Student ID: S3508880

Student Name: Kyrollos Iskandar

I certify that this is all my own original work. If I took any parts from elsewhere, then they were non-essential parts of the assignment, and they are clearly attributed in my submission. I will show I agree to this honour code by typing "Yes": *Yes*.

Please start your report since here ….

k-nearest neighbour classifier for human activity recognition based on data from a chest-mounted accelerometer

Kyrollos Iskandar[a]

[a]124 La Trobe St, Melbourne VIC 3000, RMIT University

Email address: S3508880@student.rmit.edu.au

13 June 2020

<u>**Table of contents**</u>

Abstract 3

1. Introduction 3

2. Methodology 3

3. Results 4

3.1. Q1. Is there anything special with each participant? 4

3.2. Q2. Which acceleration vector component contributes the most to the acceleration magnitude? How much do the other two components contribute to it? 5

3.3. Q3. How do the distributions of acceleration vector components and acceleration magnitudes vary with Class Label? 5

3.3.1: Class label 1: Working at a computer 5

3.3.2: Class label 2: Standing up, walking and going up/down stairs 6

3.3.3: Class label 3: Standing 6

3.3.4: Class label 4: Walking 6

3.3.5: Class label 5: Going up/down stairs 6

3.3.6: Class label 6: Walking and talking with someone 6

3.3.7: Class label 7: Talking while standing 6

3.4: The k-nearest neighbour and Decision Tree classifiers 6

4. Discussion 7

5. Conclusion 7

References 7

Abstract

Recognition of human activities by computers can be useful for various machine learning applications. A machine learning algorithm may use only motion data, such as chest acceleration data, for recognising a number of common activities. It is the aim of this work to compare the k-nearest neighbour and Decision Tree classifiers in recognising the activities participants did based on their chest accelerations. Accelerometer data was obtained from Casale, Pujol [1], cleaned, explored and modelled. It was found that the activities were done in a controlled environment given their common order across participants. There were also some variations of acceleration vector components and magnitudes across participants for any given activity. It was found that the k-nearest neighbours classifier more accurately recognised activities by around 1.333% compared to the Decision Tree classifier when used as a 51-nearest neighbours classifier. Thus, the 51-nearest neighbours classifier was recommended as the classifier to use for activity recognition in this dataset out of the classifiers considered.

1. Introduction

The dataset consists of 15 groups of accelerometer data. Each group of data contains accelerations in the three Cartesian axes, with the x-, y- and z-axes pointing left, up and forwards, respectively, with respect to the participant's chest on which each accelerometer was mounted [2]. Each group of data also contains Class labels numbered from 1 to 7 that specify activities that participants did. These activities are specified by Casale, Pujol [1].

The aim of using the data is to make a model that recognises what activities participants while being informed of the acceleration vectors and acceleration magnitudes of their chests. It is the aim of this work to investigate relationships between features in the data, model the dataset using the k-nearest neighbour (KNN) and Decision Tree classifiers and to compare the accuracies of the classifiers to each other. Thus, the more accurate classifier is to be recommended for modelling the dataset.

2. Methodology

The data was obtained from Casale, Pujol [1] in 15 groups, that is, one group of data per participant, in the form of three orthogonal Cartesian accelerations and a Class label feature. The data was cleaned by checking for missing and impossible values and fixing them if any were found. There were no missing values, which was confirmed by checking the difference between how many rows columns had and how many of them were filled. Impossible values in the acceleration vector component features were checked by sorting their contents in ascending order. Any entry that had letters in them were expected to appear at the bottom of the sorted list. None were found, so the features were considered clean. However, there were some rows in the data that had a Class label of 0, which was not defined [1]. These rows were dropped from the data considering that the aim of using the data was to recognise what activities participants did based on their chests' acceleration vectors and magnitudes. That meant that the Class label was the most important feature in the data. For some cleaning of the data, the 15 groups of data were merged into one big group.

