World Scientific Research Journal Volume 7 Issue 5, 2021
ISSN: 2472-3703 DOI: 10.6911/WSRJ.202105_7(5).0009

Cluster Analysis and Realization of Mixed Data

Caili Liang

College of Economics, Jinan University, Guangzhou, 510632, China

Abstract

Mixed data is often generated in daily production and life. In order to improve the efficiency of data mining, it is necessary to establish a cluster analysis method for mixed data. We introduced the details of the calculation method of the comprehensive distance for the mixed data, the choice of the number of clusters and the choice of the clustering method. In the empirical analysis, we select the mixed data set, use "gower" distance function to calculate the distance of the mixed data first, then select the appropriate number of clusters according to the size of the average silhouette width, finally use the PAM and CLARA algorithm to realize the cluster analysis of the mixed data. We find that the clustering results of PAM algorithm and CLARA algorithm are different, and the clustering results of CLARA algorithm perform better.

Keywords

Mixed Data; Cluster Analysis; PAM algorithm; CLARA algorithm; R.

1. INTRODUCTION

Cluster analysis has developed rapidly in recent years. It is not only a common technique for data analysis, but also an important part of data mining. Cluster analysis technology is applied in the fields of machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression and computer graphics. The basic idea of cluster analysis is that, gather similar objects in the same cluster, and disperse dissimilar objects in different clusters [1]. The results of cluster analysis can reveal the distribution of samples, then dig out potentially useful information. Since the data objects we encounter are often mixed data sets, which contains continuous variables, nominal variables, and sequential variables. For example, the browsing status of a website may include fields such as gender, age, occupation, and the number of pages viewed. If cluster analysis can be performed on these mixed data sets, then important information can be mined based on the characteristics of different clusters.

The clustering analysis of mixed data sets mainly adopts the partition method, which includes PAM algorithm, CLARA algorithm, CLARANS algorithm and K-prototypes algorithm. As we all know, PAM algorithm and CLARA algorithm are two classic K-medoids algorithms. Compared with the K-means algorithm, the PAM and CLARA algorithms randomly select k sample objects as the clustering centers, which makes them more robust [2]. In addition, the K-means algorithm is only applicable to continuous variable sets, but the PAM algorithm and CLARA algorithm can convert nominal variables and sequential variables into continuous variables, and then perform cluster analysis.

Since some classic algorithms, such as the K-means algorithm, can only deal with continuous data. However, in today's data explosion, many data sets contain both continuous variables and categorical variables, showing mixed characteristics. Convert categorical variables into continuous variables, from K-means algorithm to K-medoids algorithm, which could handle mixed data sets.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202105_7(5).0009

In 1987, Kaufman Rousseevw proposed the PAM (Partitioning Around Methods) algorithm, which is the earliest and most classic K-medoids algorithm [3]. PAM algorithm can do cluster analysis on mixed data sets. Because the PAM algorithm is only suitable for small data sets, and does not have scalability for large data sets, scholars at home and abroad have been committed to improve the performance of the PAM clustering algorithm. In 1990, Kaufman Rousseevw proposed the CLARA (Clustering LARger Application) algorithm based on the PAM algorithm [4]. This algorithm extracts a small part of the samples from the original large data set, and then adopts the PAM algorithm. In 1994, R.Ng and J.Han proposed CLARANS (Clustering Larger Application based upon RANdomized Search), which combines sampling technology and PAM algorithm [5]. CLARANS algorithm is an improvement on the clustering quality and scalability compared to CLARA algorithm. In 1999, Huang et al. combined the K-modes algorithm with the K-means algorithm, proposed the K-prototypes algorithm [6]. It uses the frequency method to select cluster centers, then apply cluster analysis to mixed data sets composed of sequential variables, nominal variables and continuous variables. Subsequently, the fuzzy K-prototypes algorithm was proposed based on the K-prototypes algorithm.

