Scalable Systems Programming

Project Report

Marco Baglio
x20199520@student.ncirl.ie

Abstract— **This report is written for the project of Scalable Systems Programming for the postgraduate diploma in science in Data Analytics. The report shows an analysis of the data obtained from monitoring of the human physical activities by body sensors. First, an algorithm of activity classification is trained on the data from nine subjects. Second, a procedure of clustering aims to define the similarity among the subjects by considering only the physical parameters. The analysis is conducted on a large-scale environment by application of the Apache Spark architecture.**

I. Introduction

This report is written for the final project in Scalable Systems Programming for Data Analytics. The purpose of this project is the analysis of the data obtained from monitoring of the physical activity of nine subjects.

In particular, this report investigates the following research questions:

1. The calibration of an algorithm capable to classify the human activity based on the monitoring of the physical parameters.
2. Localize the differences among different subjects by categorizing the data obtained from the physical parameters. The scope is to group the subjects with similar physical conditions.

The report is structured with the following logic: first, there is a literature review, second a description of the methodology, third the interpretation of the results and finally the reflections on the future works.

II. Related Work

The authors of the datasets employed in this analysis have proposed five different classifiers [1]: decision tree, boosted decision tree, bagging decision tree, naïve Bayes and kNN. The perforamnce achieved are ~ 90% with high variance among the inviduals, determining that a personalized approach can be beneficial.

A subject-independent approach is proposed for the Human Activity Recognition (HAR) systems [2]. The HAR is applied to the SelfBACK: a system designed to assist people with low back pain (LBP) by monitoring of physical activity. The task of classify the physical activity introduces two important aspects: representation and personalization. The representation involves the capacity of provide generalized non-contraddictory results independent from the activity and sensors employed. The personalization regards the capacity to generalize a classifier without using a subject-dependent approach. The authors proposes a kNN algorithm based on the data collected from 50 volunteer participants aged between 18-54 years involved in 9 different activities. The system selects the individuals that are most similar to the target providing a subject-independent approach with high level of personalization and representation. The improvement from a general model is up to 5% of F1 score.

In the study of disease associated with obesity and lack of physical activity, it has been proposed a system to detect if an user is indoors or outside [3]. In situation of distrupted global positioning system (GPS), the application of multiple light and temperature sensors can be employed. The data were collected for 20 days with an average of two hours a day sampled with a frequency of 1 Hz. The kNN algorithm shows high preicision for this purpose, an error of 0.003 is claimed by the authors. The different weather conditions depending on the seasonlity requires to train the algorithm ad-hoc for the season analyzed.

An approach based on 7 wearable sensors placed in different body locations has been proposed [4]. Three healthy subjects with age between 25 and 35 performed experimental movement 10 times while the sensors recorded acceleration and rotations. The kNN classifier provided 98.4% accuracy on recognizing activities such as “stand to sit”, “look back” or “turn clockwise”.

Another approach to track the physical activities of the obese or overweight patients has been analuyzed [5]. The accelerometer in the smartphones was used to detect activities such as walking, jogging, sitting, etc. The kNN algorithm performed with high accuracy (99%) with the application 10 fold cross validation.

The application of triboeletric motion sensor for HAR has been proposed [6]. Five common activty were monitored: sitting, walking, climbing upstairs, downstairs and running. The kNN algorithm achieves the 80% of accuracy. The triboletric sensor can also be used to collect electric energy.

A method to detect outliers for k-means clustering of physicial activities has been proposed [7]. The algorithm is called FilterK and it is a combination of “distance based”, “density based” and “clustering based” detection techniques. FilterK was used to assign a degree of outlierness to the records of accelerations and its performance requires validation with new datasets.

Another paper provides a system for management of the blood sugar level for diabete patients based on optimization from a kNN algorithm[8]. The system monitors the blood sugar; if the level is high, it proposed to exercise, if the level is low proposes to eat and when the level is consistently too high the system prescribes to take the insulin. The system was trained with a kNN classifier base on seven factors: time, blood sugar, systolic blood pressure, diastolic blood pressure, amount of exercise, amount of meals and target consumption calories.

