Exploratory data analysis using MapReduce for human activity classification

Mehmet Ege OĞUZMAN
StudentId: 20257961
Scalable Systems Programming, MSc in Data Analytics
National College of Ireland Dublin, IRELAND
Email: x20257961@student.ncirl.ie. URL: www.ncirl.ie

***Abstract*—The use of wireless sensors to track physical activity can help determine a person's postural orientation and movements in the real world. In this study, by using a publicly available human activity recognition dataset named PAMAP2 in the UCI Machine Learning Repository, feature selection and dimensionality reduction will be applied for multiclass classification problem on human activities. In the study, Principal Component Analysis will be used for feature reduction and the data required for the implementation of PCA will be prepared on a distributed system, in a scalable manner, using the MapReduce approach. In addition, the insight required for exploratory data analysis and data cleaning will be provided with descriptive statistics prepared with the MapReduce approach. At the end of the study, one of the state-of-art machine learning models will be selected, tested on the transformed and the original data and the results will be compared. Another aim of the study is to determine the features that should be used in the PAMAP2 dataset with certain statistical methods, data analysis and machine learning algorithms in order to solve the human activity recognition problem.**

I. Introduction

Wearable technology has become increasingly popular in recent years. Wearable technology has been incorporated into health and wellness applications, such as self-management and self-care, to improve the health and well-being of users. Sensors such as accelerometers, gyroscopes, magnetometers, heart rate sensors, and other similar types of sensors are commonly found in these wearable devices. Therefore, wearable sensor-based human motion modeling is a growing field with numerous potential applications, including Human Activity Recognition, biometrics, and health. Human activity recognition has gotten a lot of attention lately because of the abundance of data and the variety of uses for which it can be put.

Recognizing people's activities is a relatively new area of study that has recently grown in importance to academics as well as business researchers. It's important to develop a human activity recognition system that can be used for a variety of purposes, including home-care, prisoner monitoring and physical therapy and rehabilitation, public safety and military use.

Signal data collected from wearable sensor technologies can be very large and high-dimensional. Therefore, in this study, data used for human activity recognition will be analyzed by applying data processing approaches on scalable and distributed systems, necessary dimensionality reductions will be applied and feature importances will be obtained, then the transformed data set will be tested with machine learning algorithms.

II. Literature Review

There are two broad categories of methods for identifying activities: sensor-based and vision-based. Sensor-based systems employ a wide range of sensors that are attached to the subject being monitored. Activity recognition systems that rely solely on images and video sequences are called "vision-based" systems. The inherent nature of both approaches poses challenges to each. When it comes to low-power wearable devices, sensors require classification algorithms to be as fast as possible, but accurate and reliable vision-based methods are still a challenge no matter the technology available today. Therefore, signal processing, feature extraction, and activity classification using machine learning algorithms are the most important aspects.

Since each sensor creates a different feature in the dataset, multidimensional datasets are formed in wearable data collection devices, which can slow down the classification algorithm to be used and reduce the accuracy rate. Therefore, methods such as feature selection or dimensionality reduction have been frequently applied in the literature [1]. For example, a correlation-based feature selection method was used in the study [2]. The correlation threshold was 0.25 and features with a correlation value of 0.25 or more were added to the training and test datasets. In another study [3], the sequential scatter search with a reduced greedy combination (SS-RGC) algorithm was used to find the most optimal feature subsets.

In another study [4], 4 different feature selection methods were tried using genetic algorithms. Attribute selection, then Dimension Selection After Feature Sensibility (Some-or-None) which is dimension selection with its selected attributes, Dimension Selection After Feature Sensibility (Take-It-All or Leave-It) which takes all the attributes of selected dimension and Dimensions Selection Without Feature Sensibility which is taking all relevant features. As a result of the study, the number

of features of the dataset was halved and 97.45% accuracy was obtained.

In the study [5], a comparative analysis was performed between classification machine learning models using 3 human activity recognition datasets (PAMAP2, mHealth and SWELL). In the study, it was concluded that tree-based models (Random Forest, XGBoost) performed better against models such as Naive Bayes and Logistic Regression.

