Machine Learning based Blood Pressure Prediction and Health Behavior Recommendation

Kunmao Li, Sujit Dey (PI)

Department of Electrical and Computer Engineering
University of California, San Diego
kuli@eng.ucsd.edu, dey@ece.ucsd.edu

Abstract— **In order to break the limits of impracticality of traditional cuff-based blood pressure measurement and inaccuracy of photopletysmography-based approaches, this project aims to construct the model for both efficient and accurate blood pressure prediction with the data from wearable devices. The corresponding results can be utilized to facilitate the daily blood pressure monitoring with effortless operation and personalized health behavior recommendation. This paper is the summary of the first stage of this project.**

I. Introduction

Well-known as the ”the silent killer” with no symptons, high blood pressure (hypertension) has contributed to many deaths, which manifests the significance of regular measurement of blood pressure (BP). Since the conventional approach using mercury sphygmomanometers is impractical for the application of home daily measurement[1], this project seeks for alternate BP estimation with the support of machine learning techniques in order to facilitate the BP monitoring but ensure the accuracy simultaneously.

Integrated with Artificial Intelligence and Machine Learning, wearable devices have the potential for shaping the future of healthcare[2] thanks to the characteristics like portability, in-time monitoring, comprehensive health and fitness features tracking.

The ultimate objective of the project is to maximize automated and continuous user data collection, build clean and comprehensive data set with additional features extracted and construct machine learning models for personalized BP analysis and prediction. The health behavior guidance tailored for the users on the chronic conditions can thus be given based on the corresponding results. During this summer internship program, I have completed the first-stage work, which included the data parsing, data set building and preliminary training model design.

II. Data Parsing and Data Set Building

A. Data Collection

We monitored and collected health and exercise related data from up to 90 participants in 3 consecutive months, the duration set for them uniformly. The summarized procedure of data collection that can be mainly divided into two parts.

1) Omron Blood Pressure: The application Omron Connects on the mobiles phones of participants helped to transfer the data from the BP monitors to the Omron Cloud Service. We could either utilize the Omron Application Programming Interface (API) built to download the data from Omron Cloud Service to our server or directly request the data from Omron Connect by OAuth authentication.

2) Samsung Health Data: The health and fitness related attributes such as calories, steps, floors, assorted kinds of exercises, sleep and heart rate could be recorded by Galaxy Watch. For the access of data, the Samsung Health mobile application was connected to Samsung Cloud Service, from which we were able to download all data.

B. Data Preprocessing

1) Blood Pressure: The raw text data requested from Omron could be easily converted to the primitive data type like dict in python, which would then be further parsed into the data frames. The amount of BP records for one participant in the same day was no smaller than two, the minimum measurement times stipulated out of the consideration of viability and practicality. The corresponding features extracted include the timestamps in Pacific time zone, systolic BP (when the heart is contracting, diastolic (when the heart fills with blood and gets oxygen), pulse and etc.

2) Health and Fitness Data: The health and fitness data collected by Galaxy Watch and requested from Samsung S Health were much thornier to deal with. First of all, there were a large amount of different features containing different attributes that were not shared across the features. Thus, different features were processed separately into disparate data frames, after which they could be further merged into one. In addition, the original data requested were segmented and compressed based on the record time and event. The data for different features were also compressed in different manners. For instance, one segment of heart rate data contained 2000 records, each of which could be decompressed into multiple data points with the interval of one minute. For exercises, both measurements (i.e., timestamps, duration, heart rate, speed, distance, calories) and contextual data (i.e., longitude, latitude, altitude, exercise type, user identity) were compressed. Extra effort for the match of timestamps, measurements and contextual data was needed.

In the next step, the abnormal data points would be regarded as outliers and then discarded. Data points were classified as abnormal if they were unreasonable (i.e., heart rate lower than 40), experienced abrupt changes or had insufficient amount (i.e., daily sleep time less than 3 hours).

The high rate of missing measurements was also observed. We expected to collect the heart rate data covering all day while the daily total duration was 14 to 15 hours per day. Similarly, the other types of data were also not continuous as they were not in the fixed-width intervals. The intervals ranged from minutes to hours with the data points missing, presumably because the connection was lost or the participant did not wear the device. To solve this problem, the cubic spline interpolation was utilized to re-sample the data into fixed intervals of 1 minute.