A single tri-axial accelerometer placed on waist was used to record acceleration data for human phsysical activity classfication [9]. 24 subjects performed real-life activities without external intervention. The study proposes the

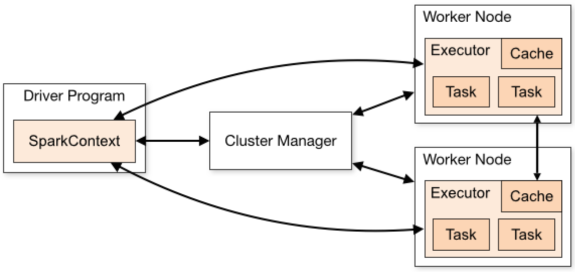


Figure 1 Overview of Spark architecture

application of dection tree or bayesian classification. The accuracy achieved in the activity classification is around 80% and the bayesian is preferred because of the flexibility provided with new future data.

III. Methodology

The analysis shown on this paper are carried out with Apache Spark, an analytics engine for large-scale processing. The system runs independent processes on a cluster coordinated by the SparkContext (Figure 1). The SparkContext is connected to a cluster manager which allocate resources across the applications. Once the connection is established, spark acquires executors on the nodes in the cluster and can send the tasks in order to be processed. Each application of this architecture is isolated from the others because the application owns unique executor processes; therefore, the data cannot be easily shared.

Databrick Community Edition is the environment used for this analysis. Databrick is a cloud-based big data platform that allows access to cluster, cluster manager and notebook environment. For demonstrative purposes, Figure 2 shows the summary of the activity carried out by the executors during the analysis of this paper.

The dataset used is freely available for academic research and there are no constraints on using the data for scientific purposes [1]. The subjects are 8 males and 1 female, the age is 27.22 ± 3.31 years and the BMI is 25.11 ± 2.62 kgm^{-2}. All the individuals followed a protocol of 12 activities (lie, sit, stand, walk, run, cycle, Nordic walk, iron, vacuum clean, rope jump and using stairs). The data collected includes 54 features:

1. Time
2. Activity ID
3. Heart rate
4. Inertial measurement unit (IMU) Hand
5. IMU chest
6. IMU angle

Each IMU contains:

- Temperature
- 3D acceleration
- 3D gyroscope data
- 3D magnetometer data
- Orientation

Ten hours of data were collected for each subjects. The nine datasets joined together includes 198371x54 data points. The amount of data involved requires the application of a scalable distributed computing approach and PySpark is used for this purpose.

The next sub-sections examine the analysis conducted for each research question.

A. *Activity classification*

The first task regards the classification of the activity based on the values obtained from the monitoring of the physical activities. The Figure 3 shows the workflow for the task 1:

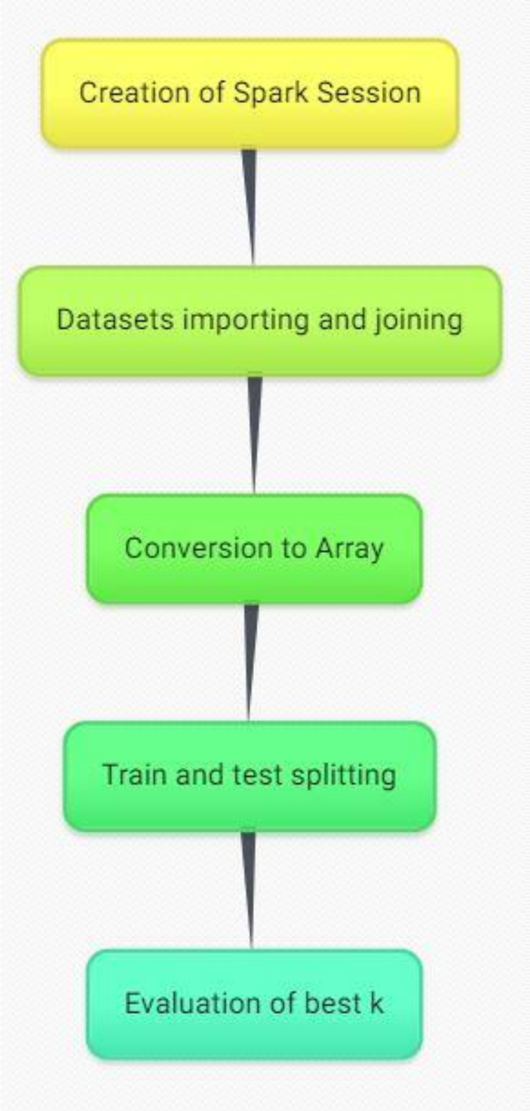
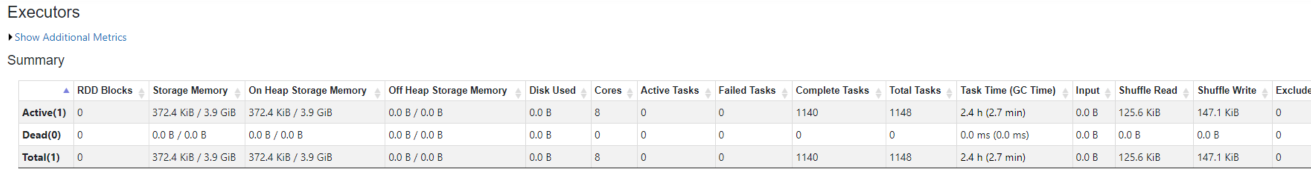


