Sleep Stages Classification

Brittany Pine, Shivangi Singh, Ruifeng Wang, Eric Atwood

Overview

The application classifies sleep stages (light, deep, awake) using heart rate in beats per minute from the Huawei Watch 2 and accelerometer readings from an android phone. Both of these metrics are useful in classifying sleep stages because sleep stages can be classified by movement (by acceleration) and brain waves (with heart rate as a proxy). While there are more sleep stages than the three listed, those are the three we are focusing on for this classification algorithm. We would like to reproduce something similar to fitbit's sleep app.

Figure 1: Example of typical Sleep Cycle from fitbit

Data Collection

- Acceleration: data collected at high frequency though phone accelerometer
- Heart Rate: bpm reading every five minutes from the watch through Heart Trace app
- True values: data collected from SleepTime app

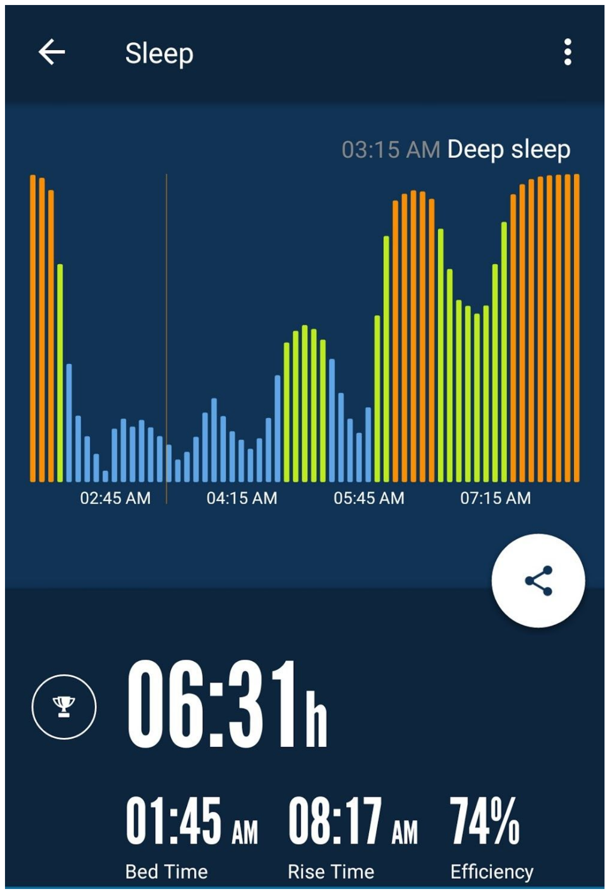


Figure 2: A night's worth of "true" values for training data from SleepTime

For labelling the "truth" values for the sleep stages we used data provided by the SleepTime app on the phone. We ran the app while collecting our data through the accelerometer and the watch. The application classified the sleep into deep sleep, light sleep, and awake. The phone was kept face-down on mattress, next to the head so that it could track acceleration. As we do not know the internal workings of the application, we are not aware of the features it uses apart from the accelerometer readings. However, we do know that this app does not use heart rate readings, so it is possible that the "true" values that we are comparing to are actually less definite than what we would be able to detect using the additional sensor.

Problems

Initially, we could not collect accelerometer data throughout the night because the phone would disconnect from the server so data collection was interrupted. We made changes to the application so that the phone stored data internally instead of sending it to a server.

We initially wanted to collect audio data, but we decided that there would be too much noise interference for it to be useful because we all live in shared living spaces.

There is a great discrepancy in the amount of readings collected for acceleration and heart rate as the accelerometer collects readings per millisecond and the watch collects one heart rate reading per 5 minutes. We have approximately 3,000 accelerometer readings per one heart rate reading in our data. To account for this, we mapped a single heart rate reading to all corresponding acceleration values based on matching time intervals. (All accelerometer readings after a given heart rate reading corresponded to that heart rate until the two time stamps were equal. Then the next corresponding heart rate in the data was used.)

Originally, the timestamp for the acceleration readings was the uptime (milliseconds since the phone was last turned on). We did not find this metric useful because it had to be mathematically converted each time. We changed the data collection so that the time would be used instead of the uptime. This created another problem; when reading in the file, all types were expected to be integers. We had to change it to take in both integers and datetime object, which proved to be difficult.

