Using Body Motion and Vital Signs to predict whether a human is Moving or Stationary

Sharhad Bashar[1]

***Abstract*— Detecting weather a person is Stationary or in Motion using sensor data is an important task with a wide range of applications. This paper focuses on designing a model that can accurately classify if a person is moving or still based on sensor readings from ten subjects performing various activity. The dataset and problem requirements posed several open ended questions, which is addressed in this paper. The paper explores and compares various Classification models as well as a custom dense neural network that are suited for the problem. Exploratory Data Analysis is also performed on the dataset to visualize and better understand the dataset, and identify the features pertinent to problem at hand. Key choices were made to better identify if a person is moving or not, and the justification for those choices are discussed below. Finally business applications of the results are also discussed.**

The source code is publicly available at `https://github.com/SharhadBashar/Health-Rhythms`.

I. INTRODUCTION

Detecting if a person is stationary, or moving is a very simple, yet important task. It has a wide range of applications in numerous settings. It can be used in a hospital for at risk patients, to make sure they are conscious, and dispatch medical help if not. Or it can be used by a Health App to determine how active a person is. Using sensory data for a persons motion and vitals can give us an effective dataset to classify their activities. This classification task can be approached several different ways using many different Machine Learning techniques.

Motion and Stationary can have different definition based on the use case. If its a user wanting to know how active they are during a time period, then activities such as sitting, lying, standing, crouching or moving their arms while staying in one place can be classified as being stationary, while activities such as walking, jumping or running are considered being in motion. However in a health care situation such as patient monitoring in a hospital or old home where it is critical to know if a person is conscious or not will require different definitions. For this case, the activities that were classified as being stationary in the previous use case can be considered as movement, ensuring that the user is conscious and dont require any immediate medical attention.

Various sensory data monitoring a persons movement and vitals can be used to collect data to help create a dataset to learn and answer the question is a person stationary or moving. For the purposes of this paper, we focus on the Mobile Health dataset from UCI Machine learning repository []. The Mobile HEALTH dataset comprises body motion and vital signs recordings for ten volunteers of diverse profile while performing several physical activities. Sensors placed on the subject's chest, right wrist and left ankle are used to measure the motion experienced by diverse body parts, namely, acceleration, rate of turn and magnetic field orientation. The sensor positioned on the chest also provides 2-lead ECG measurements, measuring a persons vitals while these activities are performed. Exploratory data analysis is performed in the next section to visualize the data from the sensors for each subject, activity and classification label, and to determine which features are important in the classification task.

This paper proposes a model designed to be used in a health care setting for tasks such as remote patient monitoring. Keeping this business application in mind, several key decisions had to be made, which are discussed in the Rationale section.

Various classification models were tested on the dataset to determine which model performs best. Models from Scikit Learn's library are used as well as custom neural networks. A Dummy Classifier is also used to ensure model performance. Section IV outlines these in further details. Followed by a discussions of these results and further research suggestions.

II. Exploratory Data Analysis

The collected dataset comprises body motion and vital signs recordings for ten volunteers of diverse profile while performing 12 physical activities. Activities 1, 2 and 3 are classified as Stationary, while the rest are Movement. The activities are:

1) Standing still (1 min)
2) Sitting and relaxing (1 min)
3) Lying down (1 min)
4) Walking (1 min)
5) Climbing stairs (1 min)
6) Waist bends forward (20x)
7) Frontal elevation of arms (20x)
8) Knees bending (crouching) (20x)
9) Cycling (1 min)
10) Jogging (1 min)
11) Running (1 min)
12) Jump front and back (20x)

Sensors were respectively placed on the subject's chest, right wrist and left ankle. The use of multiple sensors allows measurement of motion experienced by diverse body parts, thus better capturing body dynamics. The sensor positioned on the chest also provides 2-lead ECG measurements for monitoring users vitals. All sensing modalities are recorded

[1] Sharhad is a Masters student in Math and Computer Science, University of Waterloo Email: sharhad.bashar@uwaterloo.ca

at a sampling rate of 50 Hz. The activities were collected in an out-of-lab environment with no constraints on the way these must be executed, with the exception that the subject should try their best when executing them.

