Clustering cardiac rehabilitation data: a preliminary study

Daphne T. C. Lai[1], Syazwina Yasmin[1], Seng Khiong Jong[2], Sok King Ong[3] and Chean Lin Chong[2]*

[1]Faculty of Science, Universiti Brunei Darussalam, Jalan Tungku Link, Gadong, BE1410, Brunei Darussalam
[2]Raja Isteri Pengiran Anak Saleha Hospital (RIPASH), BE1518, Brunei
[3]Ministry of Health, BB3910, Brunei

**corresponding author email: daphne.lai@ubd.edu.bn*

Abstract

In this paper, a clustering framework is used on cardiac rehabilitation data to discover meaningful patterns. The data were collected in three phases. Kmeans clusters were generated and evaluated for stability. Visual assessment of the clusters using PCA plots was also done. A scoring system was developed to quantify improvement in the patients' health across the three phases. With the scores, association and correlation measures were employed to assess the meaningfulness of the clusters. Two distinct clusters were found and they were shown to have moderate clinical association (Cramer's V score=0.27) with the improvement scores.

Index Terms: data clustering, longitudinal data, cardiac rehabilitation

1. Introduction

Clustering is widely used to discover hidden structures in unlabeled datasets using a pre-defined similarity measure. Its application covers a wide range of data types such as numerical, binary, image, textual, videos, ordinal and so on. While its application was predominantly on static data, it has long gained grounds in time-based ones such as time series[1] and longitudinal[2] data. It is worth noting that they are different. Time-series data contains uni- or multi-variate data of an incident (such as unemployment, earthquakes), often collected at regular time intervals while longitudinal data contains multivariate data collected from the same subjects with repeated measurements at different time intervals.

Liao[1] has provided an extensive piece on the different clustering algorithms applied to time series, univariate and multivariate data. According to him, there are three different clustering approaches; raw-data-based, feature-based and model-based. Heggeseth[2] has described two main approaches for longitudinal data; nonparametric and model-based. For both time-based datatypes, Kmeans was employed.

For this reason, as preliminary work, we applied a nonparametric approach using Kmeans on a longitudinal patient dataset acquired from a cardiac rehabilitation programme (CRP). So far, statistical techniques have been applied on this dataset (publication is in preparation) but, no clustering techniques have been tried. The aim of this work is to discover clinically meaningful clusters in the dataset. The clusters generated were assessed for stability and clinical relevance.

In this work, two stable clusters were found with moderate clinical relevance (Cramer's V score=0.27)[3,4] to improvement scores based on a collective measure of patients' parameters. Interestingly, these scores were found to have little correlation to each individual parameter but, was able to represent the time-based nature of the data to be used for statistical assessment of the clusters.

2. Experimental approach

Dataset

The dataset contains 180 records of RIPASH CRP patients enrolled between the years 2009-2013. It consists of the following parameters: Short Form

36 Health Survey (SF36), Exercise Test Time (ETT), Resting Blood Pressure (SBP/DBP), Resting Heart Rate (HR), Fasting Blood Sugar (FBS), Total Cholesterol (TC), High Density Lipoprotein (HDL), Low Density Lipoprotein (LDL), Triglyceride (TG), Weight (WT), Waist Circumference (WC) and smoking status (SS). The data is considered longitudinal as the mentioned parameters were recorded from the same patients at enrolment (P1), end of Phase II (P2), which last 8 weeks and end of Phase III (P3), which last another 18 months.

Methodology

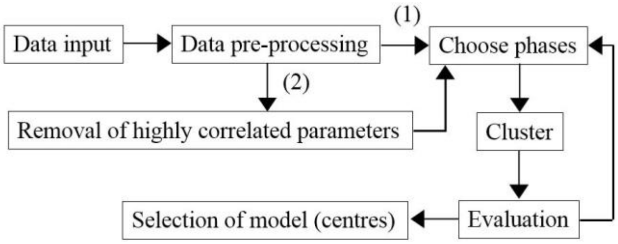
Data input
Data pre-processing
(1)
Choose phases
(2)
Removal of highly correlated parameters
Cluster
Evaluation
Selection of model (centres)

Figure 1. Clustering Framework using (1) raw-data-based and (2) feature-based approaches.

Figure 1 illustrates the steps taken to discover and assess meaningful clusters in the data set. After data pre-processing, we obtained 63 complete and normalised data records, out of the original 180 records. Next, we experimented using the raw-data-based and feature-based approaches. In the later approach, Pearson's correlation was applied on all parameters across the three phases and those that were highly correlated were removed.

