Human Activity and Personal Attribute Recognition from Smartphone-based Sensor Data

Aditi Sony (200757766), Msc. Data Science, Newcastle University, Jan 2022

Abstract

With easy availability of data from day-to-day life, it is now possible to gain more in-depth insights into individual's life. Ability to infer certain attributes like BMI or any deviations from normal behaviour can help in identifying other health related issues or certain underlying causes early and thus can be used to improve one's health.

In this project main motivation is to find out if it is possible to predict the personal attributes of the subjects, alongside classifying the activity from the available smartphone-based sensor data.

Various classifiers using both traditional machine learning algorithms as well as deep Learning techniques were built and fitted on the data and prediction was done on test data. Performance was measured using accuracy as metrics. Some feature selection and engineering were employed to improve the performance.

Classification of Activity was achieved with high accuracy (93%), however prediction of Gender (62%), Age (51%) or BMI (78%) from the same input data (smartphone sensor-based data) could not be done with similar accuracy (Figure-1). Nonetheless, work showcased that it is possible and with some further improvements using domain knowledge and hyperparameter tuning, performance of those predictive models could also be improved.

Some of the available literature have similar work but for different purpose (e.g., for person identification or health assessments) using different approach (mostly by employing feature extraction using statistical properties) and for different set of attributes however none to the knowledge have included all of the personal attributes as focused in this work.

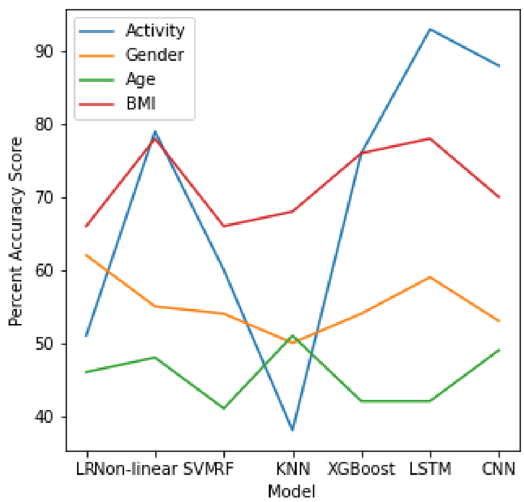

Figure-1: Comparison of performance of different models for various classification tasks

1. Introduction

Smartphones have become must have necessity these days and with time they have evolved from just being means of communication to one-stop-shop gadget with wide variety of features and applications. These days most of the smartphones have built-in Accelerometer, Gyroscope, Magnetometer and lots of applications use these to monitor person's activities.

Recently there has been an increasing trend for use of such applications for monitoring one's health or fitness. These provide many benefits from monitoring the fitness levels to providing intervention or guidance at the correct time for improving one's health. Different type of machine learning algorithms have been used in prediction or classification of activities based on the sensor data from the smartphones. My motivation here is to check if alongside recognising the activity, is it also possible to predict some of the personal attributes (e.g., height, weight, age, gender) of the individual performing those activities from the same sensor data collected from the smartphone. If it is, then maybe these can help in understanding other underlying health conditions like high BMI levels or other movement disorders early.

First any available related literature will be discussed, next dataset and methodology used in this project will be discussed, followed by results and analysis leading to the last section conclusions, any future implications or improvements and critical reflection.

2. Related Work

There is numerous literature available on work done in the Human Activity Recognition using sensor-based data. There is also some literature available for work done on identification of personal attributes from the smartphone-based sensor data [2].

However, most of the available literature [5] is based on identification of subject (or person identification) and involve statistical feature engineering techniques which require strong domain knowledge.

One of the closely resembling work done by Ibrar, K. et al. [4], involves identification of personal attributes like age, gender and weight by first identifying the subject. This was done by feature extraction using hand-crafting methods based on statistical summary from the raw data. Also, data included was for only walking activity.

Another closely related work done by Yao,Y. et al.[3], on prediction of Body Mass Index of smartphone users involves use of Motion-entropy based filtering strategy to select parts of sensor data that met certain thresholds for training the model. Then Conv-LSTM hybrid model is used for prediction.