Results

We are compiling average accuracy, precision and recall on the test data over a 10-fold cross validation. As we have to manually export data from the phone's internal storage, we do not have an integrated app to classify data in real time, but we can later predict sleep stages by running a python script. We have collected sleep data from five different people to incorporate different sleeping patterns. Our results for different classifiers and varying parameters are shown below. The best classifier for our data data was a random forest with n=100 and max depth of 10.

```
Average accuracy: 0.784
Average precision awake: 0.742
Average recall awake: 0.786
Average precision light sleep: 0.815
Average recall light sleep: 0.701
Average precision deep sleep: 0.864
Average recall deep sleep: 0.799
```

Decision tree (criterion="entropy", max_depth=5, max_features = 5)
(Data from two people, the rest of results are from four people)

```
average accuracy: 0.913
average precision awake: 0.942
average recall awake: 0.956
average precision light sleep: 0.637
average recall light sleep: 0.946
average precision deep sleep: 0.930
average recall deep sleep: 0.766
```

Decision tree (criterion="entropy", max_depth=5, max_features = 5)

```
average accuracy: 0.915
average precision awake: 0.948
average recall awake: 0.988
average precision light sleep: 0.495
average recall light sleep: 0.978
average precision deep sleep: 0.886
average recall deep sleep: 0.920
```

Decision tree (criterion="entropy", max_depth=5, max_features = n_features)

```
average accuracy: 0.968
average precision awake: 0.963
average recall awake: 0.980
average precision light sleep: 0.930
average recall light sleep: 0.990
average precision deep sleep: 0.978
average recall deep sleep: 0.877
```

SVM (C-Support Vector Classification, all default parameters)

```
average accuracy: 0.919
average precision awake: 0.900
average recall awake: 0.942
average precision light sleep: 0.856
average recall light sleep: 0.982
average precision deep sleep: 0.949
average recall deep sleep: 0.695
```

Random Forest (n=50, “entropy”, max_depth = 5)

```
average accuracy: 0.979
average precision awake: 0.983
average recall awake: 0.980
average precision light sleep: 0.968
average recall light sleep: 0.988
average precision deep sleep: 0.994
average recall deep sleep: 0.895
```

Random Forest (n=50, “entropy”, max_depth = 10)

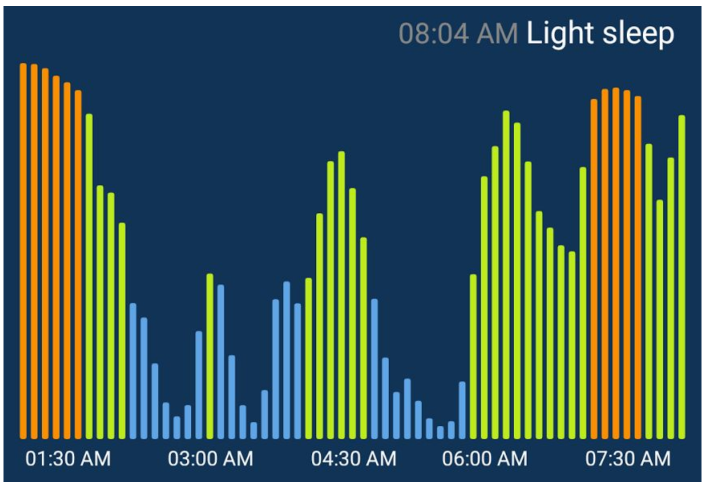
```
average accuracy: 0.920
average precision awake: 0.900
average recall awake: 0.944
average precision light sleep: 0.858
average recall light sleep: 0.983
average precision deep sleep: 0.953
average recall deep sleep: 0.691
```

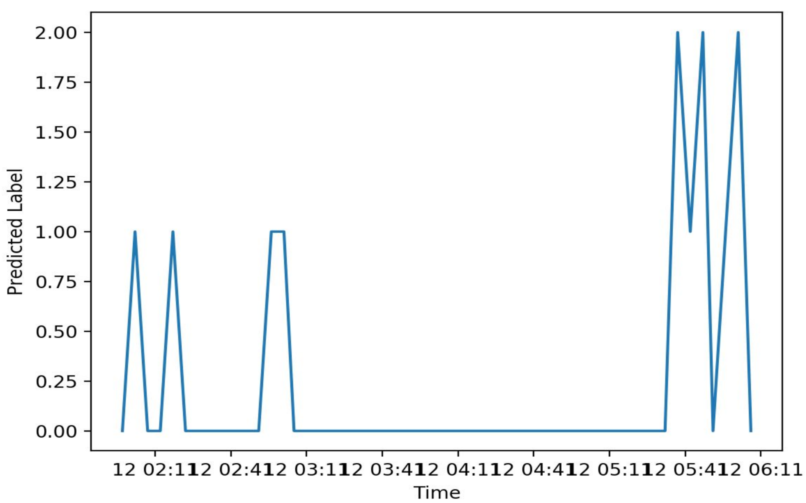
Random Forest (n=100, “entropy”, max_depth = 5)

```
average accuracy: 0.980
average precision awake: 0.983
average recall awake: 0.981
average precision light sleep: 0.969
average recall light sleep: 0.988
average precision deep sleep: 0.994
average recall deep sleep: 0.903
```

