Classification of Sleep States in Mice using Generic Compression Algorithms

Owen Mayer*, Diane C. Lim†, Allan I. Pack†, and Matthew C. Stamm*

*Department of Electrical and Computer Engineering, Drexel University, Philadelphia, PA 19104
†Center for Sleep and Circadian Neurobiology, Division of Sleep Medicine/Department of Medicine, University of Pennsylvania School of Medicine, Philadelphia, PA

Sleep is associated with a variety of chronic diseases as well as most psychiatric, addiction and mood disorders. To analyze sleep patterns in rodents, researchers analyze polysomnogram data containing electroencephalographs (EEG) and electromyographs (EMG). However, the analysis is performed manually by a expert human scorer, which is a slow, time consuming, and expensive process that is also subject to known human error and inter-scorer inconsistency [1]. To address this, researchers have developed a variety of techniques to automatically classify rodent sleep states using features extracted from EEG and EMG signals [2]. In many approaches, researchers extract a variety of heuristic features from explicitly chosen spectral bands of the EEG and EMG signals [3]. However, human designed, heuristic features often do not capture complete salient sleep-state information, which leads to inferior classification performance.

In this work, we propose a new sleep-state classification system that eliminates the explicit feature extraction step used in previous approaches. To do this, we use the Normalized Compression Distance (NCD) to measure similarity between a test epoch (time-segment) of polysomnogram data to a dictionary of epochs with known sleep-states. The Normalized Compression Distance is an estimate of the (non-computable) Normalized Information Distance, which is a metric of the amount of common information between two strings [4]. Two sequences of polysomnogram data of the same sleep-state share more common information (i.e. sleep-state) than two sequences from different sleep-states, and thus have a smaller information distance. A dictionary is populated with epochs with known sleep-states. To classify epochs, we calculate the NCD to each dictionary entry using a generic compression algorithm, such as *bzip2*, thereby creating a vector of distances. The vector of distances is then used as the basis for classification. To test the efficacy of our proposed classification system, we conduct a series of experiments to assess classification accuracy. We also test the effects of using different compression algorithms for NCD calculation, as well as different classification methods. We achieve a 90.5% correct classification rate when using *bzip2* compression to calculate the NCD, with a dictionary containing 1500 epochs and a support vector machine classifier.

References

[1] K.-M. Rytkönen, J. Zitting, and T. Porkka-Heiskanen, "Automated sleep scoring in rats and mice using the naive Bayes classifier," *Journal of neuroscience methods*, vol. 202, no. 1, pp. 60–64, 2011.

[2] C. Robert, C. Guilpin, and A. Limoge, "Automated sleep staging systems in rats," *Journal of neuroscience methods*, vol. 88, no. 2, pp. 111–122, 1999.

[3] V.-M. Katsageorgiou, G. Lassi, V. Tucci, V. Murino, and D. Sona, "Sleep-stage scoring in mice: The influence of data pre-processing on a system's performance," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 598–601, IEEE, 2015.

[4] R. Cilibrasi and P. Vitanyi, "Clustering by compression," *Information Theory, IEEE Transactions on*, vol. 51, no. 4, pp. 1523–1545, 2005.

[5] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A symbolic representation of time series, with implications for streaming algorithms," in *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pp. 2–11, ACM, 2003.

Classification of Sleep States in Mice Using Generic Compression Algorithms

Owen Mayer*, Diane C. Lim†, Alan I. Pack†, Matthew C. Stamm*

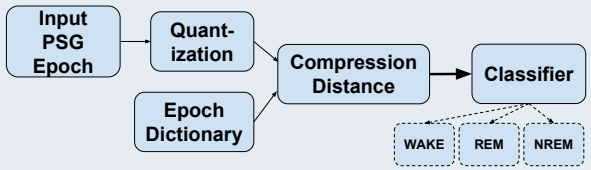
*Department of Electrical and Computer Engineering, Drexel University

†Center for Sleep and Circadian Neurobiology, Division of Sleep Medicine, University of Pennsylvania School of Medicine

Introduction

- Sleep behavior is key to the diagnosis and treatment of many chronic diseases as well as psychiatric, addiction and mood disorders
- Analysis of sleep data of mice is done manually
  - time consuming and expensive
  - prone to human inconsistency
- We propose a new method for automatic classification of mice sleep states
  - uses **compression based distances**
  - does not require explicit feature selection
  - parameter free (except choice of compression algorithm, classifier hyperparameters)

