Model-Based Identification of EEG Markers for Learning Opportunities in an Associative Learning Task with Delayed Feedback

Felix Putze[1,⋆], Daniel V. Holt[2], Tanja Schultz[1], and Joachim Funke[2]

[1] Karlsruhe Institute of Technology, Institute of Anthropomatics and Robotics, Karlsruhe, Germany
[2] University of Heidelberg, Institute of Psychology, Heidelberg, Germany
{felix.putze,tanja.schultz}@kit.edu,
{daniel.holt,funke}@psychologie.uni-heidelberg.de

Abstract. This paper combines a reinforcement learning (RL) model and EEG data analysis to identify learning situations in a associative learning task with delayed feedback. We investigated neural correlates in occipital alpha and prefrontal theta band power of learning opportunities, identified by the RL model. We show that those parameters can also be used to differentiate between learning opportunities which lead to correct learning and those which do not. Finally, we show that learning situations can also be identified on a single trial basis.

Keywords: Reinforcement Learning, learning situations, EEG, Frequency Analysis.

1 Introduction

Reinforcement learning (RL) is a fundamental mechanism of adaptive behavior in humans. It is often implicitly involved in Human-Computer Interaction (e.g. when users learn to operate a new software) but can also be explicitly employed as part of a predictive user model for adaptive systems. The underlying models of the learning progress are usually individually calibrated through behavioral data (e.g., response probabilities). In recent years biosignals generated by neural activity (as measured by EEG or fMRI methods) have become another relevant source of information for real-time user modeling. The practical utility of this combined approach was illustrated by [1], who showed how the prediction of mental user states in an intelligent tutoring system for an algebra-isomorph can be substantially improved by blending predictions of a cognitive task model with neurally derived information. However, in order to successfully apply this approach, neural markers need to be identified that can be integrated into user

⋆ This project was partially funded by the Heidelberg Karlsruhe Research Partnership (HEiKA), a co-operation between the Ruprecht-Karls-University of Heidelberg and the Karlsruhe Institute of Technology.

S. Wermter et al. (Eds.): ICANN 2014, LNCS 8681, pp. 387–394, 2014.

models in a principled manner. In this paper, we employ a simple reinforcement learning model to establish EEG markers for learning opportunities in an associative learning task with delayed feedback.

2 Related Work

In RL organisms learn to select sequences of actions that maximize their subjective reward over time based on the reward signals (feedback) associated with different outcomes. This can be achieved through temporal difference learning (TD), which assigns credit based on the temporal proximity of actions to outcomes. The authors of [4] demonstrated how a TD-based RL model can predict learning performance by TD-based reward propagation in a complex associative learning task with delayed feedback. One neurophysiological approach for studying RL is to analyze the Feedback Related Negativity (FRN). The FRN is a frontocentral neural response appearing 200-300ms after the presentation of feedback indicating prediction errors (i.e., a mismatch between mental model and observation). [15] documents that prediction error can be used in a task with delayed feedback to predict the occurrence of FRN for task states immediately followed by feedback as well as intermediate states. The authors present this effect as evidence for credit assignment to intermediate states from future rewards. [2] moves from time domain analysis to frequency analysis and links prefrontal theta synchronization to adaption effects in a probabilistic reinforcement learning task. A Q-Learning model was used to estimate prediction errors, which indicated whether a situation reflects a learning opportunity. While the work mentioned above explicitly addresses the processing of prediction errors, there are other cognitive processes and corresponding neurological markers related to learning events, for example working memory activity [3]. Early work on the relation of EEG synchronization/de-synchronization and memory processes has identified theta synchronization and alpha desynchronization during supposed memory processes [7,5,16,10]. Regarding alpha oscillations, following research has also identified "paradoxial" alpha synchronization during cognitive activity, which in subsequent work [6,11,9,14] was reinterpreted as a possible inhibition of task irrelevant cortical processes or conscious inhibition of cognitive processes impeding the task.

In this paper, we establish neurological markers of learning opportunities in a complex associative learning task, particularly considering memory encoding and feedback processing. We selected a complex learning task where a sequence of interdependent decisions is required to achieve a desired outcome. Tasks of this type that do not not involve probabilistic outcomes have so far not been considered in EEG studies of RL. However, learning such action sequences is both common and important in human-computer interaction, for example when trying to achieve a particular result with an unfamiliar software.

