2E-04

Leveraging consumer activity trackers for sleep stage prediction

Zilu Liang† Mario Alberto Chapa-Martell‡

†Kyoto University of Advanced Science ‡Silver Egg Technology

Introduction

Consumer activity trackers such as Fitbit have been increasingly used in longitudinal studies to track sleep patterns [1-3]. Data collected with these devices are often used for further secondary analysis [4-6]. However, these devices are known to be inaccurate especially for measuring sleep stages. In this study we propose a two-stage classification method to predict more accurate sleep stage from data that can be readily measured by Fitbit [7, 8]. Support vector machine was used in stage-1 classification to predict whether a Fitbit measurement is correct. If a Fitbit measurement is predicted as incorrect, it gets passed to the stage-2 model where the XGBoost algorithm is used to re-classify it into one of the four sleep stages: wakefulness, light sleep, deep sleep and REM sleep. In this paper we present promising results from a pilot evaluation study.

Data Preparation

We used a Fitbit Charge 2 activity tracker and a medical-grade single channel EEG device simultaneously to collect sleep data from 23 healthy young adults. We also used the Pittsburgh Sleep Quality Index (PSQI) questionnaire to collect demographic information (i.e. sex and age) and perceived sleep quality. From Fitbit public API we retrieved intraday data including sleep stages and heart rate at high resolution (below10 seconds). We developed a C# program to synchronize the sleep time series data with the heart rate time series data and to interpolate the data at a constant resolution of 1 second. Raw EEG signal data measured by the medical device was sent back to the company for analysis. The raw signals were firstly automatically scored using proprietary software and then manually inspected by doctors. The medical analysis results were used to label the dataset.

Model Construction

In total 21 features were constructed using the raw data from Fitbit and PSQI questionnaire. There are two types of features: time-dependent features and time-independent features. For time-dependent features, we first segmented the time series data into 30s epochs and then computed features in each epoch. These features included epoch ID, 30s-average sleep stage measured with Fitbit in each epoch, 30s-average heart rate measured with Fitbit in each epoch, the average sleep stage over the previous three epochs and the subsequent three epochs measured by Fitbit, and the changes in heart rate. Time-independent features include sex, age, PSQI, total sleep time, total wake time, sleep efficiency and ratio of each sleep stage.

The synchronized medical data was also segmented into 30s epochs and the average sleep stage was used to compute labels. For stage-1 classification, the labels were either 0 (i.e. Fitbit consistent with medical device) or 1 (i.e. Fitbit deviating from medical device). For stage-2 classification, the medical data were directly taken as labels, which were either 1 (denoting deep sleep), 2 (denoting light sleep), 3 (denoting REM sleep) or 4 (denoting wakefulness).

Support vector machine with linear kernel was used to achieve the stage-1 classification, while XGBoost algorithm was used to achieve the stage-2 classification. We used nested 10-fold cross-validation to tune the parameters of the models. In each iteration, the data of one participant was left aside as test set, while the data of the rest all participants were merged into a large training set. The training set was further divided into training set (70%) and validation set (30%) during 10-fold cross-validation. This process was repeated 23 times because the data of each participant was kept apart once as test set. One thing worth mentioning is that down sampling technique was performed during cross-validation due to the imbalanced nature of the dataset—one night of human sleep consists of dominantly more epochs of light sleep compared to the number of epochs of other sleep stages. These algorithms and techniques were selected because our previous studies demonstrated their advantages over other algorithms [9].

Evaluation

We used multiple measures to evaluate the performance of the model, including Cohen's *Kappa*, Matthews correlation coefficient *MCC* and micro-average F-score $F_{mic\text{-}avg}$. These measures are more suited for imbalanced and multiclass problems than traditional measures such as overall sensitivity, specificity. Table 1 shows a comparison between the proposed model and the baseline model (i.e. Fitbit proprietary algorithm) in terms of the three performance measures. The results showed that the proposed model improved *Kappa* and *MCC* by 14.3% and 13.3% respectively.

The box-whisker plots in Figure 1~3 demonstrates the distribution of the whole cohort on the three performance measures. From bottom up, a box-whisker plot displays the minimum, first quartile, median, third quartile and maximum. The plots

indicate that the proposed model outperforms the baseline model with statistical significance.

Table 1: Performance comparison between proposed model and baseline model (Fitbit proprietary algorithm).