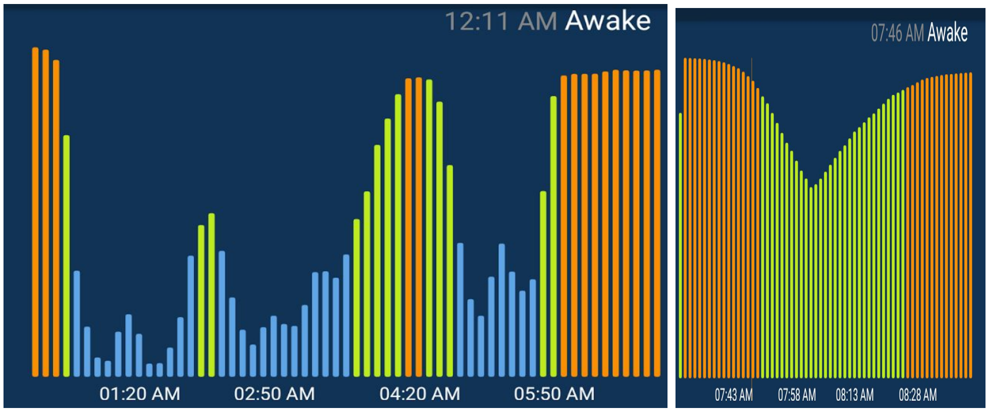
Random Forest (n=100, “entropy”, max_depth = 10)
With every metric above 90 percent, this was the best classifier.

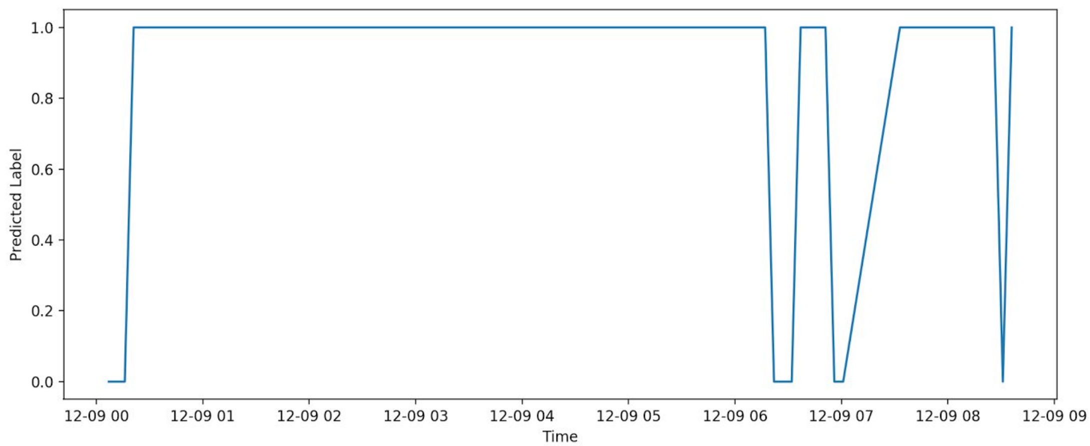
We used the data from the fifth person to compare their predicted sleep stages to true sleep stages. The following graphs should show an inverse relationship because we classified “2” as deep sleep (so it is the highest point on the line chart), but in SleepTime a shorter bar means deep sleep. While our confusion matrices on the training data show very good classification, the predictions did not come out well. First is the comparison of the SleepTime classifications compared to our classifications for a sample that was not included in the test set. The predictions were that the participant was mostly awake during the night which is untrue. Then is a comparison for data was in the test set. While never predicting deep sleep, it is generally differentiating between sleep and awake.





Figures 3 and 4: SleepTime classification vs our own for new data





Figures 5 and 6: SleepTime classification vs our own for data in the test set

Future Possibilities

We would want to collect more training data because while the accuracy, precision, and recall are all quite high, the predictions on non-labelled data were not very good.

Currently, we do not have a UI that integrates the watch's heart rate and the phone's accelerometer readings in real time to make predictions about the sleep stages. That would be the next step to make this a usable app.

References:
https://pub.tik.ee.ethz.ch/students/2016-FS/BA-2016-02.pdf