3 Methods

The behavioral task employed is an modified version of the task used in [4]. Formally, it is an abstract tree-search which requires three binary decisions to move from the root node to a leaf node. Feedback about the success of a decision sequence is provided when reaching a leaf node. When reaching a non-target leaf node (failure), participants are moved back to the last node where they were still on path to the target. When reaching the target leaf node node (success) one learning trial is complete and the participant is returned to the root node for the next trial. Semantically, the task is framed as a "strange machine", which has four buttons (red, yellow, green, blue) and a display showing its current state in a "unknown language" (a pronounceable German non-word such as "Tarfe"). See Figure 1 for a summary of the internal structure and the display of a node. In each state two of the buttons are active to move the machine into the next state. After three button presses, the machine either reaches the target state or a failure state and is reset as described above. The task goal is to learn to reach the target state as consistently as possible without failures. To increase learning load, each state node has three possible labels associated with different response options. At each visit of a node one of these sets is randomly selected and displayed to the participant.

The procedure consisted of brief instructions followed by 15 practice trials and a main learning phase with 100 trials[1]. If participants completed the main learning phase in less than 45 minutes, a second learning phase with a differently labeled version of the machine was conducted.

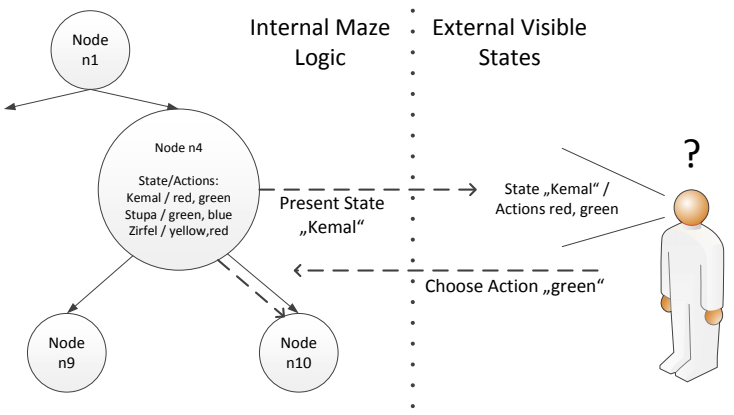

Fig. 1. Internal structure and external view of the "strange machine" task

[1] For the first 8 participants the main learning phase lasted 120 or 160 trials, which due to ceiling effects was subsequently reduced to 100.

Data were collected from 34 university students (23 female, mean age 23.1 years). Participants gave written consent and were paid for their participation. 18 participants completed two machines, 16 completed only one.

EEG was recorded from 29 scalp electrodes placed according to the international 10-20 system using actiCAP active electrodes and actiCHamp amplifiers (Brain Products, Germany) at a sampling rate of 500 Hz with Cz as recording reference. The EEG data were re-referenced to a common average reference and segmented into windows of 400ms length starting 100ms after a new state is displayed. Data segments containing ocular artifacts were identified and removed by testing for correlation of electrodes Fp1 and Fp2 above a threshold of 0.97 within the regarded time frame. This procedure rejects approximately 4.5% of all trials. This means each window contains data from processing the feedback (either a new state of direct feedback at a leaf node) following a decision step. Each window was normalized by subtracting the mean from 250-150ms before stimulus. For band power analysis, we used the Thomson's multitaper power spectral density (PSD) estimate [13]. The relevant (sub-)bands for analysis were estimated on an individual basis following the method of [8]. The averaged PSD was then z-normalized for each subject.

Similar to [4], we used a Reinforcement Learning approach to model human learning behavior. We employed the Naive Q-Learning (NQL) algorithm, a variant of Watkin's $Q(\lambda)$ [12] to model the participants' learning progress. NQL is a Temporal Difference (TD) method with eligibility traces. The work of [15] demonstrates that TD methods are capable of reproducing human learning behavior and predict the generation of propagated FRNs. This work also demonstrated the benefit of eligibility traces for the purpose of closely fitting human behavioral data. Reward was selected to be $+7$ for the target node, -1 for the dead-end nodes and 0 for any inner nodes. Temperature and λ were fixed at 1.0 and 0.1, respectively. Learning rate α was optimized between 0.02 and 0.3 for each subject individually to account for the large inter-subject variance in performance. Each state label (not the node itself) is a state of the RL model, with two possible actions corresponding to the buttons of that label. For each session, a new model was initialized and trained using the action sequence as denoted in the corresponding maze log file. This allowed us to trace the learning from observation in each individual session. To quantify learning opportunities, we define *uncertainty* as the entropy of the Softmax probability distribution [12] resulting from the action Q-scores for a specific state. Until any non-zero feedback has been propagated to a state, this will result in a maximum uncertainty value of $\log 2$. When a state accumulates propagated rewards, uncertainty converges towards zero. As we can use this definition only for correct states, we define certain incorrect nodes to have a negative Q-score $< -\epsilon$ for both actions. The benefit of the notion of uncertainty compared to the classic notion of prediction error - which is defined as the update delta of the Q-score of the outgoing state for a certain step (see for example [15]) - is that it is defined in terms of states and not in terms of steps. Therefore, it can help a tutoring system to identify states which are not yet sufficiently well learned.