	Kappa	*MCC*	$F_{mic\text{-}avg}$
Baseline model	0.37	0.39	0.50
Proposed model	0.42	0.44	0.50

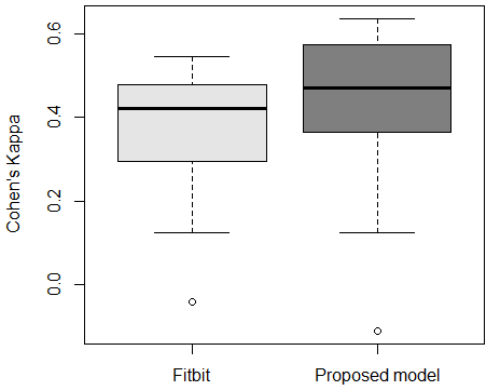

Figure 1: Comparison between proposed model and Fitbit proprietary algorithm on Cohen's *Kappa*.

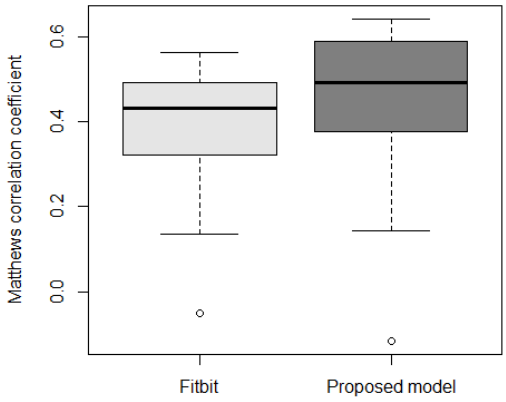

Figure 2: Comparison between proposed model and Fitbit proprietary algorithm on *MCC*.

Conclusion

In this study we proposed a two-stage model to predict more accurate sleep stage from Fitbit data. Performance evaluation showed that the proposed model improved *Kappa* and *MCC* byto 14.3% and 13.3% respectively, and the proposed model outperformed Fitbit with statistical significance.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number 16H07469 and 19K20141.

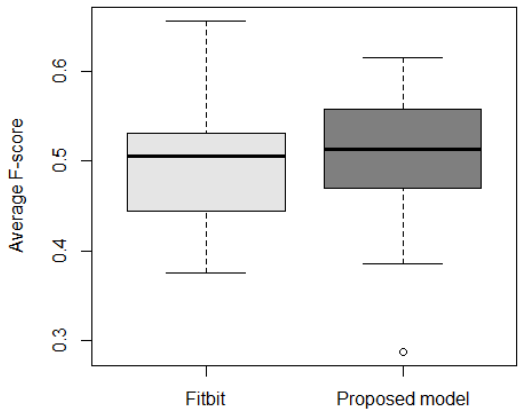

Figure 3: Comparison between proposed model and Fitbit proprietary algorithm on average F-score.

References

[1] G. Weaver, M. Beets, M. Perry, et al, "Changes in children's sleep and physical activity during a one-week versus a three-week break from school: a natural experiment," *Sleep,* vol. zsy205, 2018.

[2] J. Weatherall, Y. Paprocki, T. M. Meyer, et al, "Sleep tracking and exercise in patients with type 2 diabetes mellitus (step-D): pilot study to determine correlations between Fitbit data and patient-reported outcomes," *JMIR Mhealth Uhealth,* vol. 6, no. 6, pp. e131, 2018.

[3] Z. Liang, B. Ploderer, W. Liu, et al, "SleepExplorer: a visualization tool to make sense of correlations between personal sleep data and contextual factors," in Personal Ubiquitous Comput., 2016, pp. 985-1000.

[4] Z. Liang, M. A. Chapa-Martell, and T. Nishimura, "A personalized approach for detecting unusual sleep from time series sleep-tracking data," in Proc of the ICHI'16, Chicago, US, 2016.

[5] Z. Liang, M. A. Chapa-Martell, and T. Nishimura, "Mining hidden correlations between sleep and lifestyle factors from quantified-self data," in Proc of UbiComp'16: Adjunct, Germany, 2016, pp. 547-552.

[6] Z. Liang, B. Ploderer, M. A. Chapa-Martell, and T. Nishimura, "A cloud-based intelligent computing system for contextual exploration on personal sleep-tracking data using association rule mining," *Intelligent Computing Systems. Communications in Computer and Information Science*, Springer, 2016.

[7] Z. Liang, and M. A. Chapa-Martell, "Validity of consumer activity wristbands and wearable EEG for measuring overall sleep parameters and sleep structure in free-living conditions " *Journal of Healthcare Informatics Research*, pp. 1-27, 2018.

[8] Z. Liang, and M. A. Chapa-Martell, "Accuracy of Fitbit wristbands in measuring sleep stage transitions and the effect of user-specific factors," *JMIR Mhealth Uhealth,* vol. 7, no. 6, pp. e13384, 2019.

[9] Z. Liang, and M. A. Chapa-Martell, "Achieving accurate ubiquitous sleep sensing with consumer wearable activity wristbands using multi-class imbalanced classification," in Proc of PICOM'19, Fukuoka, Japan, 2019.