The dataset after combining all the instances from all 10 users has over 1.25 million rows as 23 features, from the sensors. Out of these, there are over 800,000 null activities (0 label) which represents no particular activity is recorded.

The following bar graph, figure 1 shows the distribution of the entire dataset into the three labels: Stationary (activity 1, 2, 3), Movement(activity 4 - 12) and Unknown (activity 0). As figure 1 shows, majority of the data is identified as Null. These instances are ignored. It also shows how uneven the data is, and will thus require oversampling. more on it in the following sections.

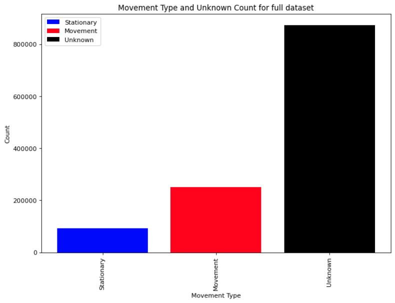

Fig. 1: Dataset label distribution

Figure 2 shows the distribution of the activities (except 0) for all the ten subjects. All the activities have equal number of readings, except for the final activity of jumping back and forth. Blue represents activities that are considered stationary, while red are activities for movement. Please refer to the Health Rhythms EDA notebook for activity distribution for each subject individually.

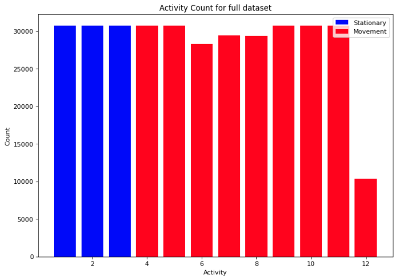

Fig. 2: Activities distribution for all subject

To further dive into the data, we perform analysis on the individual sensors themselves. Figure 3 shows the individual sensors data for Activity 1. Each sensor has a x, y, and z values (except ECG), which were used to plot the data. The data shows how different sensors report data for activities that are associated to stationary and activities associated to movement. Once more, blue is for stationary, red is for movement. The rest of the activities as well as 0 are all reported in the Health Rhythms EDA notebook.

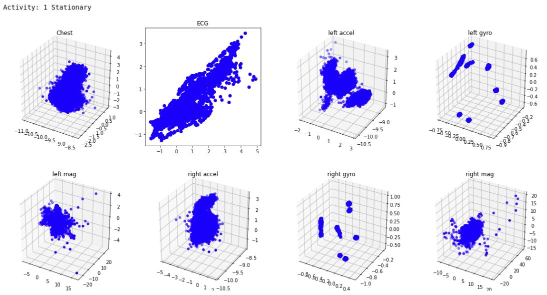

Fig. 3: Individual sensor data for Activity 1

Next we perform a few different analysis to better understand the data for learning purposes. We first create box plots for every sensor to identify outliers in the dataset. There are 23 different plots, one for each feature. Each plot contains data for all the 12 activities. Data for each activity for each feature are plotted in a box or as a individual point. The individual points corresponds to outliers in the data. Some features have more outliers than other. Figure 4 shows the plots for 4 features: ECG2, left accel x, y and z. As we can see from the plots, ECG2 has a lot more outliers than the other 3. To see all the outliers of all the features, please refer to the Health Rhythms EDA notebook.

