Journal of Neuroscience Methods 264 (2016) 33–39

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

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

Multiple classifier systems for automatic sleep scoring in mice

Vance Gao*, Fred Turek, Martha Vitaterna

Center for Sleep and Circadian Biology, Northwestern University, Department of Neurobiology, 2205 Tech Drive Hogan 2-160, Evanston, IL 60208, United States

HIGHLIGHTS

- Six machine-learning classifiers were combined into a multiple classifier system.
- Using multiple classifiers improves accuracy of automatic sleep scoring.
- At 1% rejection rate, the algorithm matches the accuracy of a human scorer.

ARTICLE INFO

Article history:
Received 23 October 2015
Received in revised form 11 February 2016
Accepted 23 February 2016
Available online 27 February 2016

Keywords:
Sleep
Sleep scoring
Autoscoring
Electroencephalogram
Mouse
Machine learning
Multiple classifier system

ABSTRACT

Background: Electroencephalogram (EEG) and electromyogram (EMG) recordings are often used in rodents to study sleep architecture and sleep-associated neural activity. These recordings must be scored to designate what sleep/wake state the animal is in at each time point. Manual sleep-scoring is very time-consuming, so machine-learning classifier algorithms have been used to automate scoring.
New method: Instead of using single classifiers, we implement a multiple classifier system. The multiple classifier is built from six base classifiers: decision tree, *k*-nearest neighbors, naïve Bayes, support vector machine, neural net, and linear discriminant analysis. Decision tree and *k*-nearest neighbors were improved into ensemble classifiers by using bagging and random subspace. Confidence scores from each classifier were combined to determine the final classification. Ambiguous epochs can be rejected and left for a human to classify.
Results: Support vector machine was the most accurate base classifier, and had error rate of 0.054. The multiple classifier system reduced the error rate to 0.049, which was not significantly different from a second human scorer. When 10% of epochs were rejected, the remaining epochs' error rate dropped to 0.018.
Comparison with existing method(s): Compared with the most accurate single classifier (support vector machine), the multiple classifier reduced errors by 9.4%. The multiple classifier surpassed the accuracy of a second human scorer after rejecting only 2% of epochs.
Conclusions: Multiple classifier systems are an effective way to increase automated sleep scoring accuracy. Improvements in autoscoring will allow sleep researchers to increase sample sizes and recording lengths, opening new experimental possibilities.

1. Introduction

Electroencephalogram (EEG) and electromyogram (EMG) combination recordings are often used to study sleep and circadian rhythms in both humans and animals. Using such EEG/EMG recordings, researchers can determine what sleep/wake state the animal is in at each time point during the recording period. This allows researchers to quantify an animal's sleep architecture, i.e. the timing and duration of different sleep stages, as well as to study the brain's electrophysiological activity during sleep. The recordings obtained this way are usually divided into short time segments, called epochs, which are manually scored to label what sleep/wake stage the animal is in.

However, sleep scoring is a very time-consuming, subjective, and monotonous process. Because of the high labor cost of scoring, sleep researchers have rarely done long-term EEG recordings. Circadian biologists, in contrast, commonly record activity continuously for a month or longer, and this disparity may be one reason why sleep science and circadian biology developed so separately in the past several decades (Dement, 2011). In addition to

* Corresponding author. Tel.: +1 847 491 5687; fax: +1 647 467 4065.
E-mail address: v-gao@u.northwestern.edu (V. Gao).

http://dx.doi.org/10.1016/j.jneumeth.2016.02.016
0165-0270/

recording length, scoring limitations also restrict sample size. Modern genomics (and other "-omics") studies typically require sample sizes of several hundred (e.g. Winrow et al., 2009), which quickly become burdensome to score.

Methods to automate sleep-scoring have been proposed to solve this problem. Early techniques were mainly based on logic-based threshold rules, with amplitude and frequency-derived features as inputs (Van Gelder et al., 1991; Itil et al., 1969; Neuhaus and Borbely, 1978). More recently, machine-learning classification algorithms have been applied to the task (Sunagawa et al., 2013 contains a good summary). These classifier algorithms are usually supervised learners, meaning that an algorithm is first "trained" on manually-scored example epochs, from which parameters for classification are derived; the rest of the recording is then scored based on these parameters. The supervised learning process works well for sleep scoring because it allows the algorithms enough flexibility to adapt to the unique characteristics of each animal. Supervised classifier algorithms such as support vector machine (Crisler et al., 2008), linear discriminant analysis, decision tree (Brankačk et al., 2010), neural nets (Robert et al., 1997), and Naive Bayes (Rytkönen et al., 2011) have previously been applied to sleep-scoring in rodents. Machine learning classification has also been used for other types of EEG analyses, such as for brain-computer interfaces (Müller et al., 2008) and epilepsy diagnosis (Subasi, 2007).

