CFAS420 CW2

35219427

April 23, 2020

1. Introduction:

The quality of life measures of patients are significant factors to help doctors infer patient conditions. And since these measurements are easily accessible, it is very helpful for limited medical resources. The purpose of this report is to examine whether it is possible to divided patients into different groups according to their quality of life variables and whether clustering is helpful for understanding issues facing the patients in the hospital.

2. Data Pre-Processing:

377 samples about the quality of life variables of hospital patients construct the original dataset. Because the study only needs the quality of life features, therefore the ID of each sample and three additional variables: Sex, Age and Relationship were removed from the dataset. All cases with missing value -9 were removed, therefore 299 cases with 22 features constructed a new data frame for the study. Since all quality of life variables are on the same scales, thus the process of standardizing can be omitted.

3. Kmeans:

Kmeans is partitioning methods with distanced-based of clustering algorithm. Samples are assigned into k clusters. The means value of the objects in the cluster is defined as the centroid of one cluster. There are k points was selected randomly as the center of each cluster, kmeans use Euclidean distance to calculate the distance between each object with each cluster center in this case, the data point is arranged into the most similar cluster. Then each cluster calculates the new mean and use it as the new center of cluster, and assign all points again. The above iterations process would continue until no more samples are reassigned to another clusters.

A good result for kmeans is that samples in one cluster should be more similar, and samples in different clusters should be more different. The total within-cluster sum of squares can be the criterion of find best model in this study, it is “tot.withiness” in output of kmeans function in R. In order to get best solution, each procedure goes through 200 times iterations with different starting value from 1 to 100 in this study.

Figure 5 in appendix shows that the kmeans result of this study. Because there are 22 variables, in order to shows the result in 2D, fviz function of R perform PCA and show the data point based on two first principle components. The optimal result of kmeans with minimum total within sums of squares, tot.withiness=3404.28. Figure 1 shows 4 trajectories of 22 variables’ mean value of 4 clusters by kmeans algorithm. As seen in Figure 1 with the output of kmeans model, 42 objects belong to cluster1, they have highest mean value of all life variables except diarrhea among 4 clusters, the mean value of work and hobby are close to 4, it means most patients have quite a bit or very much quality for most of life variables in cluster 1.96 patients belong to cluster 2, Most people who have quite a bit of work, hobby, trouble sleeping, tired and illness or medication interfered with social activities. And in cluster 2, the mean value of vomit dropped obviously compare with cluster 1. The cluster 3 has 61 objects, all patients’ measures are less than cluster1 but they also have a little or quite a bit symptom of each variables. 100 patients belong to cluster 4, the mean values of all life features less than 2, patients don’t have most qualities at all in cluster 4.

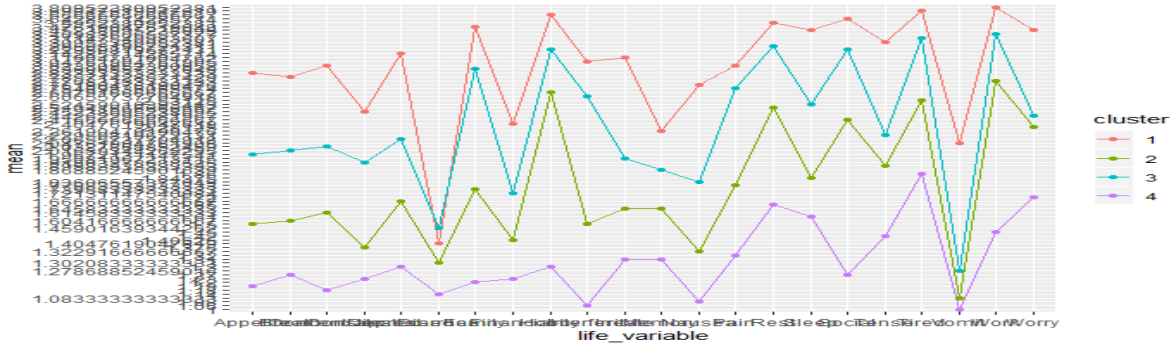

Figure 1: Trajectories of 22 life variables' mean value of 4 clusters(Kmeans)

4. PAM:

Partitioning around medoids (PAM) algorithm is another partition method based on distance of clustering. PAM doesn't use mean value to determine center of cluster. PAM also select k objects as initial medoids(centers) randomly in the beginning, and other objects are assigned to the cluster with nearest medoids objects based on distance. Then one non-medoid is selected randomly except current medoid objects. If the cost of using the non-medoid object replace previous medoid is less than 0, this non-medoid object becomes new center of cluster. Above process would repeat until no replacement happing. The purpose of PAM is to find k medoids which minimize the sum of the dissimilarities of the objects to their closest medoid object.