Various multiclass classification methods are available in the literature to solve the human activity recognition problem. A number of cutting-edge machine learning methods were initially benchmarked in order to find the best classification approach, including binary decision trees, support vector machines, deep neural networks, random forests, and Adaboost.

III. Methodology

A. Dataset

The dataset used in this study was gathered from real-world applications and is critical for implementation. Trivisio, Germany-made IMUs (inertial measurement units) and a heart rate monitor were worn by nine healthy human subjects in the PAMAP2 (Physical Activity Monitoring for Aging People 2) dataset [6]. Each of the three IMUs has an accelerometer, gyroscope, and magnetometer to measure temperature and 3D data. A 2.4 GHz wireless network transmits the data at a rate of 100 Hz. One IMU was worn on each subject's dominant wrist, one on each subject's dominant ankle, and one on the subject's chest during the experiment. On all 9 test subjects, referred to as "subject101" through "subject109," these methods have been successfully applied. The PAMAP2 is a 54-attribute multivariate time series data set and data of all these subjects consists of 2872532 records. Dataset is publicly available in the UCI Machine Learning Repository [1].

Independent Features
- 1 Timestamp (s)
- 2 activityID (encoded activities)
- 3 heart-rate (bpm)
- 4-20 IMU hand
- 21-37 IMU chest
- 38-54 IMU ankle

Target Feature : activityID
- 1 lying
- 2 sitting
- 3 standing
- 4 walking
- 5 running
- 6 cycling
- 7 Nordic walking
- 9 watching TV
- 10 computer work
- 11 car driving
- 12 ascending stairs
- 13 descending stairs
- 16 vacuum cleaning
- 17 ironing
- 18 folding laundry
- 19 house cleaning
- 20 playing soccer
- 24 rope jumping
- 0 other (transient activities)

[1]Public PAMAP2 dataset is available at UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/pamap2+physical+activity+monitoring.

B. Model Overview

The study will be carried out in six different stages.

Independent Features
1) Exploratory Data Analysis
2) Data Cleaning
3) Principal Component Analysis
4) Dimensionality Reduction
5) Train Random Forest
6) Compare the results

According to the MapReduce paradigm, extracting descriptive statistics, correlation matrix and covariance matrix to apply PCA are completely parallelized. In order to run our MapReduce schema in a distributed fashion across a cluster of machines, Hadoop is the software infrastructure that takes the challenges of distributed computing into account. MapReduce is implemented using the Python library MRJob. Open source library MRJob can run on Hadoop and AWS EMR platforms, making it easier to write MapReduce code blocks.

Data sets provided by wearable sensor technologies with millions of rows may not be processed adequately by a single-core computer's processing power. Instead, these computations were carried out on a cloud-based virtual server using the MapReduce technique.

C. Exploratory Data Analysis

1) Descriptive Statistics with MapReduce

Descriptive statistics required for data analysis were obtained with scalable and distributed methods. Jobs developed using the Python MRJob library were developed to find values such as missing value counter, mean 1, median 2, min-max, kurtosis 6, skewness 5, variance 4 and standard deviation 3.

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} a_i = \frac{a_1 + a_2 + \cdots + a_n}{n} \tag{1}$$

$$odd - median = \frac{(N+1)}{(2)}, even - median = \frac{\frac{n}{2} + (\frac{n}{2} + 1)}{2} \tag{2}$$

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2} \tag{3}$$

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n} \tag{4}$$

$$skewness = \frac{1}{(N-1)(N-2)} \sum_{i=1}^{N} (\frac{x_i - \overline{x}}{S})^3 \tag{5}$$

$$kurtosis = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^4}{(\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2)^2} \tag{6}$$

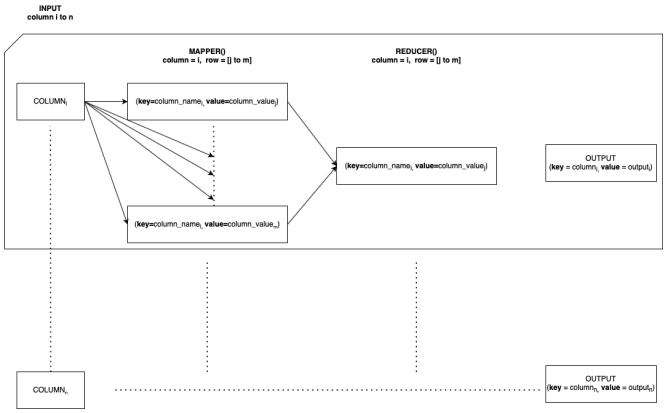

Fig. 1: Operation diagram of MapReduce jobs for calculating descriptive statistics.