To determine clustering solutions of interest, data from different individual phase or combination of phases were chosen to be clustered using Kmeans.

The number of clusters is determined by looking at the "elbow" in the sum of squared error (SSE) scree plot, and the clustering is performed for several runs. The stability of the clusters found across the runs was assessed using the within-sum-of-square (WSS) measure.

As part of clinical evaluation, a scoring system involving parameter comparison between those in 1) P1 and P2 and 2) P2 and P3 is used. For the parameters DBP, HR, FBS, WC, LDL, TG, SS, SBP, WT, BMI and TC, a decrease between two compared phases is regarded as an improvement. The opposite is true for the parameters: ETT, SF36 and HDL. A score of '1' assigned for improvement while a score of '0' indicates otherwise. The scores are then summed for each patient. The Cramer's V coefficient is used to measure clinical association between two nominal variables; cluster assignments and improvement scores of patients. If the level of association is low (<0.2),[3] a new experiment was conducted using different phases. PCA plots were also used to assess the clinical relevance of the clusters.

Experiments

Five experiments on different phases were carried out to find clustering solutions of interest:

1. Clustering individual phases
2. Clustering differences between phases
3. Clustering phase 1 and phase 2 without highly correlated parameters
4. Clustering phase 1 and phase 2 with all parameters
5. Clustering three phases together with all parameters

For each experiment, we conducted 10 runs. All experiments are implemented in **R**.[6]

3. Results and Discussion

In this section, we will present results from experiment 4 and 5 in greater detail, as they are found to be most interesting.

For experiment 1 to 3, the "elbow" on the scree plot was at k=3. For this reason, we test running Kmeans 10 times with k=2, 3 and 4. Solutions with k=2 are most stable. Further investigation into experiments 1 to 3 were stopped because, either clusters found were from individual phases and did not demonstrate observable trend across the phases. Furthermore, the removal of highly correlated parameters prior to clustering means the parameter for a particular phase is not represented. More complex approaches such as those detailed[1,2] are required to be investigated in order to proceed with such experiments.

Table 1. Phase comparison for cluster centres 1 and 2 based on P1 and P2.

Clus1	***ET***	***SF36***	***SBP***	***DBP***	***HR***	***FBS***	***WT***	***BMI***	***WC***	***TC***	***HDL***	***LDL***	***TG***	***SS***
P1	8.26	69.49	132.51	81.54	72.37	6.05	77.27	29.93	38.38	4.19	1.05	2.38	1.65	0.57
P2	9.88	78.26	122.57	74.86	69.31	5.89	77.33	29.85	37.81	3.85	1.04	2.09	1.58	0.06
P1vsP2	↑H	↑	↓H	↓	↓	↓H	↑H	↓H	↓H	↓	↓	↓	↓H	↓
Clus2	***ET***	***SF36***	***SBP***	***DBP***	***HR***	***FBS***	***WT***	***BMI***	***WC***	***TC***	***HDL***	***LDL***	***TG***	***SS***
P1	7.19	73.21	115.57	74.14	73.61	5.36	65.46	26.04	34.84	4.10	1.23	2.30	1.25	0.11
P2	8.79	76.11	119.93	76.43	69.04	5.27	65.69	26.14	34.42	4.09	1.21	2.34	1.19	0.00
P1vsP2	↑	↑	↑	↑	↓	↓	↑	↑	↓	↓	↓H	↑	↓	↓

Table 2. Improvement score (Total) between P1 and P2, showing only for three patients.

	ET	***SF36***	***SBP***	***DBP***	***HR***	***FBS***	***WT***	***BMI***	***WC***	***TC***	***HDL***	***LDL***	***TG***	***SS***	***Total***
1	1	0	1	1	1	0	0	0	1	1	0	1	0	0	7
2	1	1	0	1	1	0	0	0	1	0	1	1	0	1	8
3	1	0	1	0	0	0	0	0	0	1	1	1	1	0	6

