Determining Individual Characteristics from Human Activity Data (iPhone Motion Sensor)

Introduction

Ofcom's figures indicate that in 2020 45 percent of households in the UK had two mobile phones. Published figures indicate there were over 53 million smart phones in the UK in May 2021 and over 2,700 million globally (O'Dea, 2021). It is therefore important to understand how the data they are capable of gathering about our physical activities could be used to infer characteristics about the phone users. The question I am interested in answering is - Can activity data recorded on a smartphone be used to infer individual characteristics about weight, height, gender, age, and if so, to what degree of accuracy?

Related Work

Research studies have established it is possible to recognising the type of activity that the human is undertaking to a high degree of accuracy. Researchers from University of California, Irving undertook experiments with a group of 30 volunteers within an age bracket of 19-48 years. The dataset is publicly available and has been used as a data source for many research papers (Reyes-Ortiz, 2012). Each person performed six activities wearing a smartphone (Samsung Galaxy S II) on the waist. Using its embedded accelerometer and gyroscope, linear acceleration and angular velocity was captured at a constant rate of 50Hz.

On this similar dataset, support vector machine classifier and convolutional neural network machine learning models achieved almost 100 percent accuracy when determining activity type (Sun, 2019). Analysis established that within a short amount of time (1-1.5 min) the smartphone has enough data to determine what its user is doing (95%: 6 activities) or who the user is (Walking 94%: 30 participants) and even the basics of a person's specific walking style (Liftoff, 2019).

Gender and average weight prediction using this dataset (Huang, 2019).
Exploratory analysis of data subjects (Shaar, 2019).

Broader context

Cardiac abnormalities are one of the leading causes of deaths all over the globe. Smartphone-based systems may offer opportunities for real-time cardiac monitoring and early abnormality detection (Shabaan et. al, 2020). It is important to test the accuracy of predictive models utilising smartphone data.

A paper on wearable devices identified 423 unique devices from 132 different brands. With recent advances in mobile sensor technology, privately collected physical activity data can be used as an addition to existing methods for health data collection in research (Henriksen et. al, 2018). Given the popularity of wearable mobile sensors, it is important to determine whether smartphone activity data could provide additional personal attribute data about the user.

If users can be personally identified by smartphone motion data the implications are that sensor data could be biographically significant legally, under the General Data Protection Regulations (Liftoff, 2019). Therefore data privacy concerns need to be addressed when data 'relates to or is 'obviously about' a particular individual. It is important that smartphone users understand the data that can be captured. This will allow informed choices to be made and privacy trade-offs to be transparent. Users may find monitoring useful for improving the quality health interventions.

Hypothesis

The hypothesis being assessed is whether personal characteristics of weight, height, age and gender can be predicted from motion sensor captured by a smartphone. The analysis will compare model accuracy.

Experiment

An experiment at Queen Mary University of London generated the dataset being reviewed. Time-series data was generated by accelerometer and gyroscope sensors. It was collected from an iPhone 6s kept in the participant's front pocket using SensingKit which collects information from Core Motion framework on iOS devices. All data collected in 50Hz sample rate. A total of 24 participants in a range of gender, age, weight, and height performed 6 activities in 15 trials, of varying length, in the same environment and conditions: downstairs, upstairs, walking, jogging, sitting, and standing. Full study conditions are provided by the researchers to allow repetition (Malekzadeh, 2019).

Summary of analysis

Following the CRISP-DM methodology, I researched the dataset to understand context and reviewed existing analysis. I performed exploratory data analysis and generated numerical and graphical summaries to understand the data associated to the trial participants.

There were 24 data subjects in the trial, 10 were females and 14 males. The raw data from the 15 activities conducted by the 24 participants was loaded and grouped into the 6 activities. Exploratory analysis was conducted to ascertain spread of height, weight, gender, age for the trial participants as a group, and then subdivided by gender. A histogram, correlation matrix and scatter plot was generated for the whole group, then for males and females. Minimum, maximum and mean values were calculated for each attribute for the whole group, then for males and females.

I based my exploratory analysis of trail participants on the work performed by Shaar (Shaar, 2019). I reused and extended his work by cleansing the data to move code, which was a unique identifier and didn't add value to the analysis. I generated additional graphs splitting the data by gender, histogram/bar chart, correlation matrix and scatter/density plots. In addition, I produced a summary tables for minimum, maximum, average variable values for the whole dataset and also by gender.

