Workload Classification Across Subjects Using EEG

EEG data has been used to discriminate levels of mental workload when classifiers are created for each subject, but the reliability of classifiers trained on multiple subjects has yet to be investigated. Artificial neural network and naive Bayesian classifiers were trained with data from single and multiple subjects and their ability to discriminate among three difficulty conditions was tested. When trained on data from multiple subjects, both types of classifiers poorly discriminated between the three levels. However, a novel model, the naive Bayesian classifier with a hidden node, performed nearly as well as the models trained and tested on individuals.

Adaptive automation technologies promise to meliorate the demands made on mental capabilities by modern automation and computerized systems (Moray, Inagaki, & Itoh, 2000; Inagaki, 2003). A critical aspect of these adaptive technologies is accurate and reliable assessment of operator workload. Traditionally, workload is assessed by questionnaires which are quantified through statistical techniques such as factor loading, discriminant analysis, and correlation/covariance analysis (Hart & Staveland, 1988). Although progress has been made, there are no globally accepted methods for measuring and predicting workload (Lysaght et al., 1989; Rubio, Diaz, Martin, & Puente, 2004; Noyes & Bruneau, 2007). In addition, subjective measures are invasive and cannot be obtained in real-time as they require interrupting the task to complete a questionnaire. As a result, many researchers have moved towards using electrophysiological measures to predict workload (Gevins et al., 1998; Gevins & Smith, 2003). In particular, electroencephalography (EEG) has been used extensively to examine the changes in the brain's electrical activity in response to cognitive activity. The main assumption is that if brain-state classifiers can be found, then they can be used by a brain-computer interface (BCI) in real-time to detect operator mental workload (Wilson & Russell, 2007).

While a number of different classifier algorithms have been used with EEG data, such as linear discriminant analysis, support vector machines and artificial neural networks, it is not clear which method is superior (see (Bashashati, Fatourechi, Ward, & Birch, 2007) and (Lotte, Congedo, Lécuyer, Lamarche, & Arnaldi, 2007) for extensive reviews). Artificial neural networks (NN) are by far the most popular classifier and have shown success discriminating at least two levels of cognitive workload (Wilson & Fisher, 1995; Wilson & Russel, 2003a, 2003b; Wilson, Estepp, & Davis, 2009; Wilson, Estepp, & Christensen, 2010). In this paper we present the first application of naive Bayesian models to the detection of cognitive workload and compare these to NNs. To preview our results, the best workload classifiers are Bayesian and a Bayesian variant trained on data from all participants does almost as well as the set of Bayesian models trained on individual performers (i.e., one model for all participants versus a separate model for each participant).

The Promise of EEG

In principle, EEG provides an objective and relatively unobtrusive means for measuring workload. In practice, much work needs to be done in the development of quantitative methods for analyzing and interpreting EEG data. Training classifiers is time consuming and requires a lot of data, especially in situations that involve multiple subjects. Currently, the standard practice is to train a new classifier for each subject. Recent research suggests it might even be necessary to train new classifiers each day (Wilson et al., 2010). One way to potentially reduce overall training time is to train one model across subjects. The large variability between subjects poses a significant challenge to building a common classifier and has not previously been investigated. With traditional techniques such as NN, it seems likely that the classifier would not separate signal from noise across multiple subjects. However if measures can be incorporated to account for between subject variation, such a classifier might produce more robust and stable classifications.

The present paper has two goals. The primary goal is to investigate the effect of between subject variability on workload classifier accuracy. The secondary goal is to compare the performance of NNs to Bayesian graph-

ical models. In order to accomplish these goals, we first compare the accuracy of a NN, a standard naive Bayesian classifier, and a novel naive Bayesian classifier with a hidden node when each are trained on EEG data from multiple subjects and tested on individual subjects. We then compare these classifiers to the performance of NNs and standard naive Bayesian classifiers which were trained and tested on single subjects. Since naive Bayesian classifiers have not been used as workload classifiers before, a little time needs be spent discussing how they function before getting into the details of our experiment.