Figure 6 in appendix shows the clustering result by PAM with K=4. It is obviously that the result of PAM is worse than Kmeans, existing more overlapping among all clusters. The output of PAM model shows that 52 patients in cluster 1 have quite a bit or very much of work limitations. And the qualities of hobby and rest are highest among all clusters. 107 patients in cluster 2. The mean value of work, hobby, rest, tired and social greater than 3. But compare with cluster 1, some patients don't have these conditions at all (minimum value=1 for each variable).87 patients are assigned to cluster 3. All mean values of life variables are less than or equal to 2. Most patients have a little quality for each life variable in this cluster. 53 patients are assigned to cluster 4. No variable has a mean value than 1.6. Most patients have no quality at all for all life variables in cluster 4.

Average silhouette width is used to find optimal k value of PAM model. Silhouette coefficient is an approach of intrinsic methods. According to the quality of clusters are separated and how compact the clusters are, average silhouette able to get best amount of clusters K. Figure 9 in appendix shows the output of silhouette coefficient for the PAM, the K=2 is the optimal number.

Figure 7 in appendix shows PAM clustering with K=2. 93 patients belong to cluster 1. Most patients have high quality for each life variable in cluster 1 except vomit. The mean value of every life variable is greater than 2. 206 patients belong to cluster 2. Most patients have slight performance of each life variable in cluster 2.

5. GMM:

A limitation of partitioning methods is that each object only belongs to one cluster and there is no probability or uncertainty to express the quality of an object is associated with each cluster. Gaussian mixture model (GMM) is an algorithm which used to find the mixed representation of the probability distribution of multiple Gaussian models. GMM is comprised of multiple Gaussians, each Gaussian as a component or cluster for samples.

Figure 2 shows the PMM clustering with 4 Gaussians by Mclust in R. 49 patients belong to cluster 1. Except diarrhea, the mean values of all 21 variables are greater than 2. 56 patients belong to cluster 2, the qualities of

their symptoms are less than cluster 1, but the mean values of most variables still greater than 2. 98 patients belong to cluster 3, it is obviously that most people have these qualities a little or don't have them at all, except work, hobby, rest and tired, their mean value still greater than 2. 96 patients belong to cluster 4, all mean values are less than 2, most of patients don't have any symptom in it.

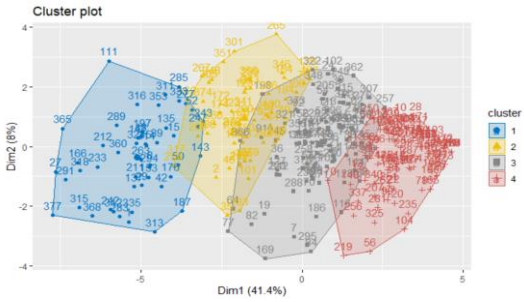

Figure 2: GMM output with K=4, fviz_cluster function perform principal component analysis (PCA) and show the data point based on the first 2 principal components.

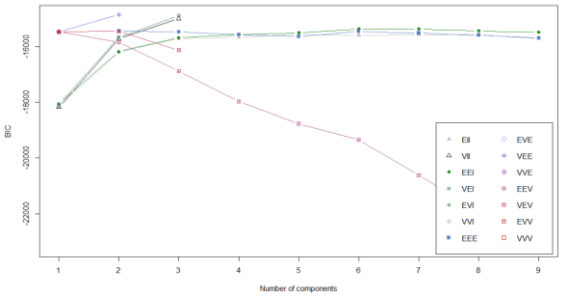

Figure 3: Mclust offer 14 models in GMM for clustering, Different model with different clustering result. The model with highest BIC is used for output. VEE model is selected in this case.