4 Analysis

We now investigate the relation between the prediction of computational RL model and empirical EEG data to identify situations in which learning occurs. We do this in two main steps: First, we use the RL model to predict learning opportunities and look at neurological correlates in the EEG data. Second, we differentiate learning opportunities between successful and unsuccessful learning attempts. This second step shows how EEG markers and computational model interact to identify learning situations better than each of them can individually.

For the analysis of EEG synchronization and desynchronization, we concentrate on two effects that are related to feeback processing and memory encoding: Theta synchronization in the prefrontal cortex and alpha synchronization in the occipital cortex. We average PSD across electrodes O1 and O2 to represent occipital activity and average PSD across electrode positions Fz, Fc1, Fc2 to represent prefrontal activity.

We assume that memory encoding occurs systematically when new information on the task is learned from the feedback at the end of certain steps. We therefore have to identify those situations which allow learning. To sort the steps into classes, we use the RL model and apply two thresholds to dichotomize uncertainty: A strict threshold t_s (selected to characterize 80% of all values as 'high uncertainty') and a tolerant t_t threshold (selected to characterize 30% of all values as 'high uncertainty'). We use t_s to label outgoing states as (un)certain and t_t to label incoming states. This choice minimizes the number of missed learning opportunities. The left half of Figure 2 summarizes the class definition: Class `LEARN` denotes a learning opportunity, class `NO-INFO` denotes absence of a learning opportunity due to missing information and class `SATURATED` denotes absence of a learning opportunity due to an already saturated knowledge. We expect to see pronounced differences between the first and the latter two classes. We expect the latter two classes to be similar. To avoid class imbalance, we only include the first five occurrences of each state in each class in our analysis. Statistics are calculated on the normalized averaged PSD distributions for the respective classes as a two-sided paired t-test. To rule out that low-frequency ocular artifacts confound the results, we checked that there was no significant difference in eye blink frequency between the different classes during preprocessing.

Figure 3 shows average occipital alpha power and average prefrontal theta power calculated for the three classes separately. We see a increase in alpha power from the `NO-INFO` class to the `LEARN` class in the occipital cortex, while there is no significant difference between `NO-INFO` and `SATURATED`. Analogously, we see a difference between `NO-INFO` class to the `LEARN` and `SATURATED` classes in the theta band for the prefrontal cortex. However, those differences in the regarded bands marginally miss statistical significance: $t(36) = 1.48$, $p = 0.07$ for occipital alpha and $t(36) = 1.62$, $p = 0.057$ for prefrontal theta. One reason for this observation is that learning opportunities denote the potential for learning, but do not always lead to memory encoding as the subject overlooks the opportunity or is not able to correctly memorize the new information.

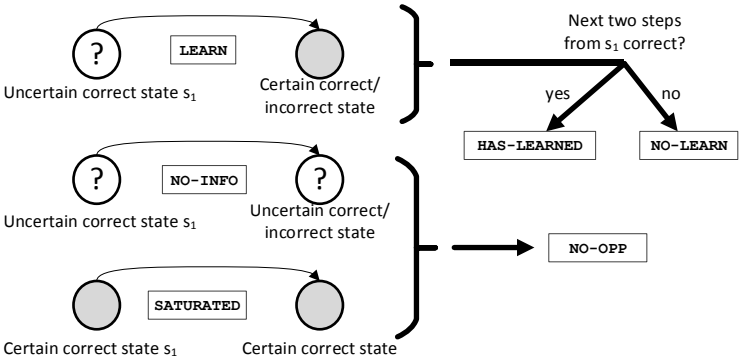

Fig. 2. Definition of learning opportunities (left) and learning situations (right) as derived from the RL model to form the classes for evaluation of classes

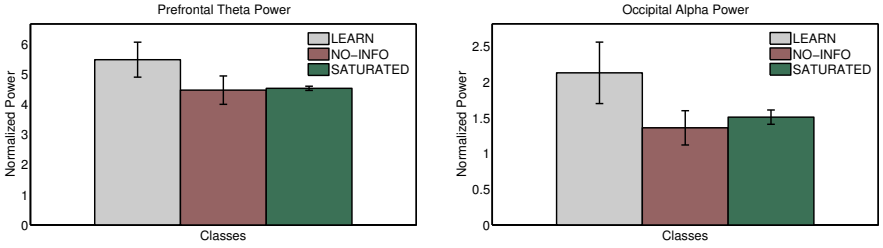

Fig. 3. Theta power at the prefrontal cortex (left) and alpha power at the occipital cortex (right) for the classes `LEARN`, `SATURATED` and `NO-INFO` for learning opportunities. Whiskers indicate standard error.