Naive Bayes Classifier

Naive Bayes Classifier (NB) is a very simple classifier based on the Bayes' theorem. Its structure is shown in Figure 1. The C node represents different classes and

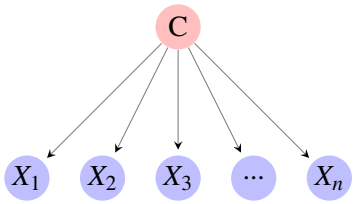

Figure 1. Naive Bayes Classifier

$X_1, X_2, \cdots, X_n$ represent different components or features of a sample. NB assumes all the feature nodes are independent of each other given the class, and typically, the feature variables are assumed to have Gaussian distribution if they are continuous. Despite its naive design and apparently over-simplified assumptions, NB has worked quite well in many complex real-world situations. Compared to other complex graphical models, it requires smaller amount of training data to accurately estimate the parameters necessary for the classification (Zhang, 2004).

The classification results are determined by the posterior probability $P(C|X_1, X_2, ..., X_n)$, which can be transformed using the chain rule and Bayes' theorem into equation 1–3.

$$P(C|X_1, X_2, \cdots, X_n) = \alpha P(X_1, X_2, \cdots, X_n|C)P(C) \quad (1)$$

$$= \alpha \prod_{i=1}^{n} P(X_i|C)P(C) \quad (2)$$

$$= \underset{C}{argmax} \prod_{i=1}^{n} P(X_i|C)P(C) \quad (3)$$

Given a test sample $(X_1, X_2, \cdots, X_n)$, the class is determined by equation 3. In our case, the class node represents three workload conditions and the feature nodes $(X_1, X_2, \cdots, X_n)$ represent the EEG frequency features.

Hidden Node Naive Bayes Classifier

The between subject variations pose a big challenge when training a common classifier for use on multiple subjects. In order to deal with these variations we introduce a novel naive Bayesian classifier with a hidden node (NB-HN).

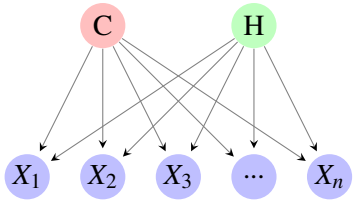

Figure 2. Hidden Node Naive Bayes Classifier

Despite a large amount of variation between subjects, it is reasonable to assume that there exists some commonalities in their brain signals in response to the task demands. By introducing a hidden node to the standard NB model we can account for both the common aspects of each subject's data as well as the individual differences. A graphical model of the NB-HN classifier is shown in Figure 2, where an additional node H is connected to each feature node. The structure of NB here models the shared attributes of different subjects and node H is used to model the inter-subject differences. The node H may represent any factor that could cause the between subject variation. The expectation-maximization (EM) algorithm is used to uncover hidden node variables (Dempster, Laird, & Rubin, 1977). No *a priori* information is needed during the training stage to compute the hidden node. Additionally, the value of the hidden node is not needed at the testing stage. The posterior probability can be computed by marginalizing over the hidden node as shown in equations 4–6.

$$p(C|X) = \frac{P(C, X)}{P(X)} \quad (4)$$

$$= \frac{\sum_H P(C, X, H)}{P(X)} \quad (5)$$

$$= \frac{\sum_H P(X|C, H)P(C)}{P(X)} \quad (6)$$

The hidden node may be a discrete node or continuous node. Typically, the larger its size is, the more infor-

mation it contains. In the present experiment we used a discrete hidden node with size of 12.

Method

The EEG data used in the present article comes from a previously published study by Wilson et al., 2010 in which eight participants (3 males; mean age 21.1 years) performed the Multi-Attribute Task Battery (MATB) (Comstock & Arnegard, 1992) on five separate sessions spread over the course of a month. The five sessions were separated by 1 day, 1 week, 3 weeks and 4 weeks. The MATB includes monitoring, communication, and resource allocation tasks which are performed concurrently in a continually changing task environment. The demands of each subtask were varied so that three levels of overall MATB difficulty were available. In an attempt to reduce learning effects, participants were trained until performance scores reached asymptote with minimal errors. Each day's session consisted of three trials where a trial was comprised of a low, medium and high difficulty block. Each block lasted five minutes and the order of blocks within each trial was random. Three of the participants did not fully complete all of the trials on day 3. For this reason, day 3 was excluded resulting in 12 complete trials for each participant.