Figure 3 Workflow adopted for the task 1


Executors

Show Additional Metrics

Summary

	RDD Blocks	Storage Memory	On Heap Storage Memory	Off Heap Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Exclude
Active(1)	0	372.4 KiB / 3.9 GiB	372.4 KiB / 3.9 GiB	0.0 B / 0.0 B	0.0 B	8	0	0	1140	1148	2.4 h (2.7 min)	0.0 B	125.6 KiB	147.1 KiB	0
Dead(0)	0	0.0 B / 0.0 B	0.0 B / 0.0 B	0.0 B / 0.0 B	0.0 B	0	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	0
Total(1)	0	372.4 KiB / 3.9 GiB	372.4 KiB / 3.9 GiB	0.0 B / 0.0 B	0.0 B	8	0	0	1140	1148	2.4 h (2.7 min)	0.0 B	125.6 KiB	147.1 KiB	0

Figure 2 Screenshot of executors activity for one of the analysis carried out.

1) Creation of the spark session

The first step is the creation of a spark session or context (see Figure 1) that is the core element of the analysis.

2) Datasets importing and joining

At this step the data are pre-processed.
First, nine datasets in format .dat were uploaded into the environment.
Second, each dataset is read and the data from each subject is stored.
Third, the non-valid values are dropped.
Finally, the processed data joined in a unique dataframe.

3) Conversion to array

This step involves the selection of the column containing the activity ID, a unique numeric value identifying the activity.
Finally, the activity values and the data from monitoring are converted from a dataframe to a numeric array to allow the computation of machine learning algorithms.

4) Train and Test splitting

The data are divided into two groups. The train data, used to calibrate the kNN algorithm and the test data, used to validate the results obtained. In this analysis, the 30% of data are used for testing and the 70% for training purposes.

5) Evaluation of best k

The last step for this task is the evaluation of the best k value for the kNN algorithm. The algorithm is trained with values of k ranging from 2 to 10 and the accuracy values are computed for both training and test set.

