02582 Computational Data Analysis
Case 2

Max Specktor - s184362
Christos Gousis - s222871
Georgios Kapakoglou - s223001
Ioannis Koumaniotis - s212887

3 May 2023

1 Introduction

The report corresponds to the second assignment of the course "Computational Data Analysis". The purpose of the project is to analyse physiological measurements based on real data coming from a series of experiments that have been taken place at the Technical University of Denmark. To achieve this, state of the art machine learning algorithms will be used in order to discover hidden patterns in the data.

In unsupervised learning, the response (y) is unknown. Hence, the goal is to label the data with clusters based on common characteristics. The algorithms used in our case to accomplish this task are Principal Component Analysis, K-means clustering and Hierarchical clustering, which will be analyzed in detail later in the report.

Taking into consideration the provided dataset over three acquisition rounds (D11, D12, D13), the research goal is to analyze the behaviour of the participants during the first two phases of the experiment. More specifically, we will try to cluster the participants based on the common patterns that they have. Furthermore, we will try to investigate the biosensor data of the participants in phase 1 compared to their data during the phase 2, in which the puzzle game took place. Also, another area of exploration is to analyze the percentage of the puzzlers and instructors for each cluster and see if there is any change of these distributions from phase one and two.

2 Data exploration

The dataset consists of survey and biosensor data for 26 subjects divided into 3 different cohorts. The cohorts were divided into teams of 2, with one "instructor" and one "puzzler" on each team. Each cohort was subject to 4 rounds of the same experiment. The experiment was divided into three 5-minute phases of biosensor data gathering:

- Pre-puzzle phase
- Puzzling phase
- Post-puzzle phase

At the end of each of these phases, a self-assessment questionnaire was filled out. The biosensor data consists of 4 different measurements, electrodermal activity (EDA), heart-rate (HR), temperature (TEMP) and blood-volume pulse (BVP).

- Electrodermal Activity (EDA), is a measure of the electrical conductance of the skin, which changes in response to emotional or physiological arousal. By detecting these changes, EDA can be used to measure emotional responses and physiological arousal in various research fields, such

as psychology, neuroscience, marketing, and human-computer interaction.
In our data the range of the values is (0.1, 46.4)

- Heart-rate (HR),is a physiological measurement that refers to the number of times the heart beats per minute. It is an important indicator of cardiovascular health, as the heart rate can vary depending on the individual, their level of physical fitness, and their current state of health.
  In our data the range of the values is (53.3, 158.0)

- Temperature (TEMP), refers to the measurement of skin temperature using a biosensor. Skin temperature is the temperature of the outer layer of skin, which can be influenced by various factors such as environmental temperature, blood flow, and metabolic activity. It can also be used as a physiological marker for stress and emotional responses, as changes in skin temperature are often associated with changes in sympathetic nervous system activity.
  In our data the range of the values is (29.2, 35.8)

- Blood-Volume Pulse (BVP) is a physiological measurement that refers to changes in blood volume within a particular region of the body, usually the fingertip.
  In our data the range of the values is (-1629, 1650)

In order to prepare the data for principal component analysis (PCA) in our project, we performed some preprocessing steps. Specifically, we downscaled certain datasets so that they had the same number of observations, but without affecting their original distribution. This was done to ensure that each variable had the same weight in the PCA analysis. Once the data was standardized, we proceeded with the PCA analysis to identify the underlying patterns and structure in the data.

3 Methods

The methods that have been used in the research projects are the Principal Component Analysis (PCA), K-Means clustering and the Hierarchical clustering. The purpose of the Principal Component Analysis is the dimensionality reduction of the features and the linear transformation of the data into a new coordinate system. Moreover, K-means and Hierarchical clustering are both unsupervised learning algorithms that aim to group similar observations together based on the similarity of their features. K-means clustering partitions the data into K clusters, where K is a user-defined parameter, by minimizing the sum of squared distances between the observations and the centroid of their assigned cluster.

$$J(C) = \sum_{k=1}^{K} \sum_{i \in C_k} ||x_i - \mu_k|| \rightarrow \min_{C} \tag{1}$$

By way of contrast, Hierarchical clustering, does not require the specification of K, but instead builds a hierarchical tree-like structure of nested clusters.In the case of Hierarchical clustering, the process of searching for the nearest cluster is conducted with Complete linkage as depicted below:

$$d(C_i, C_j) = \max_{x_i \in C_i, x_j \in C_j} ||x_i - x_j|| \tag{2}$$

Where, $|x_i - x_j||p$ represents the Euclidean distance between two data points x_i and x_j.

