Session-independent EEG-based Workload Recognition

Felix Putze, Markus Müller, Dominic Heger and Tanja Schultz
Institute of Anthropomatics, Cognitive Systems Lab, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

Keywords: EEG, Workload Recognition, Session Independence, Adaptation, Feature Selection.

Abstract: In this paper, we investigate the development of a session-independent EEG-based workload recognition system with minimal calibration time. On a corpus of ten sessions with the same subject, we investigate three different approaches: Accumulation of training data, an adaptive classifier (adaptive LDA) and feature selection algorithm (based on Mutual Information) to improve generalizability of the classifier. In a detailed evalution, we investigate how each approach performs under different conditions and show how we can use those methods to improve classification accuracy by more than 22% and make transfer of models between sessions more reliable.

1 INTRODUCTION

It is known that mental overload has a negative impact on driving performance and therefore seriously affects safety in every day traffic (Lansdown et al., 2004). Much of this load is induced by in-vehicle systems designed to facilitate, support and entertain the user while driving. Those systems could benefit from a greater insight in the driver's workload level by reacting appropriately to his inner state. Previous studies have shown that a user's workload level can be reliable recognized using electroencephalography (EEG) signals (see for example (Heger et al., 2010), (Jarvis et al., 2011), (Kothe and Makeig, 2011)). Still, achieving session-independence for EEG-based workload recognition systems is a challenge as a number of parameter may change from session to session: the exact positioning of electrodes, the physical condition of the user, environmental factors influencing the recording, etc. The large number of influencing variables makes it very difficult to design a calibration or normalization scheme. Goal of this work was to provide session-independent workload recognition capable of online recognition with minimal preparation time for the user. This implies that no global normalization schemes are allowed or techniques which require the recording of much labeled or even unlabeled training data.

In literature on active BCIs and biosignal processing, a number of methods for achieving session-independence have been proposed: (Shenoy et al., 2006) investigate how classifiers which are trained on offline calibration data perform in online conditions. They study the distributions of the features and the resulting classification models, notice systematic differences between both conditions and show that very different sessions in training and testing data may result in degraded recognition performance. They also note that "the strongest source of non-stationarity stems from the difference between calibration and feedback sessions", which leads us to concentrate on differences between sessions compared to negligible non-stationarities within a session of a few minutes length. Vidaurre et al. (Vidaurre et al., 2008) present an approach for unsupervised adaptation of an LDA classifier based on the assumption that non-stationarities influence the statistics of all classes in the same way. Their analysis on two BCI datasets indicates that there is a small advantage for supervised adaptation but also note "that adapting means with and without class-labels was not found significantly different".

2 METHODOLOGY

To collect data for training and evaluation of a session-independent workload classifier, we designed and conducted an experiment. One subject (a male student) recorded ten sessions over the course of several months. During each session, he performed a main task of operating a simple driving simulator (Mattes, 2003) and several different secondary tasks in parallel. The session was broken down in stages; for each stage, the type of secondary task

Putze F., Mülller M., Heger D. and Schultz T..
Session-independent EEG-based Workload Recognition.
DOI: 10.5220/0004250703600363
In *Proceedings of the International Conference on Bio-inspired Systems and Signal Processing* (BIOSIGNALS-2013), pages 360-363
ISBN: 978-989-8565-36-5

(if any) was kept constant. Driving stages with secondary task were labeled as high workload conditions while driving stages without secondary task were labeled as low workload condition. There are three different types of secondary tasks: A visual search task in two different difficulty levels (*Visual1* and *Visual2*), a math task with two difficulty levels (*Divide1* and *Divide2*) and a game of Tic Tac Toe against the computer (*TTT*). All secondary tasks were presented on a monitor to the right of the subject and operated by keyboard within easy reach. Each condition (driving only and driving with each of the secondary tasks) was recorded twice for six minutes each. The order of tasks was randomized between sessions to eliminate order effects. During each task, EEG was recorded using an Emotiv EPOC device. This wireless device offers a fixed layout of 14 saline electrodes sampled at 128Hz. It can be fully set up in less than two minutes by the user without help, which constitutes a benefit for our aim of preparation-free workload recognition compared to classic EEG caps. The user was told to concentrate on the task but was not instructed otherwise (e.g. on artifact avoidance) to record data under realistic conditions. In total, we collected 10 sessions with 60 minutes of EEG data each, resulting in a total corpus of 600 minutes of usable data.