C. Data Set and Features

The features used for model training and prediction were summarized in the table I.

Features	Attributes	Features	Attributes
Time	Pacific Timestamp Day of Week	Bood Pressure (BP)	Diastolic BP Systolic BP
Heart Rate (HR)	HR per Minute Resting HR Max HR HR zone	Step Count	Count Distance Speed
Floors climbed	floors	Calories	Calories per Minute Daily Calories
Sleep	Sleep Duration Wake-up Time Bed Time Sleep Stage	Contextual	Age Longitude Latitude Altitude Additional Exercise Type

TABLE I: Description of Features

There were additional attributes computed and introduced by me for further model construction. The same measurements (i.e., heart rate, distance, speed) were matched and merged if the same timestamps were observed. The explanation of some features is given as follows.

- Max HR: 220−age
- Resting HR: the heart rate right after the wake up time
- HR Zone: the ratio of heart rate to max heart rate
- Sleep Stage: 1(awaken), 2(light sleep), 3(deep sleep), 4(Rapid Eye Movement)
- Additional: the additional contextual data for exercise (i.e., stroke type and pool length for swimming)

There were two versions of the data set for short term and long term prediction respectively. The first contained the data records with the fixed sampling interval of 1 minute so that the predictions derived could correspond to a fixed amount of minutes in the future. Data with long intervals across two records were discarded rather than imputed. For small amount of missing values, they could be tackled by linear interpolation and cubic spline interpolation to make sure the same sampling intervals. For sleep stage, value of 1 representing the awaken stage was uniformly assigned to the timestamp when participant was not in sleep. For attributes like count, distance and speed, the missing values would be imputed by 0 since no movement was made by the participant.

The second data set was acquired by down sampling the first data set into the daily frequency with aggregating the data. The daily data was not from one natural day but from the past 24 hours ending at each BP record. For the attributes like heart rate and speed, average values would be computed by the data with fixed 1-minute intervals. For the attributes like distance and sleep duration, the daily value was the summation of all original values in 1-minute intervals.

Both data sets were finally sorted in the chronological order of the timestamps.

III. Training Model Construction

A. Training Procedure

Both of the data sets were split into 80% for training, 10% for validation and the final 10% for testing. The target was the blood pressure and the input was formed by all features except blood pressure. Only the input was scaled and it was time-shifted by 3 days to find out how the health behaviors in the previous few days would influence the blood pressure.

The preliminary model was based on random forest (RF) for its capabilities of handling many input variables, running efficiently on large databases and robustness to overfitting. As there were a large amount of input features in the data set and the high-dimensional data tend to be unreliable, dimensionality reduction was required to enhance the prediction accuracy. After the preliminary training, the feature importances were further ranked for identifying the most vital features essential to BP, so that the input features could be pruned. For the evaluation of the training performance, the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) were used.

B. Results

The MAE and RMSE of the prediction by the preliminary model were demonstrated in the table For the data set with 1-minute intervals, RMSE and MAE were 10.20 and 2.67 for diastolic BP and 12.03 and 2.91 for systolic BP. For the other data set, RMSE and MAE were 12.25 and 3.52 for diastolic BP and 13.16 and 3.84 for systolic BP.

In addition, the three most important features influencing the growth trend of blood pressure would be heart rate, bed time, count according to the preliminary prediction.

IV. Conclusion

I collected and parsed the health and fitness data, which were further used for building the model to predict the blood pressure in both short term and long term. According to the preliminary training results, heart rate, bed time and count were closely related to the tendency of the blood pressure. More results of my work can be viewed at https://github.com/kwanmolee/Blood-Pressure-Prediction-and-Peronalized-Health-Behavior-Recommendation.

References

[1] Buxi D., Redoute J. M., Yuce M. R. ”Cuffless blood pressure estimation from the carotid pulse arrival time using continuous wave radar.” Proceedings of Annual International Conference of the IEEE Engineering in Medicine and Biology Society; August 2015; Milano, Italy. pp. 57045707.

[2] De Pessemier, Toon, and Luc Martens. ”Measuring heart rate with mobile devices for personal health monitoring.” 12th International Conference on Internet and Web Applications and Services (ICIW 2017). 2017.