3. Methodology

3.1 Dataset used

Dataset used is MotionSense data which is part of research work done by Malekzadeh et al. [1] from Queen Mary University of London. It is a time-series data generated by accelerometer and gyroscope sensors (attitude, gravity, userAcceleration, and rotationRate). It is collected with an iPhone 6s kept in the participant's front pocket. A total of 24 participants in a range of gender, age, weight, and height performed 6 activities in 15 trials in the same environment and conditions: downstairs, upstairs, walking, jogging, sitting, and standing.

3.2 Data Pre-processing

Raw data was organised per activity & trial and per subject. First, a labelled file was prepared from the raw data. Activities were labelled as ranging from 0 to 5 for ease of classification later.

This data was then explored for any class imbalance, and it was observed that data was imbalanced for each activity type. There was more data available for certain activities like sitting, standing and walking than others. This was balanced by under-sampling of the majority classes before use in any of the models. Since there was enough data for both genders and was not too much imbalanced it was not balanced before use. Also, it was noted that subjects have age ranging from 18 to 46 however most subjects were in the age range of 24-33. Subjects have height ranging from 161cm - 190cm with median height 175cms. However, two separate peaks were visible at 165 cm and at 180cm indicating bi-modality. Weight range of subjects was 48-102kgs with mean and median value of 72kgs.

Train-Test Split: Data from subjects 1-20 for all activities was kept as training data and data for subjects 21-24 were left for test purposes. This was done this way and not by random sampling to maintain the temporal nature of the data. Also, it was important to keep data for each subject for all activities together so as to allow models to learn the relationships between input sensor readings with subject's personal attributes. This is common approach for splitting of such data in academic literature.

Data was normalised by subtracting the mean and then dividing by standard deviation of all independent variables from training set and applying it to both training and test sets. This is very helpful in bringing all input variables to similar scale and not let any of the variables dominate due to its scale.

Sliding window approach: Sliding window with overlap approach has been used for modelling this time-series activity data. Fixed window length of 100 (meaning subset of 100 rows for each window) with 50% overlap has been selected to construct the window frames. According to available literature this is the most common approach for analysing such data.

Sampling rate of the data is not known (it maybe 100Hz or 50Hz). With sampling rate of 100Hz it'll amount to window length of 1 sec and with 50Hz to 2 secs, but in any case, window length of such timeframes is considered suitable for such analysis work in the available academic literature. Overlap of 50% will help in reducing the information loss and is also well-known industry practice for such time-series analysis.

Data was reshaped into window frames and maximum count of the dependent variable was taken from each window. Also, data was standardised after this to ensure zero mean and unit variance across all input variables.

3.3 Models and Evaluation metrics used

Several models using different techniques (for e.g., tree-based ensembling, bagging, boosting etc) including- Logistic Regression, Non-linear Support Vector Machine (RBF kernel), XGBoost, Random Forest, K-Nearest Neighbours were tried for all the classification tasks and performance was evaluated using accuracy as metrics. Main motivation of using various models was to explore different techniques and their impact on prediction performance. Accuracy score is considered a good measure of overall performance of a model and is also easy to compare between different models. Though, some other metrics could have also been used like F1 score or precision, recall etc however, due to time-constraints and also for ease of understanding only one metrics was being used throughout. Only exception was for any regression tasks where evaluation metrics used was Mean-Squared Error as accuracy is only suitable for classification purpose.

3.3.1 Classification of activity

There were 6 classes of activities and hence it was a multi-class classification task. Initially different models as mentioned above were built using all variables. However, the accuracy score obtained was very poor on the test data.

Next, models with reduced number of features (keeping only UserAcceleration, Rotation rate and Attitude data) were tried. Gravity data was dropped as it was believed that it was not adding any useful information since user imparted acceleration was already present from accelerometer readings. The accuracy score improved quite a bit.

Next, hyperparameter tuning of SVM was performed (since it was the top scoring model for prediction accuracy), and best set of parameters were searched using GridSearchCV package. The hyper-parameters selected for tuning purpose were C values ranging from 0.1 to 100, gamma values ranging from 0.001 to 1 and different kernels like rbf, poly and sigmoid. However, this did not help in improving the prediction accuracy any further.