Figure 1 describes how descriptive statistics work on distributed systems. The data loaded on the HDFS system. Then it maps each column as a key and its value as a row value with the help of the mapper function. The reducer function calculates the descriptive statistics formula for each mapped column one by one and outputs it.

2) Relationship Analysis Between Features

For the most part, our classes are balanced, as you can see in the chart below 2.

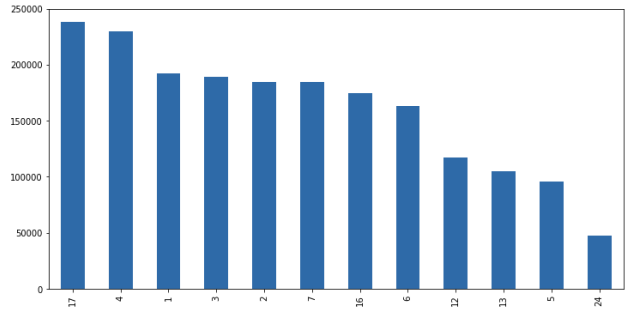

Fig. 2: Value counts of target feature ”activityID”.

When Figure 3 is examined, bar chart shows that ”rope jumping” and ”running” are the most time-consuming of the activities, since they activities with the highest heart rate.

The relationship between SubjectID and activityID was examined with the bar plot in Figure 4 and it was observed that the SubjectID property was not very discriminating. Therefore, Chi-square test was applied between activityID and SubjectID, and as a result of the test, it was revealed that there is a relationship between these two features.

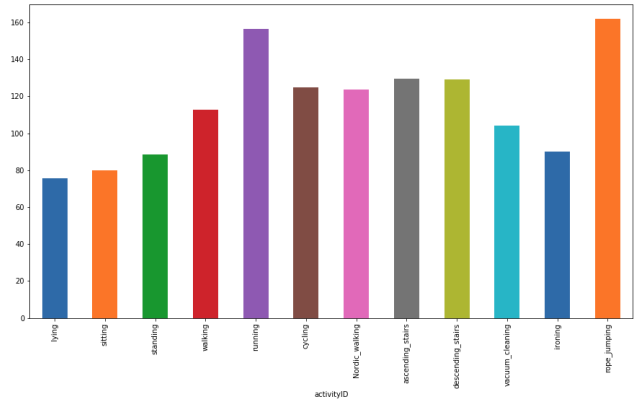

Fig. 3: Relationship between ”heart-rate” and ”activityID”.

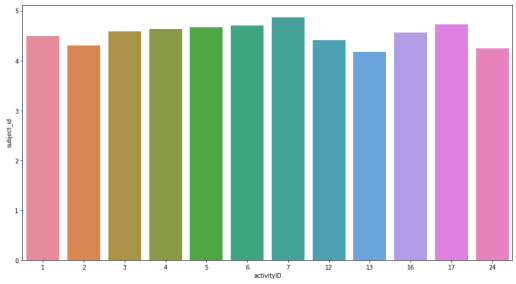

Fig. 4: Relationship between ”subjectID” and ”activityID”.

3) Correlation Matrix

The correlation matrix was obtained with the MapReduce structure seen in Figure 5.

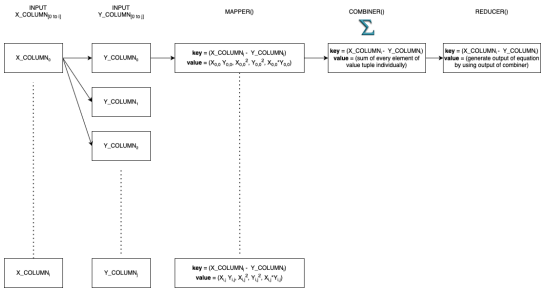

Fig. 5: MapReduce job structure to obtain correlation and covariance matrix.

$$r_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 (y_i - \overline{y})^2}} \tag{7}$$

For the MapReduce implementation of the correlation formula, each feature combination must be calculated. This is time consuming but a scalable solution for single core processors or for big data where computations cannot be done with pure Python or Pandas.