In order to select the number of clusters, Mclust assumes that data is a sampling result of one or more mixed Gaussian distributions. BIC was used in this study to find optimal model. As seen in Figure 3, Mclust outputs performances of the 14 models of 9 groups, and Mclust select the model and the number of clusters G with lowest BIC. Maximum likelihood estimate with EM algorithm are used to select optimal number. VEE model with G=2 is selected in this study. Figure 8 in appendix shows the final model of GMM algorithm with 2 clusters. GMM output is valuable result in this case. The advantage of GMM is that object is assigned to clusters based on probability, each object could belong to multiple clusters with different probability. There is more information in GMM output, such as confidence, uncertainty, mixing probabilities, mean and variance of each component, etc. All of them help us make decisions in reality.

6. LCA:

Latent class analysis (LCA) is a statistic approach about Finite Mixture Model, LCA is a based-model clustering method where construction is created and identified from latent categorical variables, these latent variables based on individual responses from multivariate categorical samples. In this study, all life variables are converted to categorical variables. Table 1 in appendix shows the log-likelihood, BIC and AIC of LCA with different number of clusters K. The model of K=3 is the suggestion of LCA based on BIC in this case.

Figure 5 shows the clustering of LCA with 3 clusters. The most patients in cluster 1 have work, hobby, tired, family, social quality very much. Only a few patients have symptom of diarrhea in cluster 1. Thus, the patients' conditions in cluster 1 are worst. Cluster 2 consists of most persons who don't have the quality of vomited, constipated, diarrhea, financial difficulties at all. Tired, work limitations and hobby limitations are still patients matters, but the quality of them is better than cluster 1. Most of patients have other quality a little or quite a bit. Thus, the patients' conditions in cluster 2 are middle. Cluster 3 consists of most persons who don't have other symptoms quality at all, except tired and trouble sleeping. Thus, the patients' conditions in cluster 3 are best.

Including covariates to achieve clustering is an advantage of LCA, covariates is helpful for perform clustering conditional based on additional features. In this case, gender of patient, age of patient and patient's relationship were used as covariates in LCA. Figure 10 in appendix shows the output with 3 covariates. Table 2 in appendix

shows the p value and t value of LCA with 3 covariates separately. It is obviously that all t value less than 1 and all p value greater than 0.05. Therefore, we don't have enough evidence to reject the null hypothesis, 3 covariates are not significant in this study.

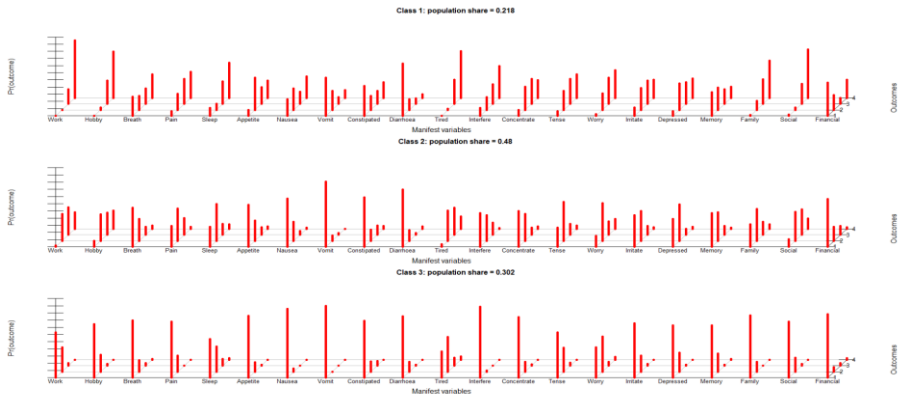

Figure 5: LCA output with 3 clusters. 'Mixing" proportions denote population share of observations response. The x axis denote variables in model, the y axis denote the probability of sample's life variable in each cluster.

7. Conclusion:

Above results come from 4 different clustering algorithms, according to their outputs, there are several points:

- According to the methods of select K and the clustering result. K=2 is the best selection in this case, samples should be divided into 2 clusters. One cluster include patients who have high quality of all variables, especially work limitations and hobby limitations. Another cluster includes patients who have a little quality or don't have quality at all for most life variables. In LCA model, based on BIC, we have K=3. The patients were divided into one group with high quality of all life variables, another group with most patients don't any symptom of all variables or slightly symptom. Third group with patients have middle quality comparing with above two clusters. This result is valuable, because the result of K=2 would ignore some detail about patients, such as the patient who have several qualities a little or quite a bit was assigned into the second cluster, some patients in poor conditions would be ignored. Thus, if we want to treat patients with different conditions detailed, K=3 is a better option than K=2, but it's up to the doctor's demand in reality.
- Most patients have symptom of work and hobby limitations, the symptom of most of them are serious. And most patients suffer from tired and need to rest.
- Only a few patients have symptom of diarrhea, the output of diarrhea variable from each cluster is same. Therefore, the patients who have diarrhea quite a bit or seriously could be extracted and analyzed separately.
- According to the outputs, model based clustering methods are suitable in this study. As we mentioned above, GMM and LCA is not hard clustering, probability is used to evaluate the assignment of each point, and more information could be used to make decision.