Then, Deep Learning models including CNN and LSTM were tried on data with reduced features. Initially only basic models with few layers and nodes were tried. Later, hyperparameter tuning was performed manually (due to computational constraints in using GridSearchCV) by changing different parameters for e.g., number of layers, number of nodes in each layer, batch size, number of epochs etc. After some tuning, final model was fitted on training set, prediction was done on the test set and prediction accuracy score was evaluated.

3.3.2 Classification of gender

There were 2 classes hence it was a binary classification task. Again, reduced dataset with limited features was used and all the models were fitted on training set. Prediction was done on the test set. However, prediction accuracy was low.

Hyperparameter tuning for Logistic Regression (as it was the top scoring model with highest test accuracy amongst all the models) of some of the parameters like penalty (L1 and L2) and C values ranging from 0.1 to 100 were tried using GridSearchCV. Solver was fixed to be as Liblinear since it was noted that it converged faster whilst default Lbfgs solver was taking very long to converge (even with high max iteration values of upto 1000). However, this did not help in improving the prediction accuracy any further.

Then LSTM model was built using similar architecture for final model from previous section (but changing the activation function to sigmoid and loss to cross-entropy as it is a binary classification task) and both training and test accuracy was evaluated. It was observed that model was overfitting with higher training accuracy scores but very low test accuracy. Hence, tried reducing the complexity (reducing number of layers, number of nodes and number of epochs), but that didn't help with improving the prediction accuracy score or reducing the overfitting issue.

Next, CNN model with similar architecture for final model from previous section (but changing the activation function to sigmoid and loss to cross-entropy as it is a binary classification task) was fitted on training set and prediction done on test set. But similar issue of overfitting and low test accuracy was observed. Again, reducing complexity by reducing number of layers, nodes, epochs, batch size etc was tried but no improvements in prediction accuracy or reduction in overfitting was observed.

Next, it was considered that maybe data from certain activities like sitting & standing was not adding much useful information in identification of gender of subject performing those and reducing such data may help with reducing complexity and improving performance. Hence some of the models (Logistic Regression and Random Forest) were tried on dataset without those activities, however it was observed that this further reduced the prediction accuracy score and hence it was not considered as useful step and no other models were tried on it further.

3.3.3 Classification of Age

Age was converted into 4 categories namely- less than 20, 20-30, 30-40 and over 40.
Thus, continuous values were converted to categorical classes ranging from 0 to 3 respectively.
Classifiers were then built and fitted on training set. Prediction on test set was done and evaluated for accuracy score. Accuracy was low for all models.
Once again CNN and LSTM were also tried but prediction accuracy score remained low and overfitting issue was also present.
On checking for class balance, it was noted that there was high imbalance in various classes as there were very few subjects in age group less than 20 or over 40 than other categories. Since there was only 1 subject for class 0, random oversampling was not possible. Also, undersampling of majority classes was not considered as that would lead to too much information loss. Hence, SMOTE technique was used to create some synthetic samples representing minority classes. After that the data was split again into train and test sets and some of the models (Logistic Regression, Random Forest and KNN) were applied to the balanced dataset, however the accuracy score decreased drastically. Not sure why accuracy went low on balanced dataset, but it was not considered useful step in this case and hence no other models were then tried further.

3.3.4 Prediction of Weight and Height

Initially regression models were tried for both prediction of weight and height individually. Models tried were Support Vector Regressor and LSTM and evaluation metrics used was “Mean Squared Error”.

3.3.5 Classification of Body Mass Index

It was considered that BMI would be a better measure from weight and height of subjects, hence BMI was calculated using the formula- weight (in kg) / (height (in m))^2. Next it was divided into categories as below-

BMI Range	Class
Less than 18.5	0
18.5 - 24.9	1
25 -29.9	2
Over 30	3

Also, it was believed that having gravity data would be useful here as there is well-known correlation between weight and gravity. Various classifiers were then built and fitted on training set. Prediction of BMI category on test set was evaluated with accuracy score.