Using the information from Casale, Pujol [2] that the data was collected at a rate of 52 Hz, the indices were transformed into a Time feature, with each data row being $\frac{1}{52}$ seconds wide. This transformation was done for each participant separately because each participant did the activities in their own time.

The Time feature was not used in modelling the data. However, it was useful in understanding the accelerations of the participants.

Using scatter plots of the data, it was found that there was an average value in each acceleration vector component. Thus, for each participant, the average x, y and z accelerations were subtracted from every x, y and z acceleration, respectively. This was done to have an idea of accelerations in opposite directions per axis, that is, in both the positive and negative directions, rather than only in one direction per axis, that is, the positive direction. These subtractions were done considering that the average x, y and z accelerations had to be zero, lest the acceleration vector components say that the participants were eventually flying at high speed, which, given the Class labels [2] and Time feature, is impossible.

For each participant, the acceleration magnitudes A_{mag} were calculated according to Equation 1.

$$A_{mag} = \left(A_x^2 + A_y^2 + A_z^2\right)^{\frac{1}{2}}$$

(Equation 1)

where A_x, A_y and A_z are the x, y and z accelerations, respectively. Thus, the acceleration magnitudes were a fourth time series, with the acceleration vector components A_x, A_y and A_z being the original time series.

The data was plotted as scatter plots including the acceleration magnitudes to see how the features of the data related to each other. For each participant, the spread of each of the acceleration vector components and acceleration magnitudes were investigated though box plots. This was also done per Class label per participant.

The data was then modelled as one big group, and not 15 individual groups, using the k-nearest neighbour (KNN) and Decision Tree classifiers. The parameters of each classifier was manually tuned to maximise their accuracies. The classifiers were compared to each other in terms of classification accuracy and the most accurate one was recommended.

The following questions were explored in the data:

- Q1. Is there anything special with each participant?
- Q2. Which acceleration vector component contributes the most to the acceleration magnitude? How much do the other two components contribute to it?
- Q3. How do the distributions of acceleration vector components and acceleration magnitudes vary with Class Label?

For the Decision Tree classifier, the Gini index criterion was used.

3. Results

3.1. Q1. Is there anything special with each participant?

A few participants' accelerations were small (Figure 3.1.1). Others' accelerations were larger.

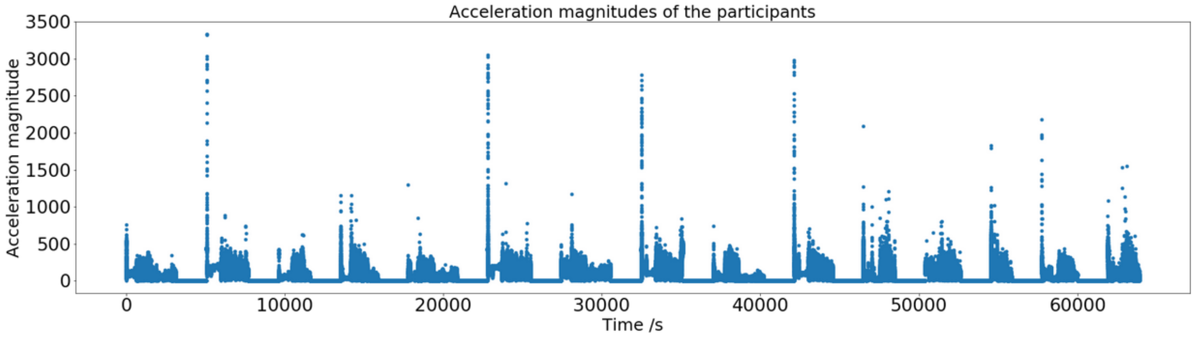

Figure 3.1.1: Acceleration magnitudes of the 15 participants. The horizontal lines of zero acceleration in between adjacent waveforms are not part of the acceleration magnitudes of the participants. They are there only to space the waveforms apart so that the acceleration magnitudes of each participant can be seen in isolation. The spacer width is 1923 seconds.