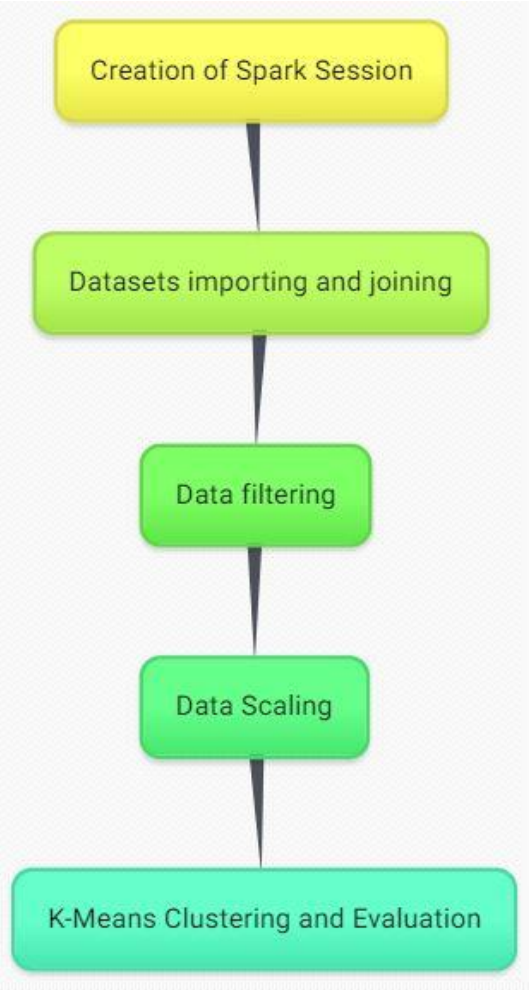

Creation of Spark Session
Datasets importing and joining
Data filtering
Data Scaling
K-Means Clustering and Evaluation

Figure 4 Workflow adopted for the task 2

B. Activity clustering

The second task regards the clustering of values obtained from the monitoring of the physical activities. Clustering has the scope to group all the similar values together without the specification of a target value, de facto applying an unsupervised algorithm to the data obtained.
This task is divided into three sub-tasks:

1. Clustering based on the heart rate during walking.
2. Clustering based on the chest temperature during walking.
3. Clustering based on the combination of chest temperature and walking activity during walking.

The Figure 4 shows the workflow for the task 2:

1) Creation of the spark session

Similarly, to the task 1, the first step is the creation of a spark session or context (see Figure 1) that is the core element of the analysis.

2) Datasets importing and joining

A dataset containing the information from the monitoring of each subject was uploaded on Databricks. Therefore, each dataset is read and pre-processed; during this step, the values non-valid are dropped.
Finally, the processed data joined in a unique dataframe.

3) Data filtering

Only the data obtained during the activity of walking are kept, all the other values are dropped. This step involves the selection of the column containing the heart rate, the chest temperature and time.
Finally, the activity values and the data from monitoring are converted from a dataframe to a numeric array to allow the computation of machine learning algorithms.

4) Data scaling

This step applies only to the third sub-task. The heart rate and the chest temperature are normalized separately and combined in a single vector.

5) K-Means Clustering and Evaluation

The last step for this task is the evaluation of the best k value for the clustering algorithm. The algorithm is trained with values of k ranging from 2 to 10 and the Silhouette score is computed for each value. The best value of k is chosen according to the best compromise of silhouette score and cluster representability.

IV. Results and Interpretation

1) Task 1

The task 1 involves the classification of the activity based on the monitoring of physical parameter.
Figure 5 shows the values of accuracy obtained on the training set for different values of k. In general, the accuracy obtained is always above 0.975, denoting that a good level of performance of the kNN algorithm for this task. The accuracy

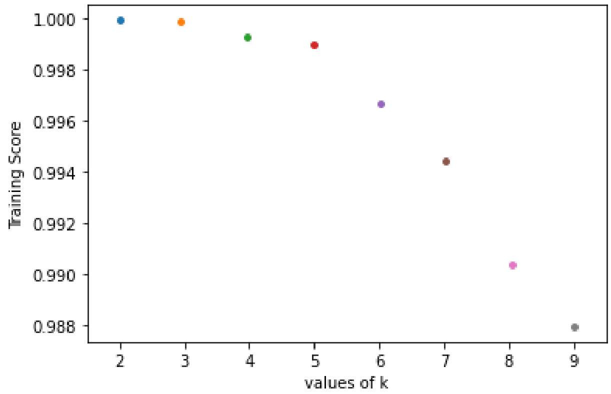


Figure 5 Values of accuracy depending on the k parameter adopted for the training set.

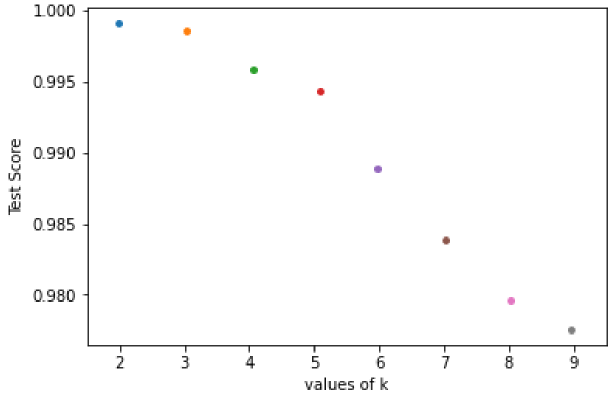


Figure 6 Values of accuracy depending on the k parameter adopted for the test set.

is stable around 0.99 for k values smaller than 6 and it starts to degrade for values larger.