There are some limitations of the K-medoids algorithm. (a) The scalability of large data sets is poor, or even impossible to handle. (b) Only numeric fields can be clustered, and non-numeric fields must be converted into numeric fields. (c) The number of clusters to be generated must be given in advance, which is more sensitive to initial conditions. (d) Only spherical clusters can be found. (e) Very sensitive to the input sequence of data.

2. METHODOLOGY

2.1. Mixed Data

In actual databases, one table is often composed of fields of different data types, and each field records different information. The mixed data set contains two or more data types among continuous variables, ordered categorical variables, symmetric binary categorical variables, asymmetric binary categorical variables, and unordered categorical variables. For example, in the website visit information table, “the time spent on the website” is a continuous variable, “the number of pages visited” is a discrete variable, “the user level” is a multivariate ordinal categorical variable, “gender” and “domestic/foreign users” are symmetric binary categorical variables, “VIP/non-VIP” and “the completion of the first purchase” are asymmetric binary categorical variable, and “occupation” is a multiple non-categorical variable. This is a relatively complex mixed data set, which contains almost all data types. If cluster analysis can be performed on such a mixed data set, the user portrait of the website can be captured, and then targeted advertising can be invested, finally website revenue will increase.

2.2. Distance Matrix of the Mixed Data

Since mixed data mainly includes continuous variables and categorical variables, when calculate the comprehensive distance of mixed data sets, we can calculate the distances of continuous variables and categorical variables separately, then calculate the comprehensive distance by weighted average. That is to say, we can map all p variables to the range of [0, 1], and then weighted average.

Assuming that there are p attributes in a mixed data set, and each attribute is a variable, the distance between two samples is

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}} \quad (1)$$

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202105_7(5).0009

If there is no observations of variable f between two samples, that is to say, x_{if} or x_{jf} is missing, or when the variable f is an asymmetric binary categorical variable and x_{if}=x_{jf}=0, then the weight item $\delta_{ij}^{(f)} = 0$. Otherwise $\delta_{ij}^{(f)} = 1$. We find that, the distance between two samples about the variable f is related to the field type of itself.

(a) When the variable f is a binary or disordered categorical variable and $x_{if} = x_{jf}$, then ${d_{ij}}^{(f)} = 0$. Otherwise, ${d_{ij}}^{(f)} = 1$.

(b) When the variable f is a continuous variable,

$$d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{\max(x_{if}) - \min(x_{if})} \tag{2}$$

(c) When the variable f is a sequential categorical variable, we convert the data object into a continuous variable on [0,1] first, the conversion formula is

$$Z_k = \frac{x_k - 1}{M_k - 1} \tag{3}$$

M_k is the maximum value of the x_k. Next, we treat it as a continuous variable as formula (2).

After calculating the distance of the mixed data object according to the variable types, bring into the distance formula and weighted average, then obtain the comprehensive distance of the mixed data set. With the comprehensive distance, it is possible to do cluster analysis on mixed data.

2.3. K-medoids Algorithm

The K-medoids algorithm is a classic partition algorithm, which divides the n samples in the data set into k cluster. K-medoids algorithm process as follows.

(a) Randomly select k samples as the initial cluster centers (O_1, O_2, ..., O_k).

(b) Calculate the comprehensive distance between the remaining samples and each center point, and assign them to the cluster closest to them.

(c) Randomly select a non-center sample and calculate the exchange cost between the non-center sample and the current center sample. If the exchange cost is small, swap the non-center point with the center point.

(d) If each cluster does not change, stop clustering.

For small data sets, we can directly use the PAM algorithm. But, when deal with large data sets, we should use the CLARA algorithm. The "small" and "large" here are mainly determined by the computer's memory and speed. When the CLARA algorithm was proposed in 1990, a "small" data set meant that the sample size was less than 100. With the growing of clustering algorithm, in 1997, when the sample size was less than 200, it was regarded as a "small" data set. By 2006, the PAM algorithm could deal with thousands of observations.