The criteria we defined in the RL model yield a reasonable prediction whether a learning situation occurs during a specific step. In the previous analysis, we assumed the definition of a learning situation as a given ground truth to investigate neurological markers for learning. However, we concluded that the computational model can only yield a noisy prediction of a successfully learning event. To quantify this predictive power, we introduce the term of a *learned state*. A learned state is a correct state s for which the next two steps starting in s stay on the correct path. 38% of all steps labeled as learning situations do not result in a learned state[2]. In the following, we combine this prediction by the computational RL model with the information of EEG to detect those missed learning opportunities. We propose that the observed alpha and theta synchronization effects are caused by cognitive processes of learning situations. This implies that when sorting learning opportunities in learned and not-learned outgoing states, we should observe a similar difference in PSD: Learned outgoing states show a level of alpha and theta synchronization which is not present for missed learning

[2] This number depends of course on the threshold applied to the uncertainty level of the outgoing step. A lower threshold leads to fewer false alarms but also increases the number of missed learning opportunities.

opportunities. To investigate this hypothesis, we sort the steps from the LEARN class of the positive and negative learning opportunities by this criterion, forming the HAS-LEARNED and the NOT-LEARNED classes. Steps which are not categorized as learning opportunities form the NO-OPP class, see the right half of Figure 2. On average, the LEARN class contains 26.1 steps, while the LEARN class contains 16.6 steps. Figure 4 shows the band power for the three different classes, now resulting in a significant ($t(35) = 2.74$, $p < 0.005$) increase in individual alpha power from the non-learned to the learned steps, as well as a significant difference in theta power ($t(35) = 1.76$, $p < 0.05$) in the prefrontal cortex. The steps in the NOT-LEARNED class are not significantly different from steps in NO-OPP for both brain regions.

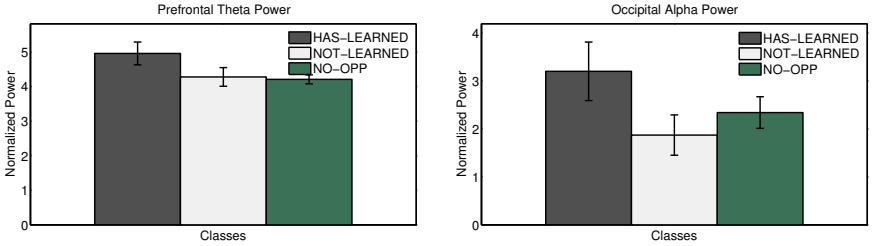

Fig. 4. Theta power at the prefrontal cortex (left) and alpha power at the occipital cortex (right) for the classes HAS-LEARNED, NOT-LEARNED and NO-OPP for learning situations. Whiskers indicate standard error.

To make this significant difference accessible for a tutoring system, we need to provide prediction of learning situations on a single trial basis. For this purpose, we train a Naive Bayes classifier to separate the HAS-LEARNED and the NOT-LEARNED class. As features, we use individual occipital alpha power and prefrontal theta power. We evaluate this classifier in a participant-dependent leave-one-out crossvalidation. To exclude cases where one class receives too few training samples, we remove the most imbalanced sessions where the majority class contains more than 70% of all samples from the analysis. The resulting classifier yields an average recognition accuracy of 71.0% which is significantly better ($t(25) = 2.49$, $p = 0.01$) than the baseline accuracy of 59.6%, as determined by a one-sided paired t-test of classification accuracy vs. size of majority class for each subject. The average improvement over the baseline is 19.7% relative.

To conclude, our results show that we can use the RL model to identify learning opportunities in an associative learning task, despite delayed feedback. We showed this by providing neural evidence for learning. We further showed that we can combine the model with such EEG markers to predict learning success. This is also feasible on a single trial basis. Future work will concentrate on reducing label noise by using a more sophisticated cognitive model (e.g. explicitly representing working memory) implemented in a cognitive architecture.