Reference:

1. Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.
2. https://www.jamleecute.com/partitional-clustering-kmeans-kmedoid/
3. https://towardsdatascience.com/gaussian-mixture-models-explained-6986aaf5a95
4. http://xuzhougeng.top/archives/Cluster-data-with-mclust

Appendix:

Figure 5: Kmeans output with K=4,, fviz_cluster function perform (PCA) 2 principal components that explain

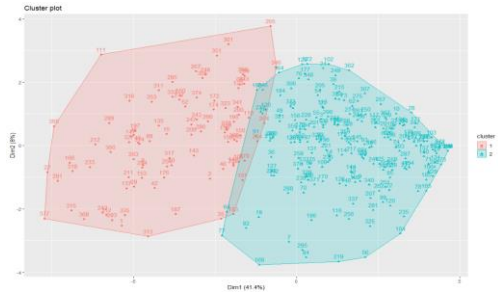

Figure 7: PAM output with K=2.

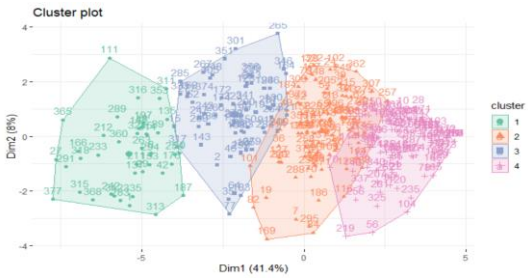

Figure 6 PAM output with K=4. fviz_cluster perform PCA 2 principal components that explain

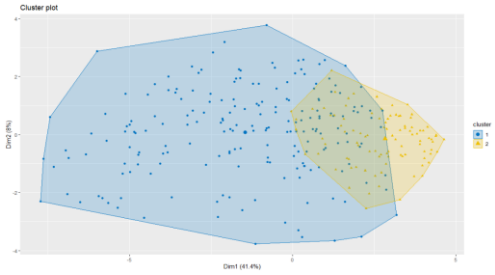

Figure 8: GMM output with K=2.

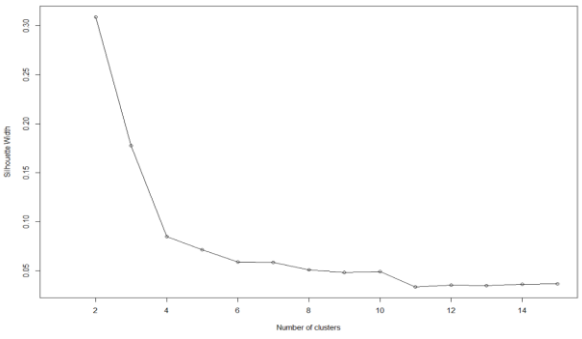

Figure 9: Output of silhouette width of PAM model.

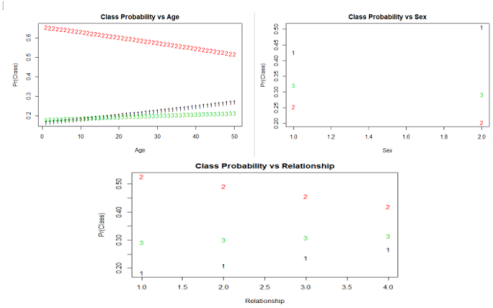

Figure 10: Covariates in LCA. Age as covariate(left), Sex as covariate(right), Relationship as covariate(below)

K	AIC	BIC	Likelihood
1	14123.95	14355.59	-6988.975
2	12622.35	13089.30	-6184.174
3	12214.11	12916.37	-5916.057
4	12122.98	13060.55	-5806.489

Table 1: AIC, BIC, Log-Likelihood of different number of clusters in LCA model

Covariate	Coefficient	t-value	p-value
Age	0.00653	0.193	0.848
Sex	-0.400	-0.551	0.583
Relation	0.099	0.229	0.819

Table 2: Coefficient, t-value, p-value of different covariate in LCA model