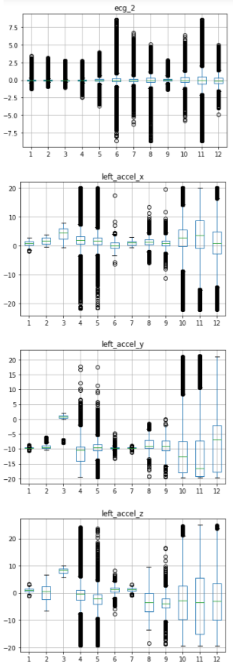

Fig. 4: Box plots to find outliers in data

Next we perform tests to determine the normality of the dataset as well as determine the importance of the features. Please refer to the end of Health Rhythms EDA notebook for the normality plots. Figure 5 shows the importance of features for the activities and for the labels of Stationary and Movement. As we can see from the plots, the data from the Left and Right gyros are the most important. This corresponds to the task, since we are trying to determine if a person is stationary or moving, and gyro data is very important in answering that question. Chest Acceleration data also plays an important role in this classification task, and is reflected as such in the feature importance plots.

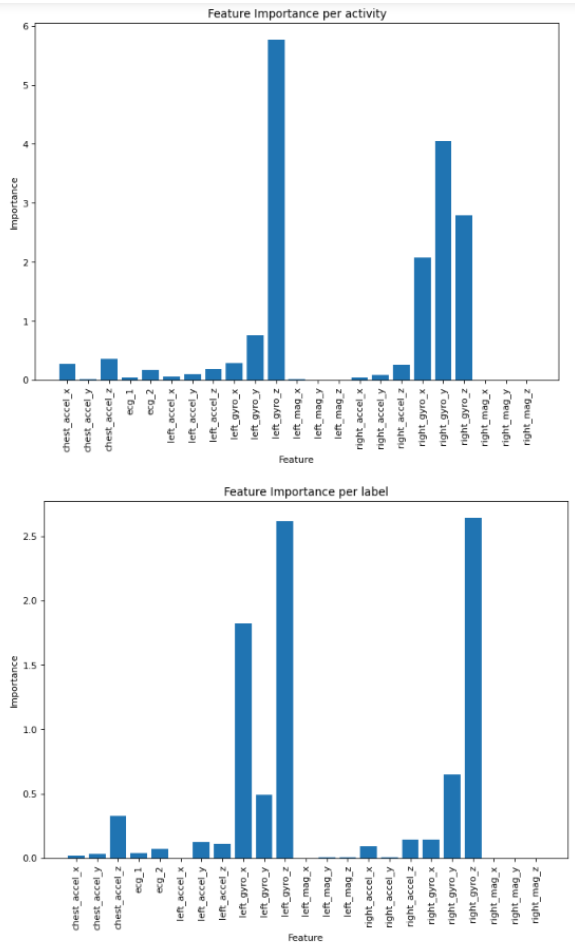

Fig. 5: Feature importance for each activity and for each label

PCA, tSNE and kMeans clustering was also performed to determine which labels to use, and which labels better relate to Stationary and Movement. More on this in the next section.

III. Rationale

For deciding which labels to keep and which to discard, as well as which activity corresponds to Stationary and which to Movement, we performed PCA, kMeans and tSNE analysis. The full code and better plots can be found in the Health Rhythms EDA notebook. The results are shown below:

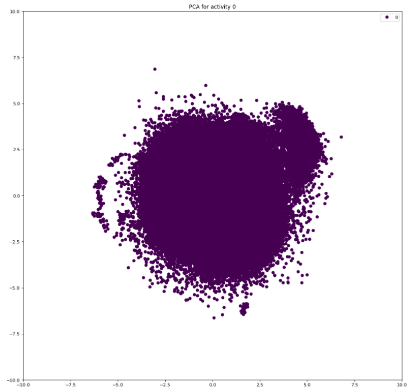

Fig. 6: PCA for activity 0

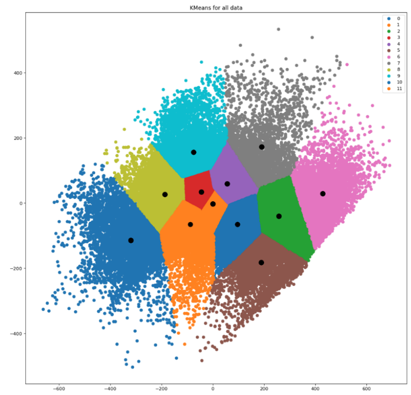

Fig. 7: KMeans for all the data as well as the cluster center

First we needed to consider if activity 0, which corresponds to no activity should be taken into the dataset for learning. It was decided to not take labels 0 into consideration, since we are unsure what type of activity it relates to. We need to do further analysis on each of the sensors data for activity 0 to see if it closely relates to the sensor data for another activity, and then classify it accordingly. Furthermore, it also depends on the business case. If we are determining if a person is conscious or not, then all the activities (from 1 - 12) should be classified as movement, and label 0 should be classified as stationary, in which case, a medical personnel will be alerted. Or if the use case was for a activity detector that tells us what percentage of time a person is moving, then we can ignore label 0, as it does not specify the type of activity. At the same time, performing PCA analysis on label 0 (figure 6) and comparing them to PCA for other activities (figures 8 9 10 11). There is no clear correlation to one single type or activity. So without further analysis, or a definitive business case, we remove all instances with label 0 from the dataset.

Now to decide which activities correspond to movement and which to stationary. As mentioned above, this depends on the business case. the outcome of the classification will depend largely on the requirments and needs of the customer.

However, we assume the most basic case for our needs. Immediately we can tell that labels 1, 2 and 3, which corresponds to standing still, sitting and lying down respectively are considered stationary, while labels 4, 5, 9 - 12 correspond to some type of motion. Performing PCA analysis on these two groups support the outcome. For stationary activities in figure 9, we can see that the plot shows data in small clumps or several droplets. In contract, PCA for movement activities in figure 10 show data in one large group. Now, we need to separate the remaining labels, 6 - 8. These correspond to movement of body parts, but the body itself remains in one place. Once again, it depends

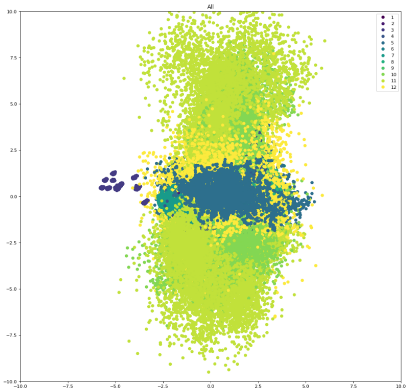

Fig. 8: PCA for activities 1 - 12

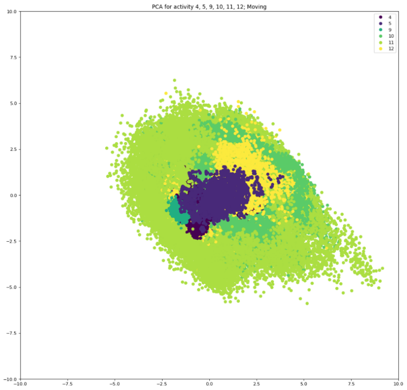

Fig. 10: PCA for activities 4, 5, 9 - 12. Movement Activities

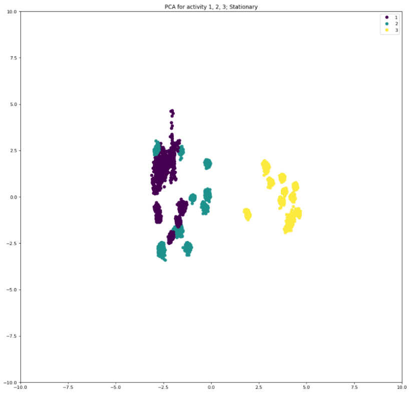

Fig. 9: PCA for labels 1, 2 and 3. Stationary Activities

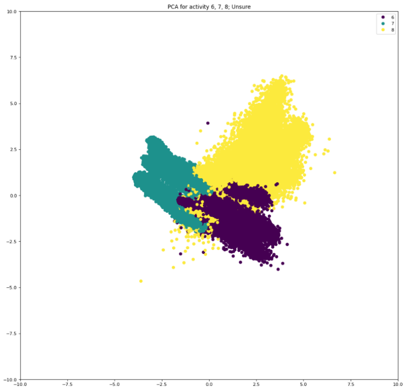

Fig. 11: PCA for activities 6, 7, 8. Unknown activities

heavily on the business case, but for our requirements, we choose follow the patterns established from PCA analysis. Performing PCA on these show that these data relate more towards the movement category than the stationary category. Thus they are labeled as such.