References

1. Anderson, J.R., Betts, S., Ferris, J.L., Fincham, J.M.: Neural imaging to track mental states while using an intelligent tutoring system. Proceedings of the National Academy of Sciences 107(15), 7018–7023 (2010)
2. Cavanagh, J.F., Frank, M.J., Klein, T.J., Allen, J.J.B.: Frontal theta links prediction errors to behavioral adaptation in reinforcement learning. NeuroImage 49(4), 3198–3209 (2010)
3. Collins, A.G.E., Frank, M.J.: How much of reinforcement learning is working memory, not reinforcement learning? a behavioral, computational, and neurogenetic analysis. European Journal of Neuroscience 35(7), 1024–1035 (2012)
4. Fu, W.-T., Anderson, J.R.: From recurrent choice to skill learning: A reinforcement-learning model. Journal of Experimental Psychology: General 135(2), 184–206 (2006)
5. Jensen, O., Tesche, C.D.: Frontal theta activity in humans increases with memory load in a working memory task. The European Journal of Neuroscience 15(8), 1395–1399 (2002)
6. Klimesch, W., Doppelmayr, M., Schwaiger, J., Auinger, P., Winkler, T.: 'Paradoxical' alpha synchronization in a memory task. Cognitive Brain Research 7(4), 493–501 (1999)
7. Klimesch, W.: Memory processes, brain oscillations and EEG synchronization. International Journal of Psychophysiology 24(1-2), 61–100 (1996)
8. Klimesch, W.: EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. Brain Research Reviews 29(2-3), 169–195 (1999)
9. Klimesch, W., Sauseng, P., Hanslmayr, S.: EEG alpha oscillations: The inhibition–timing hypothesis. Brain Research Reviews 53(1), 63–88 (2007)
10. Osipova, D., Takashima, A., Oostenveld, R., Fernández, G., Maris, E., Jensen, O.: Theta and gamma oscillations predict encoding and retrieval of declarative memory. The Journal of Neuroscience 26(28), 7523–7531 (2006); PMID: 16837600
11. Sauseng, P., Klimesch, W., Doppelmayr, M., Pecherstorfer, T., Freunberger, R., Hanslmayr, S.: EEG alpha synchronization and functional coupling during top-down processing in a working memory task. Human Brain Mapping 26(2), 148–155 (2005)
12. Sutton, R.S., Barto, A.G.: Introduction to Reinforcement Learning, 1st edn. MIT Press, Cambridge (1998)
13. Thomson, D.J.: Spectrum estimation and harmonic analysis. Proceedings of the IEEE 70(9), 1055–1096 (1982)
14. Tuladhar, A.M., ter Huurne, N., Schoffelen, J.-M., Maris, E., Oostenveld, R., Jensen, O.: Parieto-occipital sources account for the increase in alpha activity with working memory load. Human Brain Mapping 28(8), 785–792 (2007)
15. Walsh, M.M., Anderson, J.R.: Learning from delayed feedback: neural responses in temporal credit assignment. Cognitive, Affective, & Behavioral Neuroscience 11(2), 131–143 (2011)
16. Weiss, S., Müller, H.M., Rappelsberger, P.: Theta synchronization predicts efficient memory encoding of concrete and abstract nouns. Neuroreport 11(11), 2357–2361 (2000)

Stefan Wermter Cornelius Weber
Wlodzislaw Duch Timo Honkela
Petia Koprinkova-Hristova Sven Magg
Günther Palm Alessandro E.P. Villa (Eds.)

Artificial Neural Networks and Machine Learning – ICANN 2014

24th International Conference
on Artificial Neural Networks
Hamburg, Germany, September 15-19, 2014
Proceedings

Springer

Volume Editors

Stefan Wermter
Cornelius Weber
Sven Magg
University of Hamburg, Hamburg, Germany
E-mail: {wermter, weber, magg}@informatik.uni-hamburg.de

Wlodzislaw Duch
Nicolaus Copernicus University, Torun, Poland
E-mail: wduch@is.umk.pl

Timo Honkela
University of Helsinki, Helsinki, Finland
E-mail: timo.honkela@helsinki.fi

Petia Koprinkova-Hristova
Bulgarian Academy of Sciences, Sofia, Bulgaria
E-mail: pkoprinkova@bas.bg

Günther Palm
University of Ulm, Ulm, Germany
E-mail: guenther.palm@uni-ulm.de

Alessandro E.P. Villa
University of Lausanne, Lausanne, Switzerland
E-mail: alessandro.villa@unil.ch

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-319-11178-0 e-ISBN 978-3-319-11179-7
DOI 10.1007/978-3-319-11179-7
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014947446

LNCS Sublibrary: SL 1 – Theoretical Computer Science and General Issues

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)