Regarding the first method, the distance metric used for our case is the Euclidean distance as seen above:

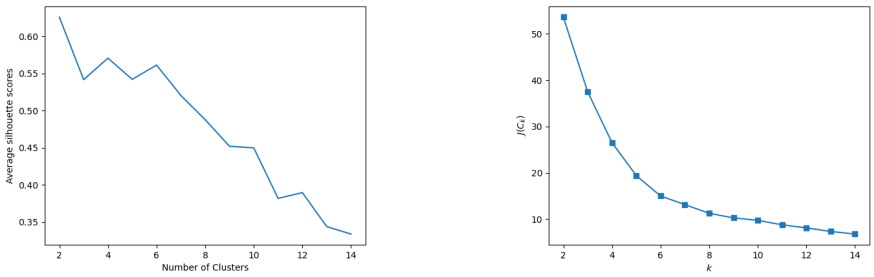
$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} \tag{3}$$

In order to find the optimal value of the number of Clusters (K), two methods were used. The first method is the silhouette (heuristic) and the second is the elbow method.

Silhouette method is expressed as seen below:

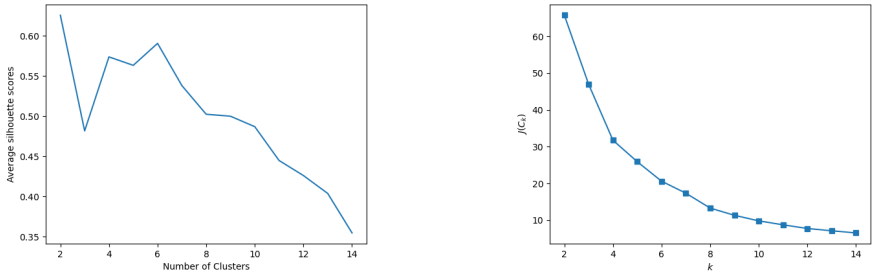
$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \tag{4}$$

Where, a(i) is the average distance between observation i and all other observations assigned to the same cluster, and b(i) is the average distance between observation i and all observations assigned to the neighboring cluster.



(a) Silhouette scores (b) Elbow method

Figure 1: Average silhouette scores and Elbow method for phase 1 bio-sensor data



(a) Silhouette scores (b) Elbow method

Figure 2: Average silhouette scores and Elbow method for phase 2 bio-sensor data

As can be seen from Figures 1 and 2, the silhouette plots show that for both phases the values 2, 4 and 6 are the most prevailing as number of clusters since they have the highest average scores. In order to find the final number of cluster, the elbow method suggests the number 6 as the optimal value for the clustering, for both phases.

4 Results

4.1 PCA Results The results of the PCA components shown in table 1, indicate that there are high values for temperature and EDA in each phase, which indicates that these variables have a positive contribution to the principal component, while the negative value for HR suggests that it has a negative contribution. The BVP variable has a relatively low contribution to the principal component compared to the other variables. Overall, this suggests that temperature and EDA are the most important variables in explaining the variance in the data, while HR has less importance, and BVP has the least importance.

	Temp.	BVP	EDA	HR
Phase 1	0.694	0.296	0.711	-0.108
Phase 2	0.699	-0.017	0.700	-0.148

Table 1: PCA components for Phase 1 and Phase 2 bio-sensor data

4.2 K-Means Clustering

Based on our clustering analysis, we observed that participants in higher clusters had higher levels of emotional arousal in response to the experiment, and this arousal steadily increases throughout the phase in the highest cluster. This finding may suggest that the experiment was particularly effective in eliciting

Phase	Cluster	Participants	Puzzler %
1	0	11, 17, 21, 25, 26	20%
	1	12, 13	100%
	2	1, 3, 5, 8, 16, 22, 24	14.3%
	3	7, 9, 10, 14	50%
	4	2, 4, 15, 18, 20, 23	50%
	5	6, 19	50%
2	0	6, 8, 16, 19, 22, 24	16.7%
	1	9, 10	0%
	2	1, 2, 3, 4, 5, 15, 17, 18, 20, 23, 25	45.5%
	3	13	100%
	4	12, 14	100%
	5	7, 11, 21, 26	100%

Table 2: K-Means clustering results

varying emotional responses from participants.2 In the clustering of phase 2, we see a concentration of Puzzlers in to certain clusters, which would support the idea that they have a different biological response because of their role in the experiment.

The clustering could also be investigated for any seasonal effects on the participants. Looking at the winter participants (first 8 participants), we can see that half of them fall into 1 cluster in phase 1, and five of them are clustered together in phase 2. A similar trend occurs in the hierarchical clustering as well.

