Ergonomics Society Annual Meeting Proceedings*, 53, 809-813.

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory & Language*, 28, 127-154.

Unsworth, N., Heitz, R.P., Schrock, J.C., & Engle, R.W. (2005). An automated version of the operation span task. *Behavior Research Methods*, 37(3), 498-505.

Van Orden, K. F., Limbert, W., Makeig, S., & Jung, T.P. (2001). Eye Activity Correlates of Workload during a Visuospatial Memory Task. *Human Factors*, 43(1), 111-121.

Wilson, G. F., & Russell, C. A. (2003). Real-Time Assessment of Mental Workload Using Psychophysiological Measures and Artificial Neural Networks. *Human Factors,* 45(4), 635-643.

Wilson, G. F., & Russell, C. A. (2007). Performance Enhancement in an Uninhabited Air Vehicle Task Using Psychophysiologically Determined Adaptive Aiding. *Human Factors*, *49*(6), 1005-1018.