For the trial output data, I added labels to allow opportunities to use supervised learning as well as unsupervised machine learning models and techniques. I cleansed, grouped, filtered, performed feature engineering and scaled the data in preparation. I ran a series of linear and non-linear, machine and deep learning models and evaluated them using a variety of metrics – accuracy, f1 score, time to compile and error rate by confusion matrix.

The multivariate time-series has 12 features: attitude.roll, attitude.pitch, attitude.yaw, gravity.x, gravity.y, gravity.z, rotationRate.x, rotationRate.y, rotationRate.z, userAcceleration.x, userAcceleration.y, userAcceleration.z. The data was loaded with added identifiers for the experiment number, participant number and activity type. Exploratory analysis was conducted to ascertain class balance, hierarchical clustering and a sample of the timeseries data (userAcceleration.x) was examined by activity type. In total there were over 1.4m observations with 16 attributes.

Best Subset Selection analysis was conducted on the whole dataset, using a Logistic Regression model. I repeated the Best Subset Selection analysis using two subsets of the data, on the Jogging activity type and on the Standing activity type. I evaluated the output using C_p, AIC, BIC and Adjusted R2.

I extended the work produced by Y.C. Huang (Huang, 2019). I utilised his approach to feature construction and engineering based on that of ROYT's (T, 2019) allowing use of all complete dataset. In addition to gender and weight prediction I extended Huang's work to include mapping for height and age. I split the data to allow for training and testing of eight classifiers. I scaled the training data and applied a variety of linear and non-linear deep and machine learning models – Support Vector Machine Linear, Support Vector Machine - Radial Basis Function (RBF), Logistic Regression (L1), Logistic Regression (L2), Decision tree, Random Forest, K- Nearest Neighbour.

The response variable was adjusted to determine model accuracy for predicting activity type, trail participant and activity type, trial participant, gender, weight, height and age.

In addition to model accuracy scores, f1 score values were produced and a confusion matrix was generated for the best performing model. Analysis of the importance of attributes was also conducted.

Sliding window, channel-normalisation, scaling were applied to a training and test dataset. Additional feature engineering was applied to get feature that could be linearly separable, low dimensional and structured so that all the classifiers could be used.

Logistic Regression is shallow learning (linear), Multi-Layer Perceptron is deep (non-linear).

Logistic regression - measure the accuracy of a regression model is by calculating the root mean square error (RMSE), a metric that tells us how far apart our predicted values are from our observed values in a model, on average

classification model - response variable is categorical, e.g. on or off. The most common way to measure the accuracy of a classification model is by simply calculating the percentage of correct classifications the model makes.

Regression is an algorithm in supervised machine learning that can be trained to predict real number outputs. Classification is an algorithm in supervised machine learning that is trained to identify categories and predict in which category they fall for new values.

If there exists a hyperplane that perfectly separates the two classes, then we call the two classes linearly separable. Few examples of linear classifiers are Logistic Regression, Perceptron, Naive Bayes, Support Vector Machines, etc. Examples of non-linear classifiers are Decision Trees, K-Nearest Neighbour, Random Forest, non-linear-kernel Support Vector Machines, etc.

Classifiers

K-Nearest Neighbour, Euclidean most popular distance, better for small data
Logistic Regression – linear classifier, linearly separable data
Linear Support Vector Machine – uses margin
Support Vector Machine – Radial B F (non-linear), uses marging
Decision Tree
Random Forest
Multi Layer Perceptron (more than one layer) – non-linear separable data, difficult to see relationship between input and output, doesn’t scale well.

Deep learning refers to neural networks with multiple hidden layers that can learn increasingly abstract representations of the input data.

A multilayer perceptron (MLP) is a class of feedforward artificial neural network (ANN). An MLP consists of at least three layers of nodes: an input layer, a hidden layer and an output layer. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP utilizes a supervised learning technique called backpropagation for training. Its multiple layers and non-linear activation distinguish MLP from a linear perceptron. It can distinguish data that is not linearly separable.

The Random Forest model performed better than the Decision Tree model, this is possibly due to the dimensionality of the data being high and decision boundary being smooth. Decision tree models can easily overfit, cope poorly with high-dimensionality data and have a block effect to decision boundaries.

Results & Conclusion

Even when the limitations of the dataset are considered, it is reasonable to conclude that personal characteristics of weight, height, age and gender can be predicted to a high degree of accuracy when using machine learning models to evaluate human activity sensor data.

The sample size of trial participants was small with less than 30. If the sample size is too small it may be difficult to detect what was intended (Kar & Ramalingam, 2013). The average age of the participants was 29. The information gathered may not be reflective of the population, i.e. the speed of movements within younger individuals may not be reflective of an older population. Therefore any predictions on age given the small sample may not be an accurate prediction of test error.