The CLARA algorithm is an improvement of the PAM algorithm, can handle larger data sets. The CLARA algorithm does not consider the entire samples in the data set, but selects a small part of the samples as the best sample for clustering, which greatly reduces the calculation time and storage requirements. Due to the large sample size, a small sample randomly selected can represent the entire data set.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202105_7(5).0009

3. EMPIRICAL ANALYSIS

3.1. Data Selection and Exploratory Analysis

3.1.1 Data Introduction

We select the Byar data set in the R package clusterMD. Byar is a mixed data set collected from a group of prostate cancer patients with stage 3 or 4 prostate cancer. The data set is a data frame containing 475 observations of the following 15 variables, see Table 1 for some details.

Table 1. The information of the input samples

Variable Name	Variable Value	Variable Type
Age	[48,89]	continuous
Weight	[69,152]	continuous
Performance Rating	0-normal activity, 1-time in bed is less than 50% of the day, 2-time in bed is more than 50% of the day, 3-only Lie down on the bed	ordinal categorical
Cardiovascular Disease History	0-no, 1-yes	binary categorical
Systolic Blood Pressure	[8,30], in units of 10	continuous
Diastolic Blood Pressure	[4,18], in units of 10	continuous
Electrocardiogram Code	0-normal, 1-benign, 2-rhythmia and electrolyte changes, 3-arrhythmia or conduction defect, 4-hardness of the heart, 5-senile myocardial infarction, 6-recent Myocardial infarction	multivariate categorical
Serum Hemoglobin	[59,182],g/100ml	continuous
Size of Primary Tumor	[0,69],cm	continuous
Index of Tumor stage and Historic Grade	[5,15]	continuous
Serum Prostatic Acid Phosphatase	[1,9999], King-Armstong unit	continuous
Bone Metastases	0-no, 1-yes	binary categorical
Stage	3 or 4	nominal
Observation	1,2,...,475	discrete
Survival Status	0-alive, 1-died from prostate cancer, 2-died from heart or vascular disease, 3-died from cerebrovascular accident, 4-died from Lung cancer, 5-died from other cancers, 6-died from respiratory diseases, 7-died from other specific non-cancer factors, 8-died from other non-specific non-cancer causes, 9-died from unknown factors	multivariate categorical

3.1.2 Preprocessing

Before calculate the comprehensive distance, the multivariate categorical variables of the Byar are converted into a factor form. The mutate function in the dplyr package can add or

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202105_7(5).0009

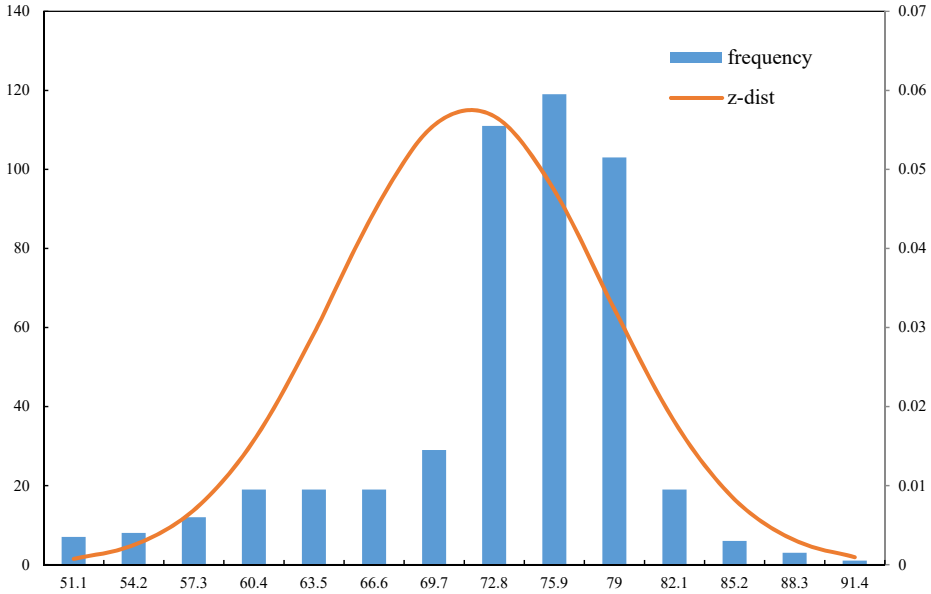
modify variables in the data frame. Here, the variable “Observation” is placed in the first column to distinguish different patients, and other variables are placed according to variable types.