The baseline system for session-dependent workload recognition is described in (Heger et al., 2010). From each window of 2 seconds, it extracts 28 spectral features in the range from 4 to 45 Hz for each electrode. The window is shifted with an overlap of 1.5s over the data stream, resulting in one data-point for each 0.5s. Before the spectral feature extraction, we perform an automatic removal of eyeblink artifacts based on Independent Component Analysis as described in (Jarvis et al., 2011) and a Canonical Correlation Analysis (Clercq et al., 2006) to remove EMG artifacts. Two classes of low and high workload are discriminated by a binary classifier based on Linear Discriminant Analysis (LDA). Results are smoothed over 3 consecutive data-points to get a more reliable workload estimate.

To achieve session independence for this baseline system, we follow two main approaches:

Session Adaptation: One way to handle differences between trained models and testing data is to actively adapt the classification model to the conditions of the current session. (Vidaurre et al., 2008) propose an unsupervised adaptation of joint statistics for both classes. The update of the selected method modifies the joint class mean $\mu(t)$ for a newly calculated feature vector $x(t)$ as follows:

$$\mu(t) = (1 - UC) \cdot \mu(t-1) + UC \cdot x(t) \qquad (1)$$

The joint mean is used to correct the bias in the feature distribution of the testing session. In formula 1, *UC* is the *update coefficient* that determines the strength of the update. Tuning the update coefficient to a correct level is a crucial aspect of this method. The approach in (Vidaurre et al., 2008) was designed to account for non-stationarities within one session and therefore uses a continuous update for the whole data stream. This seems non-optimal for adaptation between training sessions and testing sessions (which we assume to be stable due to their length of only a few minutes) for several reasons: First, a user expects a working system after a calibration phase of minimal duration. An update coefficient which is optimized to adapt the model to slow changes in the signal characteristics may result in too timid updates for inter-session adaptation. Second, when the optimal UC is estimated and evaluated on sessions of a fixed length it may be a suboptimal choice for sessions of very different duration. Therefore, we only perform adaptation on the first feature vectors of a session and keep the model constant after that. We call the number of features used for adaptation *adaptation count* (AC).

Robust Feature Accumulation: The quality of session-independent recognition highly depends on the quality and variety of the available training data. A large training set can cover a wide range of possible feature distributions and account for variability in the test set. Therefore, we can expect a more reliable recognition with multiple training sessions than with a limited training set. Of course, acquiring such a training set for each user is opposed to the goal of minimizing the effort of data collection, i.e. we have to do a cost-benefit analysis of the addition of new training sessions and also have to find ways to extract reliable models already from smaller training sets. Each recorded stage in a session is 6 minutes long, resulting in 1,440 training samples per session for training a quadratic covariance matrix of 392 dimensions (14 channels with 28 features each), resulting in more than 150,000 coefficients. This mismatch may result in overfitted models which are tuned towards the specific conditions of the training data but which do not generalize to other sessions. To mitigate this problem, we employ feature selection which tries to identify the most relevant features for a classification task. We employ a wrapper approach based on Mutual Information (MI) as described by (Ang et al., 2008). They describe the *Mutual Information based Best Individual Feature (MIBIF)* algorithm, a feature selection approach based on a high relevance criterion to reduce the feature space dimensionality. It selects the

K features with the highest Mutual Information with the ground truth. The selection of the feature count *K* is of course critical for the performance. We will investigate whether the optimal *K* is dependent on the number of available training sessions or whether there is a globally optimal *K* for the presented setting.

3 EVALUATION

For evaluation, we extract and concatenate all stages without secondary task of each session as `low` workload condition and extract and concatenate all data for one fixed secondary task of each session as `high` workload condition. Baseline performance for an all-pair evaluation (i.e. each session is used as training session for a model which is evaluated on all other sessions) averaged over all tasks is 64.9%. However, with a minimum accuracy of 49.7% and a maximum accuracy of 80.9% there is considerable variation within the results. This indicates that there is a mismatch between some pairs of training and testing session which prevents session-independent recognition in the baseline setup. Mitigating the effect of those mismatches is the main challenge of session-independent workload recognition as it makes results of each particular testing session unpredictable.