Performance of the models varied significantly with the Random Forest Classifier having the highest accuracy for each variable. Models ranged in accuracy of predicting gender 70% to 97%, height 62% to 96%, age 63% to 94% and weight from 61% to 96% accuracy.

Confusion matrices showed that the Random Tree models for gender, weight and height had a higher number of false positives and age a higher number of false negatives. The false positives incorrectly predicted that gender was male or weight and height were above average. The false negative incorrectly predicted an age lower than the average.

Evaluate success

Scaling the data improved the accuracy of the models. Due the high dimensionality of the data set is was necessary to apply feature engineering if using the whole dataset. I was able to run Best Subset Selection analysis (Appendix x) on the whole dataset using Logistic Regression, unfortunately none of the other models would compute using Google Colab resources in a timely manner. However as the whole dataset appears to be non-linear in nature the results of the Best Subset Selection didn't allow reduction of the dataset. The analysis is included for completeness.

The size of the data set by activity type differed due to the varying length of the trails. The datasets for the downstairs and jog were the smallest data sets, upstairs and standing next in order, with sitting and walking datasets the largest. Although class size varied because on varying length of the trials I chose not balance the data. Despite selecting a small dataset, the dataset associated to the activity type of Jogging could not be analysed using Google Colab resources in a timely manner.

Due to the similarities in activity types the data could have been grouped into three subsets - downstairs, upstairs and walk 2) sitting and standing 3) jogging. It is expected that model accuracy would have been improved even further.

Future implications

Reflections

Being honest this assignment, like all the others, has been challenging. Even reminding myself how to run the GitLog file took time and persistence. Figuring 'stuff' out on my own and being resilient has its rewards but feels uncomfortable and difficult. In terms of personal and professional development I really enjoy, and hate, the learning process. I struggle being out of my comfort zone, feeling very anxious, stressed and frustrated at times. I need time to absorb information and be able to gain confidence in the topics being taught. Therefore managing my time well has been very important.

Developing trusted relationships with other students has been very useful particularly in determining if I've understood the material in the way that was intended, or at least they understood it. Appreciating that others have a different learning approach has also been valuable and quite insightful. Our differences as humans make us unique and recognising there are more than several 'right' ways has been eye-opening. In future I'd like to think I will be more flexible with how I define tasks for my team to allow them more scope for personal style. When planning learning and development tasks in future I will make sure to allow sufficient time for information to be assimilated, whether for myself or others.

The process, technologies and methodologies I used involved a mixture of new and previously used. All of the approaches were still new to me.

Following feedback from previous assignments, I took more time to follow the Scientific Method and used that format for my report. I attempted to be clearer on how I had identified the problem, approached it, defined it, then selected to model it, evaluate how successful this had been and conclude what future work could be undertaken. I took more time, and words, to explain the rationale for my decisions.

I used the CRISP-DM best practice methodology. The lifecycle contains six phases; Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, Deployment. The process allows for iteration between stages and I did move between data preparation, modelling and evaluation extensively.

I created a GitHub repository using the cookiecutter Data Science template in GitHub. I used Git for version control. I developed models using Python on Google Colab with data stored on Google Drive. The cookiecutter folder structure was useful for organisation and keeping order of my work. Google Colab provided easy access to sklearn library, keras and tensorflow code. I was not able to set up RStudio locally with the Python extensions and 'reiterate' library using a virtual environment.

Google Colab resources had limitations. The dataset needed dimensional reduction to run all the models within reasonable timescales. I needed to split the code into separate notebooks to keep the size of the Python notebooks manageable. Uploading files directly from Google Colab needed the GitHub repository to be set to 'public'. Change of access and visibility needs password approval and additional text so switching this on and off was fiddly, although not difficult. As a result I limited my commits to when there was a reasonable amount of work to update within the repository. I would recommend Google Colab based on my experience. Although I had not code in Python, I did find it reasonably intuitive but handling code errors in Jupyter notebooks was not as easily identifiable or rectifiable as R. There were some sections of code I was unable to generate due to library problems in Google Colab (dtiadistance gave a C language error, for example).

I generated a significant amount of machine learning models, including sliding window, feature engineering on time-series data. I also explored more on the time-series format and generated line graphs by attribute and gender. However whilst I recognised some of the code was more advanced, and technically more difficult than the other models I had produced the output didn't add to the intent of my report. During the development of my report I cut back the content significantly to ensure it was focused. I did repeat Huang's analysis of data characterization (shape of dataset, missing values, class balance), dimension reduction or hierarchical clustering but have not included any analysis in this report (Huang, 2019).