A very effective way to improve classification accuracy is to employ a multiple classifier system (MCS). Specifically, in the classifier fusion method, a collection of algorithms each classify the set of inputs individually, and then the classifications outputted from the individual algorithms are combined to form a composite best-guess. An MCS may be composed of many repeats of a single type of base algorithm, in which case it is referred as an *ensemble* classifier, or it may be composed of several different types of base algorithms. The MCS's success can be intuitively explained by the fact that each classifier algorithm is subject to different biases and weaknesses, i.e. they are diverse, and combining diverse classifiers prevents a single classifier's misclassifications from strongly affecting the results. Multiple classifier systems have been applied with success to EEG classification in a non-sleep context (Sun et al., 2007), and also to many machine-learning tasks such as handwriting recognition (Günter and Bunke, 2005), face recognition (Czyz et al., 2004), and medical diagnosis (Sboner et al., 2003).

Here, we demonstrate a multiple classifier algorithm for sleep scoring. We show that using the MCS improves accuracy over using a single classifier, and the MCS's accuracy was on par with a second human rescoring the same recording. We also show that scoring with a modest number of rejections greatly improves accuracy at the cost of only a small amount of additional human effort.

2. Methods

2.1. Animals and recordings

We used EEG/EMG recordings from mice to test the autoscoring method ($n = 16$). The mice were a mixture of A/J ($n = 4$), C57BL/6 ($n = 2$), (A/J X C57BL/6) F1 ($n = 7$) and (A/J X C57BL/6) F2 ($n = 3$). All recordings were 24 h long and were recorded under normal, baseline conditions on a 12L:12D light-dark cycle.

To collect the recordings, we implanted mice with EEG and EMG electrodes for sleep recording while under ketamine and xylazine anesthesia. The EEG electrodes were four stainless steel screws inserted through the skull over the cerebral cortex, and the EMG electrodes were two iridium/silver alloy wires inserted bilaterally into the nuchal muscles. The electrodes were part of a pre-fabricated head mount (Part #8201, Pinnacle Technologies, Lawrence, KS), which was fixed in place with glue and dental acrylic. Two channels of EEG were collected: one from prefrontal cortex (EEG2) and the other from more posterior cortex near the hippocampus (EEG1). We used PAL 8200 Acquisition software (Pinnacle Technologies, Lawrence, KS) to obtain recordings, which were then exported to European Data Format (EDF) files. Signals were recorded at 1000 Hz, but to speed up computation time, only every fifth sample in the signal was used, which effectively reduced sampling rate to 200 Hz. Recordings were divided into 10-s epochs for scoring, and each 24 h recording consisted of 8640 epochs. We used Pinnacle PAL 8200 Acquisition and Sleep Score software for data collection and manual scoring. Protocols were approved by the Northwestern University Animal Care and Use Committee.

2.2. Human scoring

Each recording was scored by two human experts: the primary scorer was used to train the classifiers and compare computer-human agreement, while the secondary scorer was used to compare human-human agreement. Using PAL 8200 Sleep Score software (Pinnacle Technologies, Lawrence, KS), the scorer viewed each 10-s epoch and labeled it as either Wake, rapid eye movement sleep (REM), or non-rapid eye movement sleep (NREM), or excluded it from analysis if the signal contained a major artifact. Each recording was scored to completion this way. EEG2 was considered primary while EEG1 was considered supplementary when making scoring decisions. Generally speaking, Wake epochs have low-amplitude, high-frequency EEG and high-amplitude EMG; NREM epochs have high-amplitude, low-frequency EEG and low-amplitude EMG; REM epochs have low-amplitude, high-frequency EEG and low-amplitude EMG. In addition, some characteristic EEG wave-shape differences between the different sleep-wake stages also aid in scoring. Some epochs with artifacts were excluded from the analysis. These mostly consisted of "spikes" in the EEG signal, which must last 2 s or more for that epoch to be excluded. An average of 38 epochs were excluded per recording; the most had 499 artifacts, and 7 recordings had no artifacts at all. An experienced scorer requires about 4 h of time to score 24 h of recording.

2.3. Computer scoring

2.3.1. Feature selection

For feature selection, we used a procedure similar to Rytkönen et al. (2011). Only EEG2 was used; using both channels of EEG reduced the accuracy of classification, perhaps because of a surplus of non-useful features. The EEG power spectral density of each epoch was obtained by short-time Fourier transform, using a Hamming window of length equal to the length of the epoch; this was done using the "spectrogram" function in MATLAB. The power spectral density was binned into 20 logarithmically-distributed power bands between 0.5 Hz and 100 Hz, such that the lower, more biologically-relevant frequencies had more fine-grain bins. EMG power between 4 and 40 Hz was also used as a feature. The 20 EEG and 1 EMG features formed a feature vector of 21 elements in total.

2.3.2. Training epochs

We wanted our training epoch selection process to mimic how one would score training epochs in actual use, so training epochs were selected in continuous blocks rather than at random. In addition, a challenging aspect of sleep EEG classification is that REM comprises only a small minority of epochs, about 3%, so we wanted to ensure that enough REM epochs were selected as training. To select our training set, a random REM epoch was selected, and the preceding 90 epochs (15 min) and following 90 epochs were selected as training scores. This process was repeated until a total of 720 epochs (2 h) of training were selected. The remaining