The heatmap (7) depicts the degree of statistical resemblance between the various columns in the dataset. All we have to do is look at the data and see that the gyroscopes do

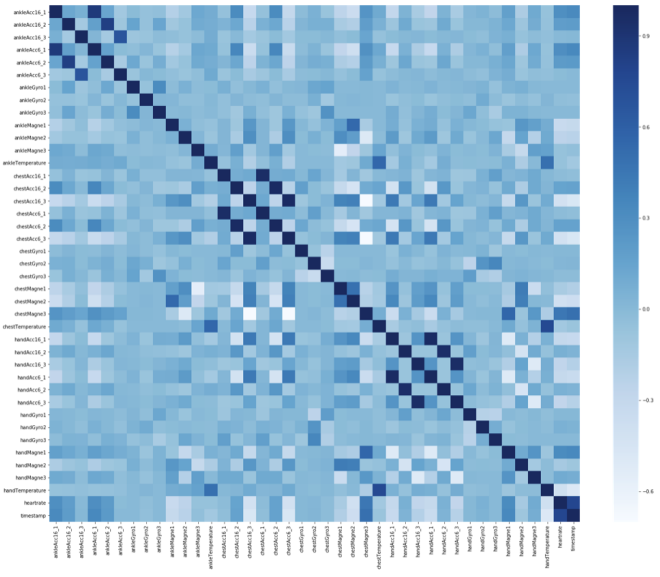

Fig. 6: Correlation matrix of numeric data.

not appear to be necessary. Contrarily, the relationship between hand accelerometers and temperature is clear. All three hand accelerometers show a strong correlation between the two. Furthermore, the chest magnetometers appear to be linked to heart rate, which is logical since they are located so close to each other.

D. Data Cleaning

1) Remove Orientation Columns

It was stated that data were collected under 16 different features for each UMI in the Data Set. For each UMI, 4 of the 16 features are Orientation data, and the data of these orientation features were declared invalid by the authors who published the dataset. Therefore, the 4 orientation columns in each UMI are dropped from the dataset.

2) Remove Transient Activities

Data records whose target feature activityID is transient with code 0 have been deleted.

3) Linear Interpolation for Heart-rate

Heart-rate can be an important indicator in classifying people's activities. For this reason, instead of deleting the missing data in the heart-rate column, linear interpolation technique was applied to fill it. Linear interpolation estimates the value to complete the linearity by looking at the filled data around the missing data, and it was decided that this is the best method to fill in the missing data, since heart rate does not give very different results in a short time.

4) Remove Irrelevant Features

The timestamp feature has been removed from the dataset because it is not a very effective feature in human activity classification.

Although the Chi-square test between subjectID and activityID shows that there is a relationship between these two features, the subjectID feature was also excluded from the dataset, considering that it would not contribute much to the model training.

E. Principal Component Analysis

In order to apply Principal Component Analysis, standardization 8 of the data was provided first. This operation could not be executed with MapReduce because the output is generated randomly as the MapReduce operation is performed as parallel tasks. This situation causes data and label matching problem. That's why the StandardScaler function of Python sckit-learn package is used.

$$z = \frac{x - \mu}{\sigma} \tag{8}$$

Then, PCA applicability was tested with standardized data. The KMO (Kaiser-Meyer-Olkin) and Bartlett tests combine all of the available data. A KMO of more than 0.5 and a significance level for the Bartlett's test of less than 0.05 indicate that there is a strong correlation in the data. To measure the degree to which two variables are closely linked, the concept of "variable collinearity" is used. In the study, the p-value of Bartlett's test was found to be 0.0 and the result of the KMO test was 0.7, thus demonstrating the applicability of PCA.