Next to compare if inclusion of gravity data made any difference to predictive performance, all classifiers were then fitted onto data without gravity data and prediction was done on test data and evaluated using accuracy score.

4. Results and Analysis

4.1 Classification of activity

Prediction accuracy on test data for various models is as shown in table-1 below. Removing gravity data certainly helped with improving the performance of models. Non-linear SVM and XGBoost were best amongst the traditional machine learning algorithms with prediction accuracy score of more than 75%.

Deep learning models (with some hyperparameter tuning) were better in performance with accuracy scores around 90%. For CNN model, it was observed that increasing the number of layers was detrimental to prediction accuracy score whilst increase in number of nodes in each layer, number of epochs, batch size and kernel size helped in improving it.

Overall, LSTM was the one with highest test accuracy score of 93%.

Table-1: Prediction Accuracy Score for Classification of Activity

Model	Accuracy Score with all variables	Accuracy Score with reduced variables
Logistic Regression	0.52	0.51
Non-linear SVM	0.68	0.79
Random Forest	0.56	0.60
XGBoost	0.78	0.76
KNN	0.41	0.38
CNN	N/A	0.88
LSTM	N/A	0.93

4.2 Classification of gender

Prediction accuracy for various models is as shown in Table-2 below. It has been low for all the models. Hyper-parameter tuning didn't help in improving the performance, nor did feature engineering by reducing data for certain activities like sitting and standing.

Deep learning models were having overfitting issues and were not generalising well, with high training accuracy but low test accuracy. Reducing the complexity of models including number of layers, epochs, nodes or batch size didn't helped in reducing the overfitting issues.

Overall, Logistic Regression model on data for all activities was found to be the one with highest prediction accuracy score of 62%. This maybe due to not much difference between the genders in performing the activities.

Table-2: Prediction Accuracy Score for Classification of Gender

Model	Accuracy Score (with sitting & standing data)	Accuracy Score (without sitting & standing data)
Logistic Regression	0.62	0.52
Non-linear SVM	0.55	N/A
Random Forest	0.54	0.54
KNN	0.50	N/A
XGBoost	0.54	N/A
LSTM	0.59	N/A
CNN	0.53	N/A

4.3 Classification of Age

Prediction accuracy for various models is as shown in Table-3 below. All models including deep learning models have very low prediction accuracy score. Deep Learning models were showing overfitting issues which didn't improved with reduction in complexity of models.

Creating synthetic data for minority classes to balance the data for each class using SMOTE technique further decreased the prediction accuracy score and performance of models drastically. Reason is not known for such a massive negative impact on predictive performance by this technique.

Overall, only 51% test accuracy score has been possible to be achieved using KNN for classification of age. It maybe due to limited training data available for certain age groups, and SMOTE technique might not have been able to generate representative synthetic samples.

Table-3: Prediction Accuracy Score for Classification of Age

Model	Accuracy Score (without SMOTE)	Accuracy Score (with SMOTE)
Logistic Regression	0.46	0.06
Non-linear SVM	0.48	N/A
Random Forest	0.41	0.08
KNN	0.51	0.11
XGBoost	0.42	N/A
LSTM	0.42	N/A
CNN	0.49	N/A

4.4 Prediction of Weight and Height

Prediction of subject's weight and height were first tried using some of the regressors like non-linear SVR and LSTM by fitting onto training set and prediction performed on test set and performance evaluated using Mean-squared Error. MSE obtained for those models are as shown in Table-4 and Table-5 below.

The MSE score was much higher for LSTM compared to non-linear SVR in both cases. However, prediction of a continuous value of weight and height doesn't seem to be appropriate approach instead categorising into BMI groups was considered to be a better alternative.