EEG

Nineteen EEG channels were recored using the International 10-20 montage (Jasper, 1958). Mastoids were used as reference and ground with electrode impedances 5K ohms or less. The EEG data was corrected for eye movement and blinks and stored at 256 samples per second. The EEG data was then down sampled to 128 samples per second prior to analysis. Discrete-time short-term Fourier transform (STFT) was performed on the down sampled EEG data using 40 second windows with 35 seconds of overlap. No taper function was applied to the windows. The magnitude of the alpha band (9-13 Hz) from the 19 sites was used as inputs to the classifiers.

Model Training

Models were trained and tested using a fivefold cross-validation setup. One fifth of the EEG data from each trial was randomly sampled for the purpose of training the models. The data not selected for training was used for testing. Data was sampled evenly across workload blocks, and for the models including multiple subjects data, evenly across subjects. This procedure was repeated for each trial.

Table 1
Fisher's LSD t-test, P value adjustment method: bonferroni, alpha: 0.05, Df Error: 35, Critical Value of t: 2.03.

		Difference	pvalue	*sig*
NB-1	NB-8	0.339	0.0013	**
NB-1	NB-HN-8	0.107	1.0000	
NB-1	NN-1	0.127	1.0000	
NB-1	NN-8	0.342	0.0012	**
NB-HN-8	NB-8	0.232	0.0571	.
NN-1	NB-8	0.212	0.1089	
NB-8	NN-8	0.003	1.0000	
NB-HN-8	NN-1	0.020	1.0000	
NB-HN-8	NN-8	0.235	0.0522	.
NN-1	NN-8	0.215	0.0998	.

Results

The mean classification accuracy for each *model* × *training* × *workload* combination is shown in Figure 3. NN-1 and NB-1 represent the performance of the neural net and naive Bayes classifiers trained and tested on individual subjects. NN-8 and NB-8 represents the performance of the neural net and naive Bayes classifiers trained on multiple subjects and tested on individual subjects. NB-HN-8 represents the performance of the hidden node native Bayes classifier trained on multiple subjects and tested on individual subjects. See Table 1 for results of Fisher's LSD t-tests. NB-1 had a significantly higher mean classification accuracy than NN-8 and NB-8. NB-HN-8 had a marginally significant higher mean classification accuracy than NN-8 and NB-8. NN-1 had a marginally significant higher mean classification accuracy than NN-8.

Discussion

We had two goals in the present paper, explore the affects of between-subject variability on classifier accuracy, and compare the performance of artificial neural networks and Bayesian networks. In order to ensure fair comparisons, identical features were used for all models; we had no interest in feature selection or optimization. Therefore, the accuracy levels achieved do not necessarily represent the best possible performance of any of the models.

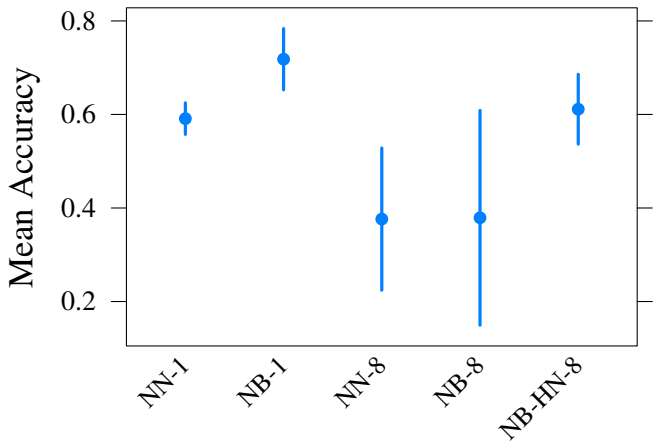

Figure 3. Mean classification accuracy from testing on each of the 8 subjects for each of the five classifiers. The error bars represent 95% confidence intervals.

Previous research has demonstrated that artificial neural networks can achieve high classification accuracy rates. In our experiments, individually trained naive Bayesian classifiers (NB-1) had a mean classification accuracy that was 12.7% higher than individually trained neural network classifiers (NN-1). Despite this difference not being significant, it is worthy to note that Bayesian classifiers can achieve comparable performance to artificial neural networks.