Activities were done in the same order across all participants (Figure 3.1.2). However, the time duration of each activity varied across participants. All activities were done in their own block of time except for standing, which was done in two or more blocks with some walking and going up or down stairs done in between.

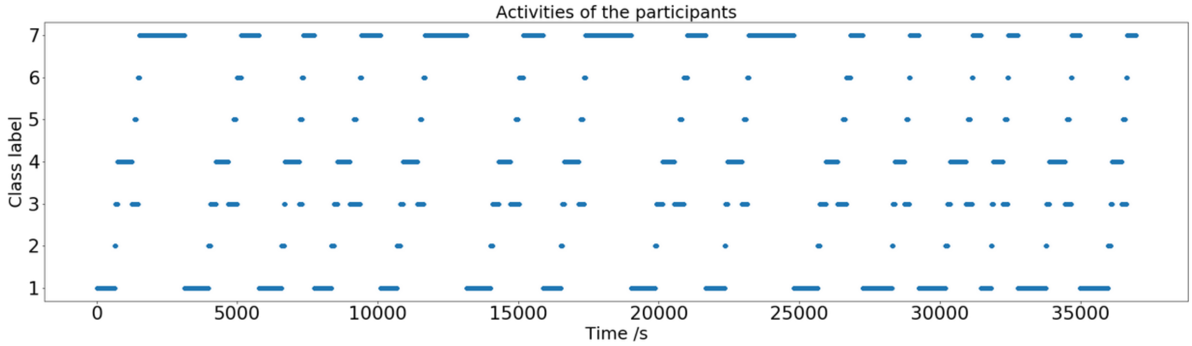

Figure 3.1.2: Activity order of all 15 participants. Participants' activity orders are placed one after the other. Each one starts at Class label 1 (working at a computer [1]). Note the variation in time duration of each activity.

3.2. Q2. Which acceleration vector component contributes the most to the acceleration magnitude? How much do the other two components contribute to it?

There seems to be the least spread in the y accelerations. They are quite packed at zero y acceleration. y accelerations were up/down. Participants had to be jumping or doing something like jumping for y accelerations to be large. The closest to jumping that the participants did was walking and going up/down stairs. Participant 7, who showed the largest range of acceleration magnitudes when they stood up, walked and went up/down stairs (Class label 2), had the largest interquartile range of y accelerations. Acceleration magnitudes are greater than or equal to 0 because they are magnitudes.

3.3. Q3. How do the distributions of acceleration vector components and acceleration magnitudes vary with Class Label?

3.3.1: Class label 1: Working at a computer

For Class label 1, working at a computer, acceleration vector components had very little spread, indicating that participants had a nearly static posture. The acceleration vector components seem to be gathered at zero. Any deviations from zero are probably due to participants leaning while sitting, assuming they were sitting. The y acceleration was usually at around zero. There were some deviations from zero. In most cases, the y acceleration was the least spread, focused at around zero units. In most cases, the x and z components were similar to each other in location and spread.

3.3.2: Class label 2: Standing up, walking and going up/down stairs

Body movement was involved in this activity. So, it is expected that acceleration vector components would deviate from zero. Participant 13 showed the largest spread in z acceleration while doing this activity. Perhaps they moved their chest forwards and backwards with much energy.

3.3.3: Class label 3: Standing

The components are very similar compared to when the participants were standing up, walking and going up/down the stairs (Class label 2). There seems to be more spread in the x (sideways) component.

3.3.4: Class label 4: Walking

This activity involved body movement. So it is expected that there be large spreads in the acceleration vector components. Participants 1 and 5 have a very large spread in y acceleration. Perhaps, when they were walking, they moved their chest up and down with much energy. The y component was more spread out compared to the y component in Class labels 2 and 3 in terms of interquartile range.

3.3.5: Class label 5: Going up/down stairs

Looks similar to Walking (Class label 4).

3.3.6: Class label 6: Walking and talking with someone

All components seem bunched up compared to in the other Class labels. Unlike the comparison between Standing and Talking while Standing, it seems that talking affected the accelerations of participants' chests more while they were walking. Perhaps body language was a factor in these variations.