So in conclusion, labels 1, 2 and 3 are Stationary, and the rest are Movement. Label 0 is discarded from the dataset.

IV. Experiments: Model Selection and Training

In this section we describe our approach and algorithms and models we use for classifying activities as Stationary and Movement.

First we consider data imbalance. There is roughly 2.7 times more movement data than stationary data. However, since the real world data balance is unknown, we perform two different tests, one for balanced data, and another for unbalanced version. For the balanced trials, we oversample the minority class so that both classes have the same amount of training data. For unbalanced, we leave the data as is.

Following common Machine Learning practises, we split our dataset into 80% training, and 20& testing. The 20% testing dataset is kept untouched till the final evaluation of the model. For the training dataset, we use K fold testing, where k is set to 5. This means the dataset is broken into 5 equal parts. 4 are used for training, and the 5th is used for model validation. This is done 5 times, and the average scores are reported.

In order to select the best model for the classification task, we choose a variety of models from Scikit Learn and pick the one with the best scores. The models used are:

1) Logistic Regression
2) Ridge Classifier
3) SGD Classifier
4) Random Forrest Classifier
5) Extra Trees Classifier
6) Ada Boost Classifier
7) Gradient Boosting Classifier
8) Linear Support Vector Classifier
9) Decision Tree
10) Bagging Classifier
11) K Nearest Neighbours
12) Gaussian Naive Bayes
13) Multi-layer Perceptron classifier
14) Complement Naive Bayes

We perform 5 K cross validation on all the models with the 80% training dataset. for each model, we calculate the following scores, in accordance to common ML practises (for both training and validation sets):

1) Accuracy
2) F1 Score
3) Precision
4) Recall
5) ROC AUC

After performing the tests, we pick the model with the best scores, which was Extra Trees Classifier. We also compare the model with a custom Neural Network called Sharhad_Net, which is a fully commected Neural Network with 2 hidden layers of 128 neurons each, input layer of 23 neurons (one for each feature) and output of 2 neurons, one for each class. We also test it against the Dummy Classifier from Scikit Learn to ensure our model actually learned, and that it does not just randomly make guesses.

We run the models on two settings: one with balanced dataset, and the second without data balancing. Note that the data is shuffled first before used for training the models. Complement Naive Bayes is a special model that is dedicated to unbalanced dataset, which was used to demonstrate model learning for special cases where the labels are well skewed towards one class. The code for this can be found in the Health Rhythms learning notebook.

V. Results

Overall, the performances of the models were close. The results and scores of the K fold cross validation can be found in the Health Rhythms learning notebook. Extra Trees Classifier had the best scores with 100% for all the evaluation parameters. This also performed well on the testing dataset as well, with only a few mislabeled data. In the notebook, please specify the model to be used for the task beside the comment (specify model to be used here).

Sharhad_Net a custom Neural Network was also ran in comparison, which got 94% on the test set. And finally comparing to the dummy classifier, which had an accuracy of 50%, both off the shelf and custom models performed better that simply guessing the outcome.

VI. Conclusion

This paper shows how to classify different activities as stationary or movement. We explare the data, define key choices for the classification task, and finally identify the best model for the task based on several scores and performances. We also show a custom Neural Network for the task and compare its performance with state of the art open source models.

VII. References

1) MHEALTH Dataset, https://archive.ics.uci.edu/ml/datasets/MHEALTH+Dataset 2014-12-07.
2) Scikit Learn Linear models, https://scikit-learn.org/stable/modules/linear_model.html