As expected, both the neural net and standard naive Bayes classifiers trained on multiple subjects (NN-8 and NB-8 respectively) performed worse than the individually trained classifiers of the same type (i.e., NN-8 < NN-1 and NB-8 < NB-1). Indeed, the performance of NN-8 and NB-8 were essentially no better than chance. These classifiers were not able to pick out the signal from the noise when presented with data from multiple subjects. However, when a hidden node that performed expectation-maximization was introduced to the standard naive Bayes classifier (NB-HN-8), its performance increased to that of a individually trained neural net classifier.

Conclusion

In this paper, we demonstrated that a classifier trained on multiple subjects can achieve performance comparable to classifiers trained on multiple subjects. This was accomplished by adding a hidden node to a naive Bayes classifier. The hidden node in this case used the expectation-maximization algorithm to account for between subject variations. These results take EEG classification one step closer to being able to discriminate workload levels on subjects that the classifier was not trained on.

References

Bashashati, A., Fatourechi, M., Ward, R. K., & Birch, G. E. (2007). A survey of signal processing algorithms in brain-computer interfaces based on electrical brain signals. *Journal of Neural Engineering*.

Comstock, J., J. Raymond, & Arnegard, R. J. (1992). *The multi-attribute task battery for human operator workload and strategic behavior research* (NASA Technical Memorandum No. 104174). Langley Research Center.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, *39*, 1-38.

Gevins, A., & Smith, M. E. (2003). Neurophysiological measures of cognitive workload during human–computer interaction. *Theoretical Issues in Ergonomics Science*, *4*, 113-131.

Gevins, A., Smith, M. E., Leong, H., McEvoy, L., Whitfield, S., Du, R., et al. (1998). Monitoring working memory load during computer-based tasks with eeg pattern recognition methods. *Human Factors*, *40*, 79-91.

Hart, S. G., & Staveland, L. E. (1988). Development of nasa-tlx (task load index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (p. 139183). Amsterdam: North-Holland.

Inagaki, T. (2003). Adaptive automation: Sharing and trading control. In E. Hollnagel (Ed.), *Handbook of cognitive task design* (p. 147-169). Lawarence Erlbaum Associates, Inc.

Jasper, H. H. (1958). Report of the committee on methods of clinical examination. *Electroencephalography and clinical Neurophysiology*, *10*, 370-375.

Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., & Arnaldi, B. (2007, June). A review of classification algorithms for eeg-based brain-computer interfaces. *Journal of Neural Engineering*, *4*(2).

Lysaght, R. J., Hill, S. G., Dick, A., Plamondon, B. D., Linton, P. M., Wierwille, W. W., et al. (1989). *Operator workload: Comprehensive review and evaluation of operator workload methodologies* (AFI Tech-

nical Report No. 851). U.S. Army Research Institute.

Moray, N., Inagaki, T., & Itoh, M. (2000). Adaptive automation, trust, and self-confidence in fault management of time-critical tasks. *Journal of Experimental Psychology: Applied*, *6*, 44-58.

Noyes, J. M., & Bruneau, D. P. J. (2007). A self-analysis of the nasa-tlx workload measure. *Ergonomics*, *50*, 514-519.

Rubio, S., Diaz, E., Martin, J., & Puente, J. M. (2004). Evaluation of subjective mental workload: A comparison of swat, nasa-tlx, and workload profile methods. *Applied Psychology An International Review*, *53*, 61-86.

Wilson, G. F., Estepp, J., & Christensen, J. C. (2010). How does day-to-day variability in psychophysiological data affect classifier accuracy. *Human Factors and Ergonomics Society Annual Meeting Proceedings*, *54*, 264-268.

Wilson, G. F., Estepp, J., & Davis, I. (2009). A comparison of performance and psychophysiological classification of complex task performance. *Human Factors and Ergonomics Society Annual Meeting Proceedings*, *53*, 141-145.

Wilson, G. F., & Fisher, F. (1995). Cognitive task classification based on topographic eeg data. *Biological Psychology*, *40*.

Wilson, G. F., & Russel, C. A. (2003a). Operator functional state classification using psychophysiological features in an air traffic control task. *Human Factors*, *45*, 381-389.

Wilson, G. F., & Russel, C. A. (2003b). Real-time assessment of mental workload using psychophysiological measures and artificial neural networks. *Human Factors*, *45*, 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*, 1005-1018.

Zhang, H. (2004). The optimality of naive bayes. In *The international flairs conference* (Vol. 17).