In order to find eigenvalues and eigenvectors, the covariance matrix of the standardized data was extracted. The formula 9 applied to extract the Covariance matrix was applied with the same structure as in Figure 5, parallelized with MapReduce.

$$S_{xy} = \frac{1}{N}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{9}$$

Eigenvalues and eigenvectors 10 are obtained from Numpy's linear algebra class using the eigh() function. The reason why the eigh() function is preferred to the eig() function in extracting the eigen values is that the symmetry of the covariance matrix has been verified.

$$\begin{aligned} A\boldsymbol{v} &= \lambda\boldsymbol{v} \\ where, & \\ \boldsymbol{A} &= S_W^{-1}S_B \\ \boldsymbol{v} &= \text{Eigenvector} \\ \lambda &= \text{Eigenvalue} \end{aligned} \tag{10}$$

Then explained variance ratio 11 is calculated. Explained variance represents how much variance each component explains.

$$explained\,variance\,of\,PC_k = \frac{eigen\,value\,of\,PC_k}{\sum_{i=1}^{p} eigen\,value\,of\,PC_i} \tag{11}$$

The calculated eigenvectors, eigenvalues, and explained variance ratios are then sorted in descending order. In this way, the process of choosing which components to use will be easier.

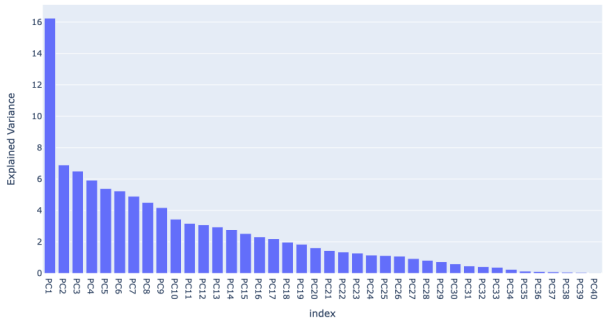

Fig. 7: Explained variance distribution.

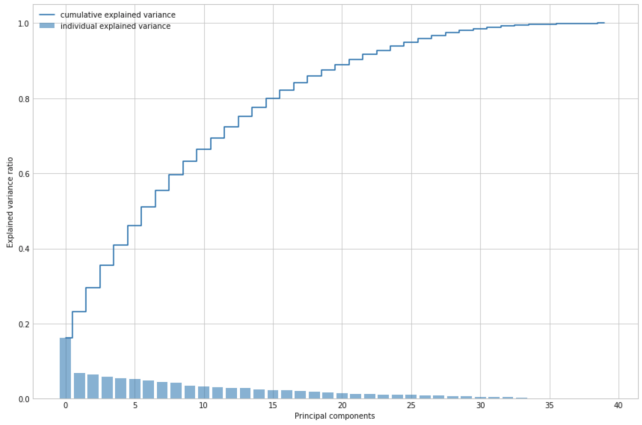

Fig. 8: Individual explained variance distribution againts cumulative explained ratio.

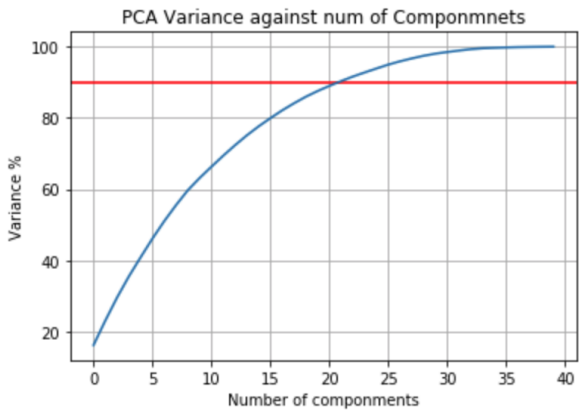

Fig. 9: Explained variance distribution againts number of components.

In Figures 7, 8 and 9, explained variance distribution, explained variance distribution against cumulative explained ratio and explained variance against component number are plotted, respectively.

About 90% to 98% of the variance can be accounted for by this model. As a result, by plotting the variance ratio against the number of components, it is possible to see how many components covered what percentage of variance. In the graph 9, it can be seen that 20 of the components fall within the range of 90% to 94% of the variance.