3.1.3 Exploratory Analysis

We do exploratory data analysis first with the cleaned Byar. Analyzing the central tendency of factorial and continuous variables, preliminarily judges the composition of these "clusters".

(1) Summary by Age

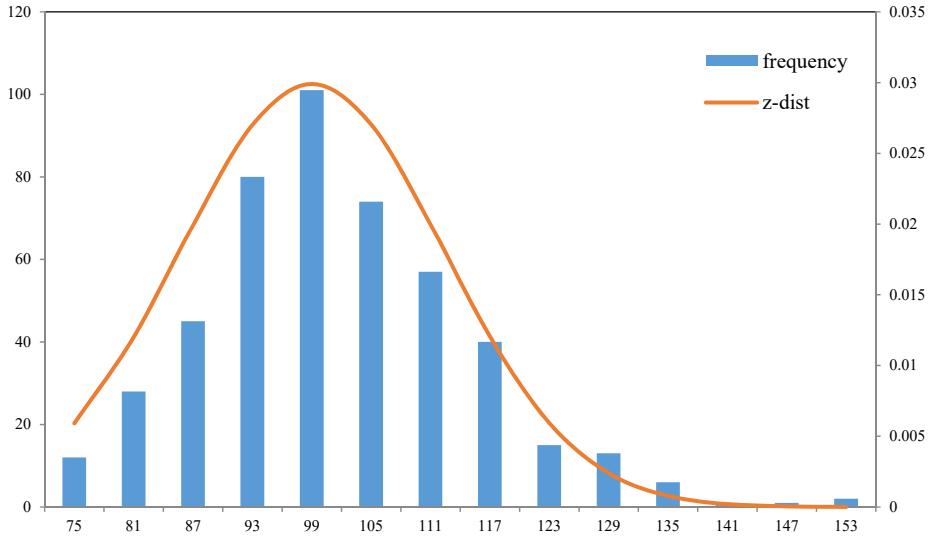
First find out the minimum age and maximum age among these patients, and then divide them into 14 groups according to "Age". Then draw a frequency distribution graph. Figure 1 reports the age distribution of samples. We find that these patients are concentrated in the 70-80 years old. In other words, most prostate cancer patients in stage 3 or 4 are 70-80 years old, and there are fewer patients under 50 or over 80 years old.



Figures 1. Age distribution of patients

(2) Summary by Weight

The lightest weight is 69 pounds, the heaviest is 152 pounds. Divide “Weight” into 14 groups, then draw a bar graph for these 14 groups. Figure 2 reports the weight distribution of patients. We find that the weight of the survey subjects is mainly distributed between 80-120 pounds.



Figures 2. Weight distribution of patients

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202105_7(5).0009

(3) Cross-analysis by Performance Rating and Stage

Draw a cross-list based on "Performance Rating" and "Stage", and the results are reported in Table 2. We find that regardless of Stage 3 or Stage 4, most patients can move normally, and the normal performance rate of Stage 3 is 7.8% higher than that of Stage 4. Conversely, Stage 4's activity restriction at each rating is higher than Stage 3. In other words, Stage 4 is more ill than Stage 3.

Table 2. Cross-List of Performance Rating and Stage

	Stage			
Performance Rating	3		4	
Normal activity (0)	255	93.4%	173	85.6%
Activity time is more than half a day (1)	15	5.5%	17	8.4%
Activity time is less than half a day (2)	3	1.1%	10	5.0%
Only Lie down on the bed (3)	0	0%	2	1%
Total	273	100%	202	100%

(4) Cross-analysis by Survival Status and Stage

Draw a cross-list based on "Survival Status" and "Stage", and the results are reported in Table 3. We found that 33% of Stage 3 survived, which was 10% higher than Stage 4. Among the dead samples, 11% of Stage 3 died of prostate cancer, which is 34% lower than Stage 4. However, 23% of Stage 3 died of heart and vascular diseases, which is 8% higher than Stage 4. We also found that Stage 3 had the highest probability of dying from heart and vascular diseases, which was 12% higher than that of prostate cancer. However, Stage 4 has the highest probability of dying from prostate cancer, which is 30% higher than that of heart and blood vessel diseases. In other words, prostate cancer is related to heart and blood vessel diseases and can be transformed into each other.