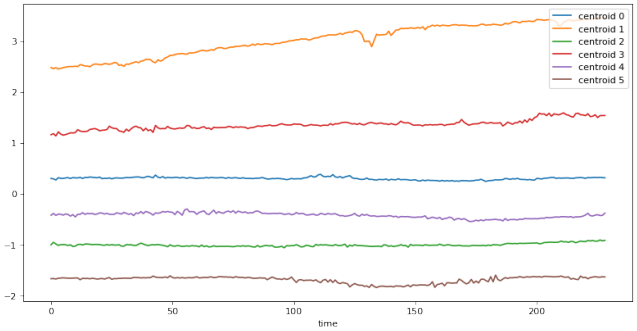


Figure 3: K-means clustering of first PCA component for phase 1 bio-sensor data

4.3 Hierarchical Clustering The results of hierarchical clustering are not as easy to visualize, unlike the centroids of K-means clustering. But the

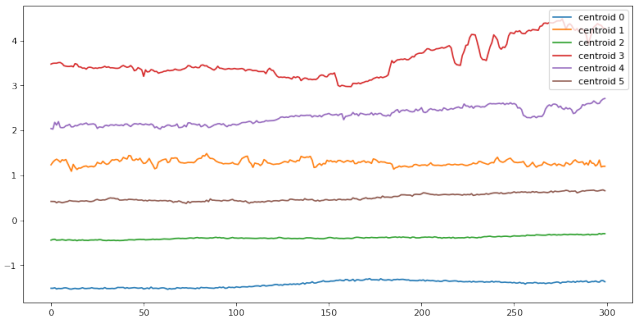


Figure 4: K-means clustering of first PCA component for phase 2 bio-sensor data

change in clustering from phase 1 to 2 does show a slightly higher concentration of the puzzlers in certain clusters, just as with the K-means method.3

Phase	**Cluster**	**Participants**	**Puzzler %**
1	0	1, 3, 5, 6, 8	20%
	1	2, 4, 15, 16, 18, 20	33.3%
	2	7, 9, 10, 14	50%
	3	11, 17, 21, 23, 25, 26	83.3%
	4	12, 13	100%
	5	19, 22, 24	0%
2	0	1, 2, 3, 4, 5	20%
	1	6, 8, 16, 19, 22, 24	16.7%
	2	7, 9, 10, 11, 21, 26	66.7%
	3	12, 14	100%
	4	13	100%
	5	15, 17, 18, 20, 23, 25	66.7%

Table 3: Hierarchical clustering results

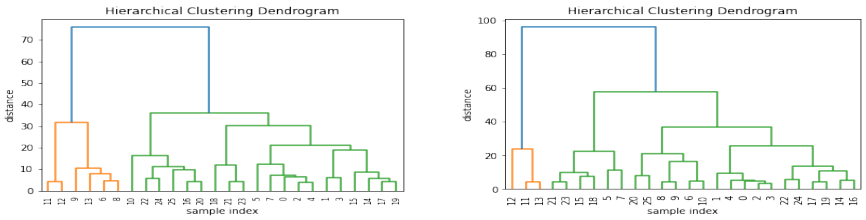


Figure 5: Hierarchical clustering of phases 1 & 2 bio-sensor data

In Phase 1, the hierarchical clustering and k-means clustering have some similarities. Both methods have a cluster with participants 1, 3, 5, 6, and 8, but the k-means clustering also includes participant 7 in this cluster. The hierarchical clustering has a cluster with participants 2, 4, 15, 16, 18, and 20, which is not present in the k-means clustering. The hierarchical clustering has a separate cluster for participants 12 and 13, while the k-means clustering includes them in a larger cluster. Overall, the hierarchical clustering has more clusters in Phase 1 compared to the k-means clustering.

In Phase 2, the hierarchical clustering and k-means clustering show some differences. The hierarchical clustering has a cluster with participants 1, 2, 3, 4, and 5, which is not present in the k-means clustering. The k-means clustering has a larger cluster with participants 6, 8, 16, 19, 22, and 24, while the hierarchical clustering splits them into three separate clusters. The hierarchical clustering has a separate cluster for participant 14, while the k-means clustering includes participant 14 in a larger cluster. The hierarchical clustering has a cluster with participants 15, 17, 18, 20, 23, and 25, which is not present in the k-means clustering. Overall, the hierarchical clustering has more large clusters in Phase 2 compared to the k-means clustering.

5 Conclusion

It is important to note that our analysis was limited to four physiological measures (temperature, EDA, HR, and BVP) and may not capture the full range of emotional experiences, as the data is rather limited. There are only 26 participants and our analysis does not attempt to incorporate all 4 rounds of the experiment. In theory, we would prefer to have more participants instead of more data from the same participants. For this reason, we chose to only include the data from round 1.

It is interesting that the PCA analysis gave a negative score to the heart rate data, although this may be because it had a significant amount of correlation to the other features and therefore wasn't useful in explaining the variance in the data.

Another research could perform inferential statistics tests such as hypothesis testing or regression analysis to determine if there is a significant difference in the distribution of participants across clusters or if there is a relationship between the clusters and other variables such as the season when experiment conducted.