Clustering P1 and P2 with all parameters

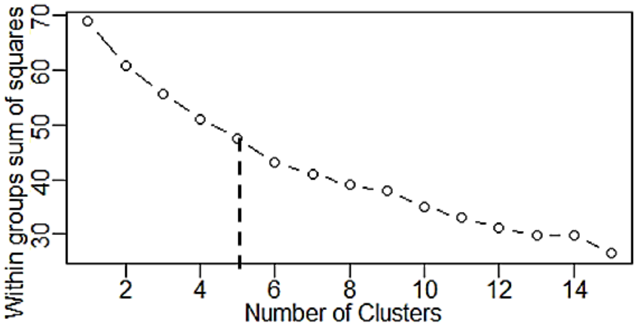

Figure 2. Scree plot of WSS against number of clusters

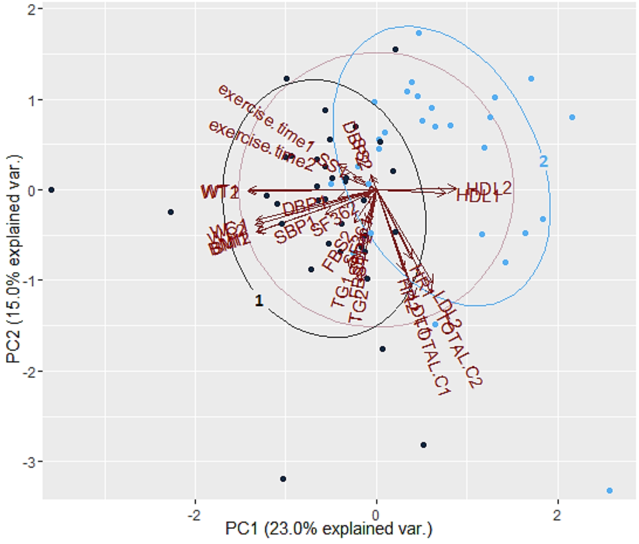

Figure 3. Two clusters found from P1 and P2 data

Based on the scree plot in ***Figure 2***, there appears to be an "elbow" at k=5. Thus, we experimented using k=2,3,4 and 5. Using WSS measure and PCA plots, we found that k=2 produced the most stable clusters across the 10 runs.

Figure 3 shows the cluster plot of the two clusters. Based on the arrows, it appears that cluster 2 is highly characterised by high HDL values while those in cluster 1 by WT and TG. These trends are consistent with the cluster centres tabulated in ***Table 1***, indicated with 'H'. However, what was not observable from the PCA plot is that patients belonging to cluster 1 has more favourable outcomes with lower (↓) SBP, DBP, BMI and LDL in P2, highlighted in green.

To evaluate the clinical associations of the clusters, scores were given for parameters that improved between P1 and P2, as shown in ***Table 2***. In ***Table 3***, it can be observed that there are more patients with total scores above 9 in cluster 1 than 2.

Table 3. Total (T.) scores of patients in cluster 1 and 2.

	T.Scores	***2***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***
Cluster	**1**	0	0	3	6	4	5	9	3	4	1
	2	1	1	2	9	7	6	0	1	0	1

Table 4. Phase comparisons for cluster centres 1 and 2 based on 3 phases.

Clus1	***ET***	***SF36***	***SBP***	***DBP***	***HR***	***FBS***	***WT***	***BMI***	***WC***	***TC***	***HDL***	***LDL***	***TG***	***SS***
P1	6.43	65.71	126.29	78.57	79.71	6.83	66.44	27.86	35.96	5.05	1.18	3.11	1.69	0.14
P2	8.46	70.00	125.71	72.14	75.71	6.79	66.96	27.90	35.45	5.17	1.17	3.07	2.00	0.00
P3	7.20	69.79	130.71	80.57	77.07	7.99	67.16	28.14	36.18	5.51	1.21	3.51	1.72	0.00
vs	↑↓	↑↓	↓↑ H	↓↑	↓↑ H	↓↑ H	↑↑	↑↑	↓↑	↑↑ H	↓↑ H	↓↑ H	↑↓ H	↓↓
Clus2	***ET***	***SF36***	***SBP***	***DBP***	***HR***	***FBS***	***WT***	***BMI***	***WC***	***TC***	***HDL***	***LDL***	***TG***	***SS***
P1	8.17	72.69	124.61	78.16	70.98	5.43	73.61	28.30	37.05	3.89	1.12	2.13	1.41	0.43
P2	9.66	79.39	120.16	76.53	67.33	5.28	73.64	28.29	36.55	3.61	1.10	1.95	1.24	0.04
P3	8.83	78.82	123.00	78.18	69.33	5.77	74.34	29.29	37.24	3.80	1.19	2.07	1.31	0.04
vs	↑↓ H	↑↓ H	↓↑ L	↓↑	↓↑ L	↓↑ L	↑↑ H	↓↑ H	↓↑ H	↓↑ L	↓↑	↓↑ L	↓↑ L	↓=

Table 5. Total (T.) scores of patients in cluster 1 and 2 based on P1,P2 and P3 data.