We first evaluate the effect of adaptation. For this purpose, we do an all-pair evaluation on seven out of the ten sessions to determine optimal values for the adaptation coefficient UC. We do this analysis separatedly for all tasks to study potential differences. Figure 1 shows the estimated optimal values for UC for different sizes AC of the adaptation window. As expected, we see a linear dependency between both values (both scales are logarithmic). While all tasks share the same trend for UC, there are considerable differences between the optimal values and we therefore continue analysis with task-specific values for UC. Some outliers, e.g. for the *Divide2* task and $AC = 32$, also indicate that estimating the free parameters of the adaptation is a delicate process sensible to the distribution of training data.

To investigate the benefit of using the estimated UC on unseen data, we perform again an all-pair evaluation on the three sessions that were held out with UC fixed to the previously determined value. For $AC = 64$, averaged over all tasks, we achieve a relative improvement in recognition accuracy of 8%. Recognition accuracy does not improve in all cases: For 26% of all instances of the cross-validation, performance degrades slightly by 3.5% relative on average. This may be the case due to unrepresentative data within the adaptation window or due to a violation of the assumption that both classes differ similarly between training and test session. The main benefit of adaptation is not the overall improvement of recognition accuracy but the mitigation of extreme mismatches between training and testing data. The minimum recognition accuracy across all pairs in the hold-out set increases from 51% to 63% when activating adaptation and the standard deviation is reduced from 8.4% to 5.6%. The size *AC* of the adaptation window does not have significant impact on recognition accuracy. While performance improves monotonically with higher *AC*, it only increases by 3% relative when going from $AC = 4$ to $AC = 512$.

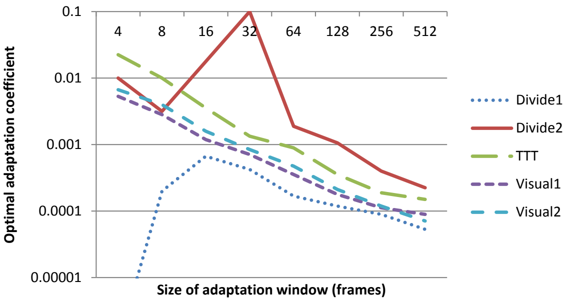

Figure 1: Optimal update coefficients for different sizes of the update window calculated for all secondary tasks.

To quantify the effect of additional training material, we performed leave-one-session-out cross-validation. In each iteration, we fixed one session as testing session and trained the classification model repeatedly on a growing training set which was generated by iteratively adding sessions in chronological order. This analysis was repeated for all secondary tasks. Figure 2 shows the recognition accuracy averaged over all tasks for different sizes of the training set. We see that overall, adding more sessions increases accuracy by more than 22% relative. The graph also indicates that accuracy may also not be saturated with a training set of nine sessions, i.e. adding more material may further increase the performance. In more than 89% of all instances in which a session was added to a training set this actually increased the resulting recognition accuracy on the fixed testing session. An analysis of the cases in which performance degrades shows that those instances correspond to pairs of sessions which also already perform with below-average accuracy in the baseline evaluation (one potentially problematic session alone accounts for more than 30% of those instances).

Contributing to the pronounced performance increase due to a larger number of training sessions may be caused by the fact that more training data allows a more robust estimation of a classification model of large dimensionality. A reduction of the fea-

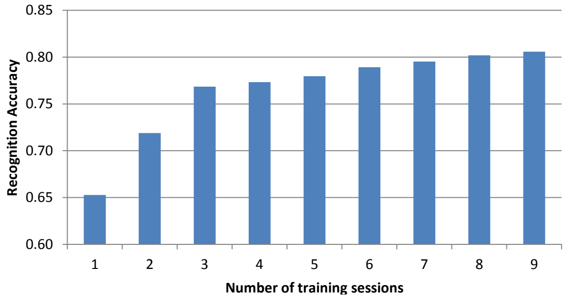

Figure 2: Average recognition performance depending on the number of available training sessions.