Figure 6 shows the values of accuracy obtained on the test set for different values of k. Similarly, to the training set, the values of accuracy are always above 0.975, validating the usage of kNN algorithm. The accuracy starts to decrease after a k value of 5, which is chosen as the optimal k parameter.

2) Task 2

The task 2 involves the clustering of the data obtained during walking activity in order to detect similarity among the subjects. This task is divided into three sub-tasks; the results are discussed in the next three sub-sections.

The parameter used to evaluate the goodness-of-fit for the clustering analysis is the silhouette score. It ranges from -1 to +1, the large value denotes that the cluster has good matching within the elements and poor matching with other clusters. Small values denotes the opposite, poor matching within the cluster and good matching with other clusters. The values are calculated for each point of the cluster; the average is used as representative value.

a) Heart rate analysis

The values of heart rate have been analyzed during walking for all the subjects in order to obtain clusters of similar subjects.

Figure 7 shows the average silhouette score depending on the value of k. The score is always above 0.65 for a k parameter between 2 and 9. The optimal k is to be researched in the trade-off between the score and the representability of the clusters. In this case, 4 is the optimal k, since larger values do not show a relevant deviation on the score.

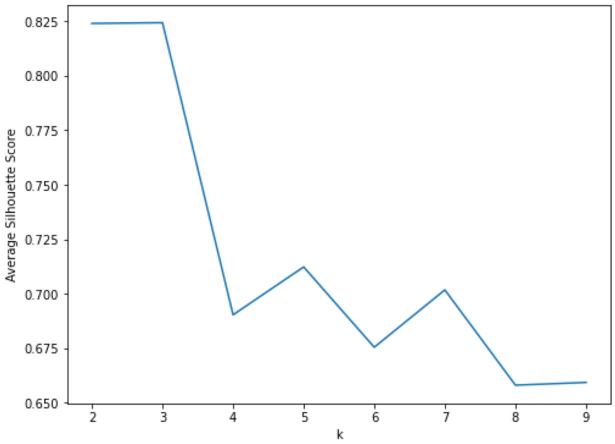


Figure 7 Heart rate analysis: values of silhouette scores for different values of k.

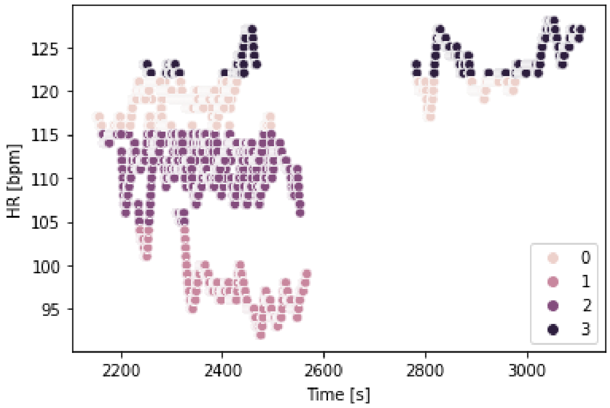


Figure 8 Heart rate analysis: graphical representation of the clusters. The scatter shows the heart rate values along time, the color legend represent the 4 clusters obtained.

Figure 8 allows a graphical representation of the clusters. The heart rate of all the subjects is plotted versus the time of recording. The clusters are represented by the colors, four bands defines the cluster area. The four bands can be interpreted as four possible difficulties of walking (i.e. speed of walking or slope) or as a difference on the health conditions of the subjects.

b) Chest temperature analysis

The values of the chest temperature have been analyzed during walking for all the subjects in order to obtain clusters of similar subjects.