	T.Scores	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***13***	***14***	***15***	***16***	***17***
Cluster	**1**	0	0	0	2	4	2	1	1	1	2	1	0
	2	1	1	1	4	6	7	9	6	3	6	4	1

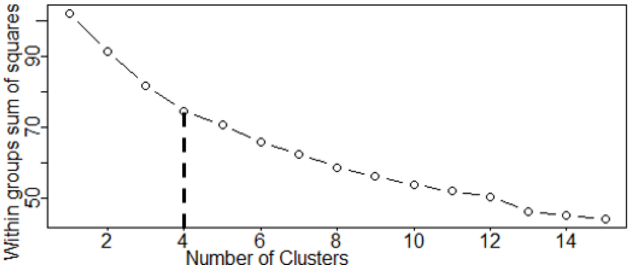

Figure 4. Scree plot of WSS against number of clusters.

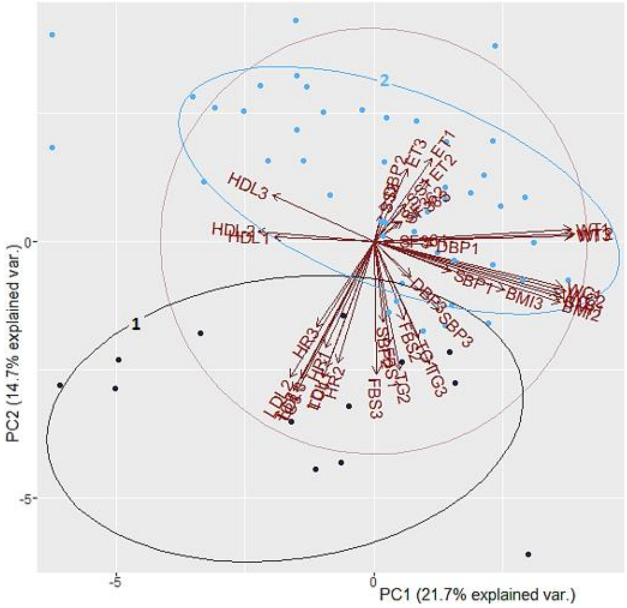

Figure 5. Two clusters found from P1, P2 and P3 data.

A Cramer's V coefficient of 0.522 was found between the clusters and total scores, suggesting relatively strong association.[3,4]

Clustering three phases with all parameters

From ***Figure 4***, the "elbow" is observed at k=4. We experimented with k=2, 3 and 4. The most stable clusters are generated at k=2.

Figure 5 shows the PCA plots for the two clusters found. Note that three points are not included to provide a clearer view of the parameters. We observed that cluster 1 is characterised by high LDL, HR, FBS and TG while cluster 2 by high ET, WT, WC, BMI and HDL. Together with higher SF36 and lower SBP, HR, FBS, TC, LDL and TG highlighted in green in ***Table 4***, it appears to suggest that patients in cluster 2 are fitter despite higher WT and BMI.

The two clusters are more well-separated than those in ***Figure 3***, where the two clusters overlap even though they are more compact, as shown in ***Table 6***. ***Table 6*** shows all the total WSS and its frequency for all 10 runs in both experiments. Clusters generated from clustering the three phases are more stable with 4 unique solutions, as opposed to 5 unique solutions from clustering two phases.

The clusters were found to be have moderate association with a Cramer's V coefficient of 0.27 with the improvement scores. The improvement scores were calculated based on improvement found between P1 and P2, and between P2 and P3.

The clinical parameter – total improvement score pair has a nonlinear and nonmonotonic relationship. This is shown in ***Table*** *7*, indicated by a low Pearson's (P) correlation with high Hoeffding's D (H) correlation whereas both low P and H values indicate random variables.[5]

This indicates that Kmeans was able to discover the hidden structures associated with improvement using a collective measure across the three phases, the improvement scores, which are not directly observable. While such structures were found, the clusters themselves (shown in ***Figure 3*** and ***Figure 5***) do not directly demonstrate improvement or otherwise.

Scatter plots were drawn to investigate in parameters with low P but medium H such as FBS in P1 (FBS1), illustrated in ***Figure 6***, as well as parameters with complete dependence such as SBP in P2 (SBP2) ***Figure 7*** using the `scatter.smooth` function in **R**.[6] This function also add a smooth curve computed by loess. Based on these two plots, there appears to be no strong correlation. Further work beyond the scope of this paper is required.

Based on the outcomes of experimental results, we consider this work to be promising as clusters with clinical association were found using a simple raw-data-based Kmeans clustering framework.