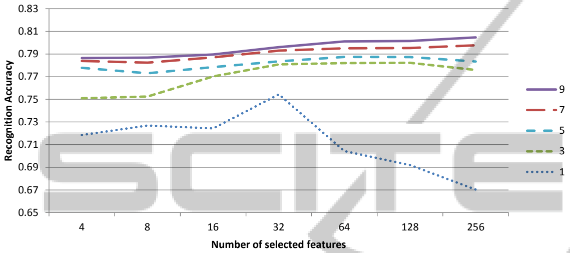

Figure 3: Average classification accuracy in dependency of numbers of selected features for different numbers of training sessions.

ture space would potentially help models trained from fewer training instances to perform better in comparison to models trained with more data. We therefore estimate classification performance when applying feature selection as described in section 2 for values of $K = 4, 8, 16, \ldots, 256$. Figure 3 shows that for a smaller number of available training sessions a smaller number of selected features yields optimal performance while the large training corpora can only be optimally exploited if more features remain. However, there is an effect of diminishing returns as the average difference between the best recognition accuracy and the one achieved with $K = 32$ is below 1%.

Figure 4 presents recognition accuracy for different sizes of the training set using the individual optimal values for K. It shows that employing feature selection indeed decreases the difference in performance between small and large training sets, while the general tendency of additional training data contributing to higher accuracy remains stable.

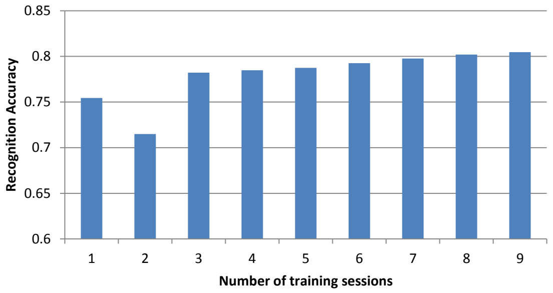

Figure 4: Average recognition performance depending on the number of available training sessions when using feature selection.

To summarize, we saw that unsupervised adaptation improved recognition accuracy and helped to make the recognition more predictable by improving accuracy especially in cases of mismatched training and testing sessions. If additional training material can be provided, accumulation of training data combined with feature selection can improve recognition accuracy substantially.

REFERENCES

Ang, K. K., Chin, Z. Y., Zhang, H., and Guan, C. (2008). Filter bank common spatial pattern (FBCSP) in brain-computer interface. In *IEEE International Joint Conference on Neural Networks. IJCNN*, pages 2390–2397. IEEE.

Clercq, W. D., Vergult, A., Vanrumste, B., Van Paesschen, W., and Van Huffel, S. (2006). Canonical correlation analysis applied to remove muscle artifacts from the electroencephalogram. *IEEE Transactions on Biomedical Engineering*, 53(12):2583 –2587.

Heger, D., Putze, F., and Schultz, T. (2010). An adaptive information system for an empathic robot using EEG data. In Ge, S., Li, H., Cabibihan, J.-J., and Tan, Y., editors, *Social Robotics*, volume 6414 of *Lecture Notes in Computer Science*, pages 151–160. Springer Berlin / Heidelberg.

Jarvis, J., Putze, F., Heger, D., and Schultz, T. (2011). Multimodal person independent recognition of workload related biosignal patterns. page 205. ACM Press.

Kothe, C. and Makeig, S. (2011). Estimation of task workload from EEG data: New and current tools and perspectives. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society,EMBC*, pages 6547 –6551.

Lansdown, T., Brook-Carter, N., and Kersloot, T. (2004). Distraction from multiple in-vehicle secondary tasks: vehicle performance and mental workload implications. *Ergonomics*, 47(1):91–104.

Mattes, S. (2003). The lane-change-task as a tool for driver distraction evaluation. In *Proceedings of IGfA*.

Shenoy, P., Krauledat, M., Blankertz, B., Rao, R. P. N., and Mller, K.-R. (2006). Towards adaptive classification for BCI. *Journal of Neural Engineering*, 3(1):R13–R23.

Vidaurre, C., Schloegl, A., Blankertz, B., Kawanabe, M., and Mller, K.-R. (2008). Unsupervised adaptation of the lda classifier for brain-computer interfaces. In *Proceedings of the 4th International Brain-Computer Interface Workshop and Training Course*, pages 122–127.