Figure 9 shows the average silhouette score depending on the value of k. The score is always above 0.70 for a k parameter between 2 and 9. In this case, there is a positive trend where the silhouette score increases within 2 and 4. The maximum silhouette score is obtained with 4 and it is 0.86.

Figure 10 allows a graphical representation of the clusters. The chest temperature of all the subjects is plotted versus the time of recording. The clusters are represented by the colors, four bands defines the cluster area. The four bands can be interpreted as four possible difficulties of walking (i.e. speed of walking or slope) or as a difference on the health conditions of the subjects.

c) Combined analysis

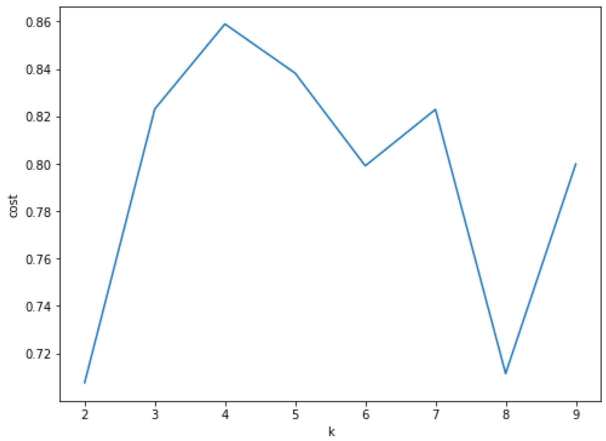


Figure 9 Chest temperature analysis: values of silhouette scores for different values of k.

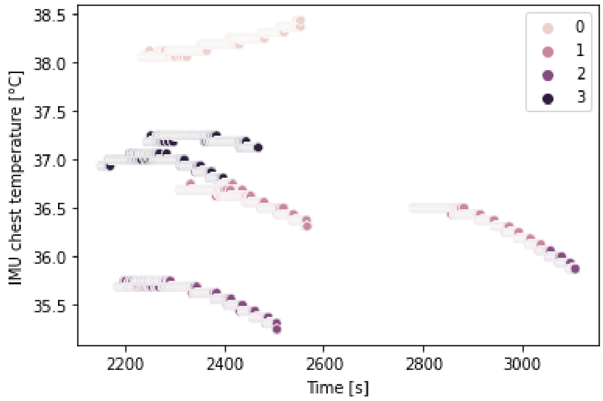


Figure 10 Chest temperature analysis: graphical representation of the clusters. The scatter shows the heart rate values along time, the color legend represent the 4 clusters obtained.

The values of heart rate and chest temperature have been analyzed during walking for all the subjects in a combined analysis to determine clusters of similar subjects.

The previous plots showed the average silhouette scores among all the clusters according to the k value. On this instance, Figure 11 and Figure 12 shows a series of graphs representing the silhouette score of each point within the cluster. The red dashed line individuates the optimal silhouette score; the optimal k value should show the most number of color bands close or larger than the threshold value.

When k is 2 or 3, the color bands are very similar and a general good result obtained, however the clusters do not provide a sufficient categorization.

The values between 6 and 9 provide similar silhouette scores with a general tendency of showing a remarkable number of points with a low silhouette score. Furthermore, none of the plots show negative silhouette scores.

The optimal k is to chosen between 4 and 5; both options are valid, a value of 5 can be adopted to increase the number of groups and possible interpretations.

V. Reflections and Future work

This paper proposed an analysis of physical activities from sensor monitoring. Nine subjects have been monitored by several sensors, providing information regarding heart rate, acceleration, temperature and rotation.

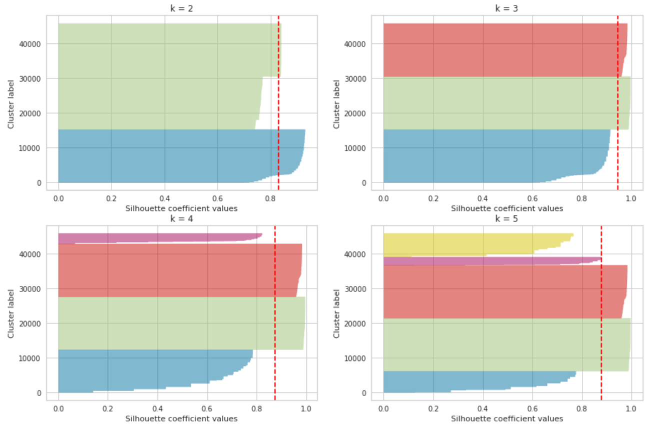


Figure 11 Combined analysis: Silhouette coefficient values for k ranging between 2 and 5.