The scatter plot of the numerical data transformed with the selected 20 components looks like the following.

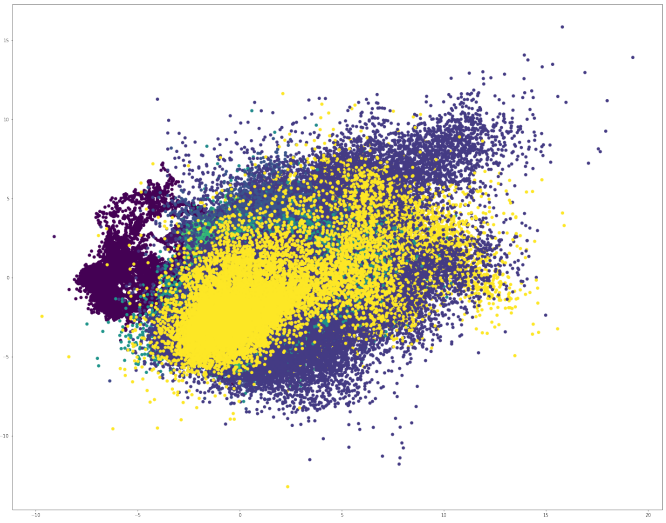

Fig. 10: Transformed numeric data.

In addition, the loading matrix, which shows the weights of the features in the components, was obtained by taking the transpose of the calculated Eigen vector matrix and is shown in Figure 11 as a heat-map.

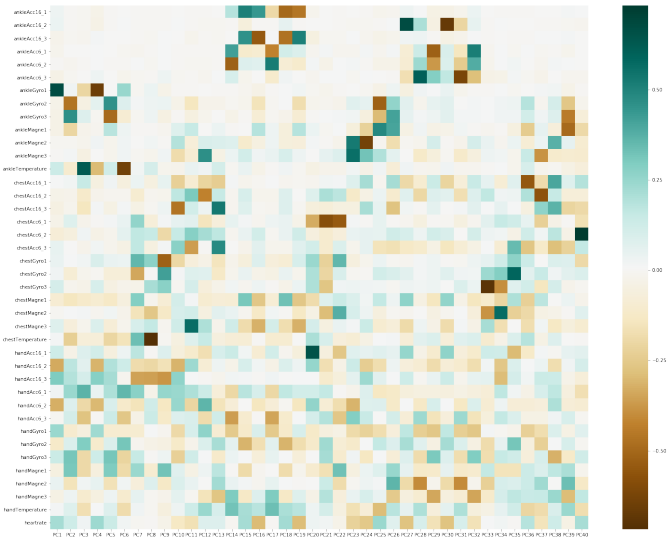

Fig. 11: Loading matrix of components.

F. Modelling

The raw data with 40 features and the pca-reduced dataset with 20 features are separately divided into test-train datasets at a rate of 20-80%. The data before and after PCA transformation was used to train the machine learning algorithm separately and the results of the models were compared.

Random Forest was chosen as the machine learning algorithm because it was stated in the literature that ensemble models work better in this human activity classification problem, and especially Random Forest algorithm was one of the most used algorithms.

TABLE I: Random Forest Test Results with Raw Data

Accuracy	Precision	Recall	F1
0.99951	0.99946	0.99940	0.99943

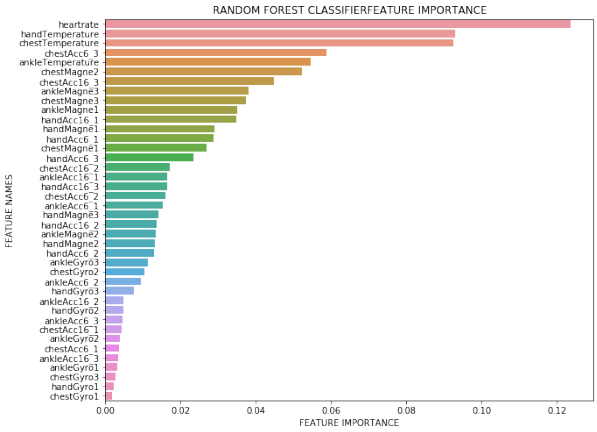

Fig. 12: Feature importances of raw data.