System Overview



1. Input Polysomnogram (PSG) epoch
   - 1 EEG channel + 1 EMG channel
   - 4 second epoch length, sampled at 256 Hz
2. Quantize the signal for compression compatibility
3. Calculate similarity via Normalized Compression Distances (NCD) to dictionary entries of epochs with known sleep-states
4. Use the vector of distances to classify the input epoch into 1 of 3 sleep states:
   (i) Wakefulness (WAKE)
   (ii) Rapid-eye-movement sleep (REM)
   or
   (iii) Non-REM sleep (NREM)

Author Contact: Owen Mayer – om82@drexel.edu

Compression Distance

The Normalized Compression Distance (NCD) is a measure of similarity between two objects in an arbitrary domain [4]

$$\mathrm{NCD}(x,y) = \frac{\mathrm{C}(xy) - \min\{\mathrm{C}(x), \mathrm{C}(y)\}}{\max\{\mathrm{C}(x), \mathrm{C}(y)\}} \quad (1)$$

- $\mathrm{C}(y)$ is the compressed size of string y via compression algorithm C (e.g. *bzip2, ppmz, lzma*)
- xy is the concatenation of strings x and y

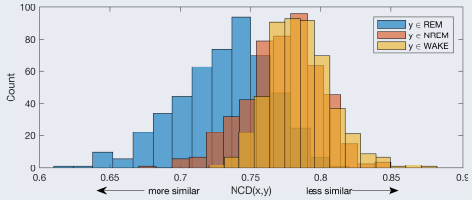
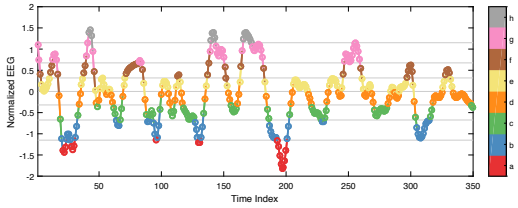


Figure 1: Histogram of normalized compression distances of a REM epoch, x, to 500 randomly chosen REM, NREM, and WAKE epochs

Polysomnogram Quantization

General compression algorithms require symbolic input

- Quantize PSG data into characters [5]
  - Linearly space bins in probability



- Concatenate quantized EEG and EMG signals
- Shift EMG symbol values to differentiate from EEG symbols

$$x = [\underbrace{aaabbabbaabbaa}_{\text{1-bit quantized EEG}} \ldots \underbrace{ccdddccccdddcdd}_{\text{1-bit quantized EMG}}]$$

- Best results achieved using 1-bit quantization

Classification

A dictionary of quantized PSG epochs with known sleep states

$$\mathbf{Y} = [y_1, y_2, \ldots, y_N]^\mathsf{T} \quad (2)$$

- y_i is i^{th} dictionary entry (quantized PSG string)

For an input PSG epoch, create a vector of distances $\mathbf{f}$

$$\mathbf{f} = [\mathrm{NCD}(x, y_1), \ldots, \mathrm{NCD}(x, y_N)] \quad (3)$$

Use $\mathbf{f}$ as input feature for classification

Results

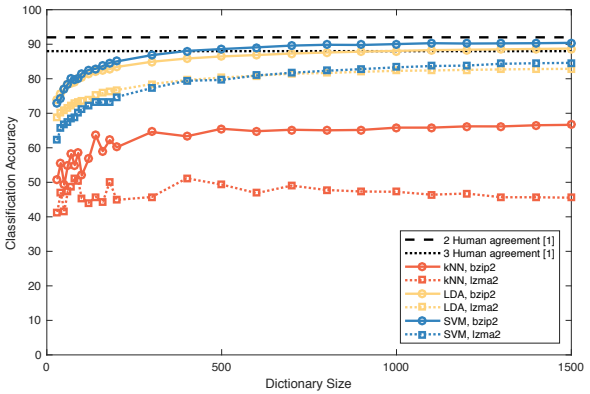


Figure 2: System classification accuracy versus dictionary size. Compression algorithm (*bzip2*, *lzma2*) and classifier type (kNN, LDA, SVM) are compared.

- Realtime classification (4s) with *bzip2* dictionary size 2000, *lzma2* dictionary size 1000

		Classifier		
		REM	NREM	WAKE
Human	REM	0.80	0.16	0.04
	NREM	0.01	0.91	0.08
	WAKE	0.00	0.09	0.91

Table 1: Confusion matrix for SVM classification with *bzip2*, 1000 dictionary entries. Overall classification accuracy is 90.2%

Dataset: 108000 epochs (120 hours) from 5 C57BL/6 mice

Training: 5-fold cross validation for SVM and LDA