3.3.7: Class label 7: Talking while standing

Some participants seemed to be leaning backwards, as shown by their negative z accelerations. The spreads in and locations of the acceleration vector components are almost identical to when the participants were merely standing (Class label 3). Any variations of the acceleration vector components from those when they merely stood may be due to them using body language while talking, which would involve some acceleration of their chests.

3.4: The k-nearest neighbour and Decision Tree classifiers

For the Minkowski metric with p = 1 and weighted distance measure, the k-nearest neighbour (KNN) classifier was found to give the highest train-test scores for the k-values that were tested when compared to the Decision Tree classifier. It was found that k = 51 in KNN gave the highest train-test score of 69.198% using a test size of 50%. However, the KNN classifier was most accurate compared to itself when the test size was only 10% of the total dataset, with a train-test score 70.123%. The KNN classifier was more accurate than the Decision Tree classifier by around 1.333%.

4. Discussion

The accelerometer is sensitive to gravitational acceleration, more specifically, its orientation with respect to the gravitational acceleration vector, which points straight down to the centre of the Earth [2]. When the participants worked at a computer (Class label 1), if there was a non-zero x-acceleration, then they were probably leaning to one side while sitting, assuming that, if they stood up, they would be standing up straight and not leaning to one side lest they may lose balance and large and sudden accelerations be seen. If there was a non-zero z acceleration, then they were probably leaning forwards or backwards.

Participants did one activity at a time. This is shown by the Class label vs. Time plots (Figure 3.1.2). The other activities involved walking and standing. So, it is expected that the x acceleration in those activities would oscillate around the zero average x acceleration. This appears to be what the data shows, except for Class label 7 (talking while standing), especially for Participant 12. Perhaps the participants, when they talked while standing, they used body language, which involved the movement of the body, probably including the chest on which the accelerometer was mounted. In general, it appears that the y accelerations oscillated a lot when participants were walking and going up/down stairs. This was the case probably because, when they were walking or going up/down stairs, their chest moved up and down as their legs and waist moved according to their walking pattern.

When they were merely standing (Class label 3), their y acceleration was nearly zero. For Participant 8, it is not known why their y acceleration was negative (-500 units) (data not shown) when they were talking while standing. Their x and z accelerations at those times were approximately centred around zero. Participant 2 seemed to oscillate their body less than the other participants. When some participants worked at a computer (Class label 1), they manifested the largest ranges of acceleration magnitudes. It is probable that they kept changing their posture while working at the computer, assuming they were sitting. For other participants, the largest range of acceleration magnitudes occurred when they went up/down stairs (Class label 5). For Participant 7, it was when they stood up, walked, and went up/down stairs (Class label 2), even though they did little of it in their total activity time. It is probable that they did this activity more vigorously compared to how they did other activities. All of the acceleration magnitudes, vs. Class labels, have points at around zero acceleration magnitude (data not shown).

Based on the higher accuracy of the KNN classifier compared to the Decision Tree classifier, it is recommended that the KNN classifier be used to classify the dataset.

5. Conclusion

It was found that the k-nearest neighbours (KNN) classifier was more accurate than the Decision Tree classifier in classifying the data by around 1.333% when using the acceleration vector components and magnitudes of the participants. Thus, it is recommended that the KNN classifier be used to classify the data.

References

- 1. Casale, P., O. Pujol, and P. Radeva. *Activity Recognition from Single Chest-Mounted Accelerometer Data Set*. 2014 [cited 2020 12 June 2020]; Available from: https://archive.ics.uci.edu/ml/datasets/Activity+Recognition+from+Single+Chest-Mounted+Accelerometer.
- 2. Casale, P., O. Pujol, and P. Radeva, *Human Activity Recognition from Accelerometer Data Using a Wearable Device*, in *IbPRIA 2011*. 2011, Springer-Verlag. p. 289-296.