TABLE II: Random Forest Test Results with Transformed Data

Accuracy	Precision	Recall	F1
0.97867	0.97827	0.97319	0.97561

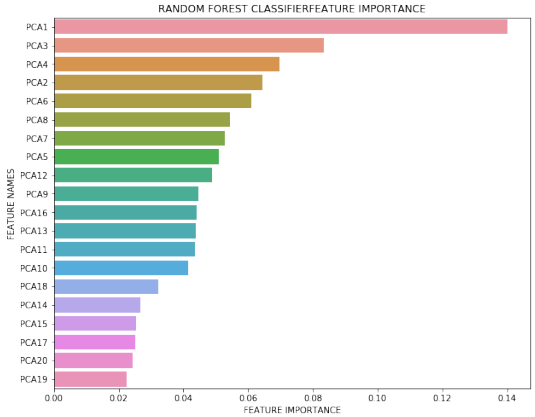

Fig. 13: Importances of components.

As a result, the test results of the model trained with the transformed data were lower by 2% compared to the test results of the model trained with raw data.

As the importance of the heart-rate feature has been proven for the solution of the problem, it has been revealed that hand and chest temperatures are also important indicators for the classification of human activity.

The list of features that contribute the most to the training of the model among the 20 separately selected components can also be seen in Figure 13. With this ranking result and the loading matrix of the components can be compared and it can be determined which feature is more important.

G. Cloud Environment

The implementation of the project has been done entirely on cloud-based virtual servers. The project is implemented using Linux instances on the OpenStack platform and the aim of the project was to observe the effect of multi-core processors on parallel programming and the MapReduce approach. All instances are provided by National College of Ireland.

The details of the environment created on the OpenStack are as follows;

OS: Linux
Distribution: Ubuntu
Version: 20.04
Instance Type: m1.xlarge
RAM: 16GB
VCPUs: 8VCPU
Disk: 160GB

IV. Conclusion

In this study, parallelized, scalable, distributed methods for data analysis, data pre-processing and dimensionality reduction for human activity classification problem and MapReduce program approach are discussed. The PAMAP2 dataset, which is a human activity recognition dataset in the UCI Machine Learning Repository, was used in the study. By applying Principal Component Analysis on the data set, dimensionality reduction was made and it was used to train Random Forest classifier after data transformation. The test accuracy results of the two Random Forest models trained separately with the transformed data and raw data were 0.97 and 0.99, respectively.

In the study, standardization and eigenvalue, eigenvector calculations, which are PCA stages, were not made with the MapReduce approach due to some technical reasons. This situation contradicts a fully scaled and parallelized structure. Sensor data can reach very large sizes and if fully scaled big data methods are not used in larger data at every stage, the project will not be able to be carried out because sufficient hardware infrastructure cannot be provided. Therefore, writing the Principal Component Analysis and transformation completely with the MapReduce approach has the potential to be one of the future studies.

References

[1] M. Gil-Martín, R. San-Segundo, F. Fernández-Martínez, and R. de Córdoba, "Human activity recognition adapted to the type of movement," *Computers Electrical Engineering*, vol. 88, p. 106822, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0045790620306789

[2] A. K. Chowdhury, D. Tjondronegoro, V. Chandran, and S. G. Trost, "Physical activity recognition using posterior-adapted class-based fusion of multiaccelerometer data," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 3, pp. 678–685, 2018.

[3] M. Arif and A. Kattan, "Physical activities monitoring using wearable acceleration sensors attached to the body." *PloS one*, vol. 10, no. 7, 2015.

[4] A. Baldominos, P. Isasi, and Y. Saez, "Feature selection for physical activity recognition using genetic algorithms," pp. 2185–2192, 2017.

[5] L. S. Ambati and O. El-Gayar, "Human activity recognition: A comparison of machine learning approaches," vol. 2021, no. 4, pp. 49–60, 2021.

[6] A. Reiss and D. Stricker, "Introducing a new benchmarked dataset for activity monitoring," 06 2012.