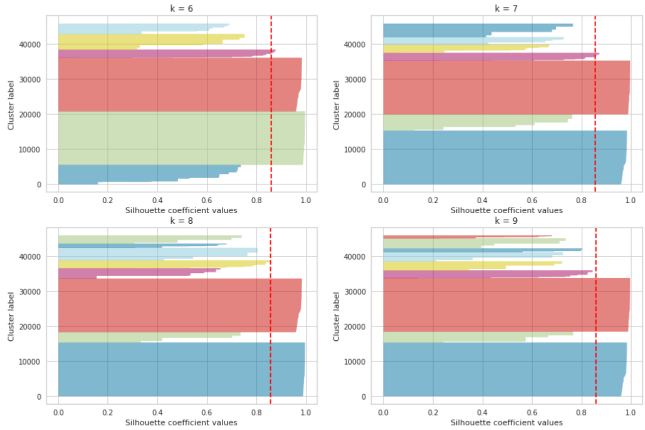


Figure 12 Combined analysis: Silhouette coefficient values for k ranging between 6 and 9.

First, the data obtained have been used to train an algorithm of activity classification. The kNN algorithm was employed and an accuracy of 0.99% was obtained on the test set.

Second, the data have been used to determine groups of similar subjects according to the heart rate and chest temperature during walking activity. The k-means clustering algorithm was applied for this purpose. Four categories was found to be the optimal representation. The categories can be related to a different level of intensity of exercise or a different health condition of the subjects.

Table I provides a succinct summary of the parameters used in this paper.

The analysis was conducted on a large-scale environment and the scalability of the analysis provides interesting studies in the future. More data can be gathered from different subjects, therefore validate the clusters obtained and provide a scientific interpretation by determining the physical condition of the subjects or the intensity of the activity. The

Table I Summary table of the optimal values obtained for each task. Note that the k parameter specified for task 1 is different from the k parameter of task 2, since they serve two different algorithms.

Task 1: Activity classification - kNN	**k = 5**
Task 2.1: HR	**k = 4**
Task 2.2: IMU chest	**k = 4**
Task 2.3: HR + IMU	**k = 5**

activity classification on this paper confirms that subject-dependent analysis provides large values of accuracy and, at the moment is the best method to classify human activities.

VI. References

[1] A. Reiss and D. Stricker. Introducing a New Benchmarked Dataset for Activity Monitoring. The 16th IEEE International Symposium on Wearable Computers (ISWC), 2012.

[2] Sani, Sadiq, et al. "kNN sampling for personalised human activity recognition." International conference on case-based reasoning. Springer, Cham, 2017.

[3] Troped, Philip J., and Jeffrey J. Evans. "Environment feature extraction and classification for Context aware Physical Activity monitoring." 2013 IEEE Sensors Applications Symposium Proceedings. IEEE, 2013.

[4] Saeedi, Ramyar, et al. "Toward seamless wearable sensing: Automatic on-body sensor localization for physical activity monitoring." 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE, 2014.

[5] Arif, Muhammad, et al. "Better physical activity classification using smartphone acceleration sensor." *Journal of medical systems* 38.9 (2014): 1-10.

[6] Huang, Hui, Xian Li, and Ye Sun. "A triboelectric motion sensor in wearable body sensor network for human activity recognition." 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2016.

[7] Jones, Petra J., et al. "FilterK: A new outlier detection method for k-means clustering of physical activity." Journal of biomedical informatics 104 (2020): 103397.

[8] Lee, Malrey, Thomas M. Gatton, and Keun-Kwang Lee. "A monitoring and advisory system for diabetes patient management using a rule-based method and KNN." Sensors 10.4 (2010): 3934-3953.

[9] Long, Xi, Bin Yin, and Ronald M. Aarts. "Single-accelerometer-based daily physical activity classification." 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE, 2009.