Indeed, high correlation is found between some of the parameters. Yet the removal of these parameters mean that the parameter for that phase is not represented. For now, we include all parameters in the clustering despite the low percentage variability represented in the two principal components in ***Figure 3*** and ***Figure 5***, which does not give a visual representation of high accuracy. In the future studies, we hope to explore other techniques to find relevant parameters as well as to determine the trajectories (groupings) within the longitudinal data.

Table 6. Total (T.) WSS (based on normalised values) and its (freq)uency for both clustering experiments.

	P1 &2 (k=2)		***P1,2 & 3 (k=2)***	
	T. WSS	***freq***	***T. WSS***	***freq***
1	60.95	5	91.65	1
2	62.60	1	90.12	3
3	61.58	2	90.09	5
4	61.98	1	91.32	1
5	62.47	1		

Table 7. Pearson's (P) and Hoeffding's D (H) correlation between Total Score with each parameters in the 3 phases.

	P1		***P2***		***P3***	
	P	***H***	***P***	***H***	***P***	***H***
ET	0.23	0.19	0.13	0.83	0.27	0.16
SF36	-0.09	0.95	0.08	0.52	0.09	0.68
SBP	0.08	0.58	-0.06	1	-0.23	0.01
DBP	0.17	0.47	-0.09	1	-0.13	1
HR	-0.17	0.23	-0.14	0.24	-0.22	0.06
FBS	0.09	0.49	-0.11	1	-0.18	0.49
WT	0.04	0.64	0.01	0.45	-0.07	0.46
BMI	0.05	0.8	-0.02	0.85	-0.05	0.52
WC	0.02	0.68	-0.02	0.66	-0.16	0.47
TC	0.23	0.25	-0.02	0.47	-0.13	0.03
HDL	-0.02	0.6	-0.04	0.66	0.17	0.83
LDL	0.20	0.35	-0.07	0.73	-0.12	0.1
TG	0.21	0.44	0.09	0.63	-0.15	0.13
SS	0.21	0.41	-0.31	1	-0.31	1

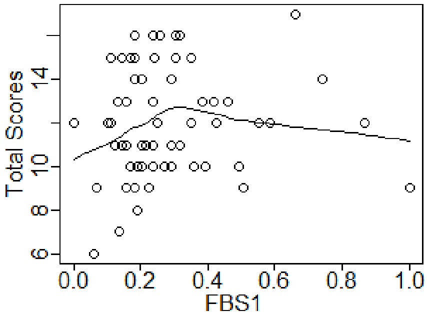

Figure 6. Scatter plot of Total Score against FBS1 (normalised).

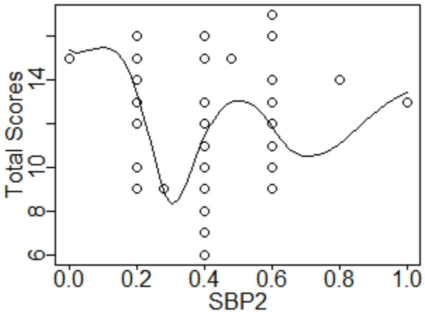

Figure 7. Scatter plot of Total Score against FBS1 (normalised).

4. Conclusion

As a preliminary study, we have experimented using a simple Kmeans clustering framework to discover clinically relevant clusters in the cardiac rehabilitation data, which contains patient data repeatedly collected from 3 different phases. Highly, clinically associated clusters were found using P1 and P2 data while moderately clinically associated clusters were found using data from all three phases. This suggests for further cluster refinement approaches to be applied, as well as exploration into other approaches such as model-based techniques, as well as application of suitable distance metric that could better model the changes across the phases, all of which, so far, has not been explored for this dataset.

Acknowledgements

This work was supported by the Universiti Brunei Darussalam under Grant UBD/PNC2/2/RG/1(311).

References

[1] T. W. Liao, *Pattern Recognition 38*, **2005**, 1857.

[2] B. Heggeseth, *Longitudinal Cluster Analysis with Applications to Growth Trajectories*, 2013.

[3] L. M. Rea and R. A. Parker, *Designing and conducting survey research*, **1992**, 203.

[4] J. W. Kotrlik, H. A. Williams and M. K. Jabor, *Journal of Agricultural Education 52.1*, **2011**, 132.

[5] D. Bhalla, http://www.listendata.com/2015/03/ detect-non-linear-and-non-monotonic.html, Last accessed: 2nd May 2016.

[6] R Core Team, *R: A language and environment for statistical computing*. https://www.R-project.org/, **2016**. Last accessed: 3rd May 2016.