Detailed findings of the exploratory analysis

There were 24 data subjects in the trial, with average weight of 72kg (11 stones 4lb), average age of 29 and average height of 174cm (5ft 8.5') Fourteen males and ten females participated.

The histograms for the data show the variation in weight, height, age and gender of the participants in the trail. The average weight of all trail participants was 72.125kg, average height was 174.2cm and average age of 28.79 years. Fourteen males and ten females participated.

Average age of the female participants was 26.2 years, average height was 166.8cm and average weight was 65.2kg. Average age of the male participants was 30.6 years, average height was 179.5cm and average weight was 77.0kg.

The oldest, heaviest and tallest participants in the trial are males. The lightest, smallest and youngest are all female participants.

There is a correlation between height and weight and a weaker correlation between age and weight for female participants. for male participants, the only and strongest correlation is between height and weight.

Histogram
Correlation Matrix
Scatterplot

Exploratory analysis of data subjects (Shaar, 2019).

References:

A. Henriksen, M.H. Mikalsen, A.Z. Woldaregay, M. Muzny, G. Hartvigsen, L.A. Hopstock, S. Grimsgaard, 2018. *Using Fitness Trackers and Smartwatches to Measure Physical Activity in Research: Analysis of Consumer Wrist-Worn Wearables (Journal of Medical Internet Research)* [ONLINE] Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5887043/ [Accessed 16 December 2021].

Y.C. Huang, 2019. *My analysis on motion sensor data (Kaggle)* [ONLINE] Available at: https://www.kaggle.com/teaprint/my-analysis-on-motion-sensor-data [Accessed 16 December 2021].

D. Liftoff, 2019. *What Does Your Smartphone Know About You? (Kaggle)* [ONLINE] Available at: https://www.kaggle.com/morrisb/what-does-your-smartphone-know-about-you [Accessed 16 December 2021].

Kar, S.S. & Ramalingam, A., 2013. Is 30 the magic number? Issues in sample size estimation. *National Journal of Community Medicine,* 4(1), p. 175.

M. Malekzadeh, 2019. *MotionSense Dataset: Smartphone Sensor Data (Kaggle)* [ONLINE] Available at: https://www.kaggle.com/malekzadeh/motionsense-dataset [Accessed 16 December 2021].

S. O'Dea, 2021. *Number of mobile phones per household in the United Kingdom (UK) in 2020 (Statista.com)* [ONLINE] Available at: https://www.statista.com/statistics/387184/number-of-mobile-phones-per-household-in-the-uk/ [Accessed 16 December 2021].

J.L. Reyes-Ortiz, D. Anguita, A. Ghio, L.Oneto, X. Parra, 2012. *Human Activity Recognition Using Smartphones Data Set (UCI Machine Learning Repository)* [ONLINE] Available at: https://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones [Accessed 16 December 2021].

R. T, 2019. *A simple Features DNN using TensorFlow (Kaggle)* [ONLINE] Available at: https://www.kaggle.com/talmanr/a-simple-features-dnn-using-tensorflow [Accessed 12 January 2022].

S.T.A. Shaar, 2019. *Starter: MotionSense Dataset : 8c09b08d-5 (Kaggle)* [ONLINE] Available at: https://www.kaggle.com/salahuddinemr/starter-motionsense-dataset-8c09b08d-5 [Accessed 16 December 2021].

M. Shabaan, K. Arshad, M. Yaqub, F. Jinchao, M.S. Zia, G.R. Boja, M. Iftikhar, U. Ghani, L.S. Ambati, R.Munir, 2020. *Survey: smartphone-based assessment of cardiovascular diseases using ECG and PPG analysis (BMC Medical Informatics and Decision Making)* [ONLINE] Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7392662/ [Accessed 16 December 2021].

X. Sun, 2019. *Human Activity Recognition Using Smartphones Sensor Data (medium.com)* [ONLINE] Available at: https://medium.com/@xiaoshansun/human-activity-recognition-using-smartphones-sensor-data-fd1af142cc81 [Accessed 16 December 2021].

G.Wang, Q. Li, L. Wang, W. Wang, M. Wu, T. Liu, 2018. *Impact of Sliding Window Length in Indoor Human Motion Modes and Pose Pattern Recognition Based on Smartphone Sensors (Sensors – Basel, Switzerland)* [ONLINE] Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6021910/ [Accessed 16 December 2021].