Table-4: Prediction Accuracy Score for Prediction of Weight

Model	MSE Score
Non-linear SVR	340.00
LSTM	5545.88

Table-5: Prediction Accuracy Score for Prediction of Height

Model	MSE Score
Non-linear SVR	92.58
LSTM	30012.09

4.5 Classification of BMI

Various classifiers tried gave prediction accuracy score as shown in Table-6 below. Inclusion or removal of gravity columns didn't make much difference in predictive performance. Most of the models were able to give accuracy score more than 70%, models with bagging (non-linear SVM) and boosting (XGBoost) algorithm were also giving equivalent performance as deep learning model (LSTM).

Table-6: Prediction Accuracy Score for Classification of BMI

Model	Accuracy Score(with gravity)	Accuracy Score(without gravity)
Logistic Regression	0.66	0.67
Non-linear SVM	0.78	0.75
Random Forest	0.66	0.74
KNN	0.68	0.66
XGBoost	0.76	0.71
LSTM	0.78	0.71
CNN	0.70	0.72

5. Conclusions

Classification of Activity was achieved with high accuracy (93%) using deep Learning model LSTM, however prediction of Gender, Age or BMI from the same input data (smartphone sensor-based data) could not be done with similar accuracy. For classification of gender, Logistic Regression model was having highest accuracy score of 62%, for classification of age it was KNN with 51% and for BMI both non-linear SVM and LSTM achieved similar prediction accuracy score of 78%.

The work completed shows it is possible to predict subject's personal attributes alongside the activity being performed from smartphone-based sensor data and can in future be used for many purposes like monitoring of BMI levels and providing correct guidance on improving fitness or overall health or identification of significant deviation from the normal behaviour typical of one's age or gender in performing activities potentially helpful for early intervention or diagnosis.

Limitations, Further Improvements and Future Implications: Further improvements to the work done could be achieved by performing feature extraction or engineering with domain expertise and by further tuning of different hyper-parameters or using different evaluation metrics which were not performed due to limited domain knowledge and time and computing power constraints.

Also, correlation between various attributes have not been studied and multilabel multioutput classification has not been explored due to lack of understanding of evaluation metrics suitable for such classification task as well time constraints, however it may be useful to understand and more suitable approach for any future work.

Critical Reflection: Working upon this project has been a steep learning curve for me with introduction to the world of machine learning and deep learning to working upon sensor-based time-series data to learning a new programming language and its libraries. Even though challenging, I have enjoyed developing myself with every problem I faced during the coursework. There were occasions when I found myself stuck onto various issues like dealing with imbalanced dataset to repeatedly getting low predictive performance of models. However, with help from numerous online resources, colleagues and teaching team, I kept building upon my understanding and knowledge on the topic. It gave me motivation to work hard and find solutions and also gave me skillset for problem solving and collaborating with others. In future I would like to try various different techniques as mentioned in the limitations section above for any such tasks.

6. References

1. Malekzadeh, M., Clegg, R. G., Cavallaro, A., & Haddadi, H. (2018, April). Protecting sensory data against sensitive inferences. In Proceedings of the 1st Workshop on Privacy by Design in Distributed Systems (p. 2). ACM.
2. Weiss, G. M., & Lockhart, J. W. (2011, August). Identifying user traits by mining smartphone accelerometer data. In Proceedings of the fifth international workshop on knowledge discovery from sensor data (pp. 61-69). ACM.
3. Yao, Y., Song, L., & Ye, J. (2020). Motion-to-BMI: Using Motion Sensors to Predict the Body Mass Index of Smartphone Users. Sensors (Basel, Switzerland), 20(4), 1134.
4. Ibrar, K., Azam, M., Ehatisham-ul-Haq, M., (2020, May). Personal Attributes Identification based on Gait Recognition using Smart Phone Sensors. In 12th International Conference on Computer and Automation Engineering (ICCAE 2020).
5. Singha, T.B., Nath, R.K., & Narsimhadhan, A.V. (2017, Nov.). Person Recognition using Smartphones' Accelerometer Data. arXiv preprint arXiv:1711.04689.
6. Lecture material and practical provided by Yu Guan on URL- yuguan1/CSC8111_code (github.com).
7. URL- www.machinelearningmastery.com
8. URL- www.stackoverflow.com
9. URL- www.towardsdatascience.com
10. URL- www.geegksforgeeks.com