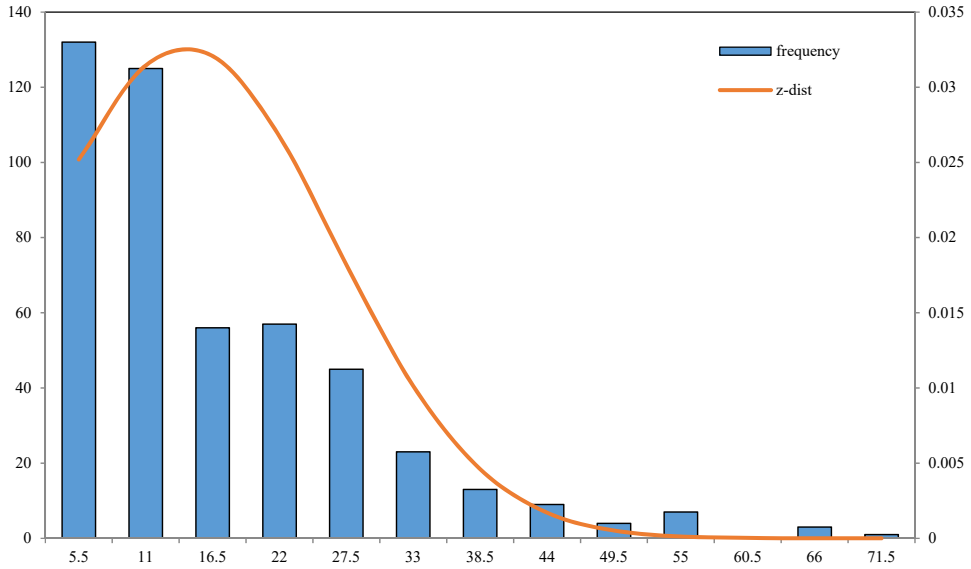
Table 3. Cross table of Survival Status and Stage

	Survival Status									
Stage	0	1	2	3	4	5	6	7	8	9
3	91	31	63	21	10	19	12	18	2	6
	0.33	0.11	0.23	0.08	0.04	0.07	0.04	0.07	0.01	0.02
4	46	90	30	10	4	5	4	9	4	0
	0.23	0.45	0.15	0.05	0.02	0.02	0.02	0.04	0.02	0

(5) Summary by Size of Primary Tumor

"Size of Primary Tumor" is a continuous variable, we draw a histogram (see Figure 3) to judge the concentrated distribution of the size of the patient's primary tumor. The histogram showed that the size of the patient's primary tumor was mainly concentrated between 0-27.5 cm, with a clear right-skewed distribution. Among them, the first two groups (the size of the primary tumor is less than 11.5 cm) accounted for 54%, nearly half. The third to fifth groups (the size of the primary tumor is between 16.5 cm and 27.5 cm) accounted for 31%. That is, the proportion of the top five groups is 85%. Most patients with prostate cancer can find the lesion and treat it in time when the tumor is small, which increases the possibility of survival. The normal distribution of primary tumor size is skewed to the right, which further proves that patients with larger primary tumors account for a minority.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202105_7(5).0009



Figures 3. Size of the primary tumor

3.2. Calculation of Distance Matrix of Byar

Table 4. Descriptive statistics of the "gower" distance

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Metric	Types
0.0202	0.2568	0.3241	0.3256	0.3968	0.6823	mixed	A,N,O,I

Like other clustering methods, the first step of our mixed data cluster analysis is to calculate the distance matrix between samples. Firstly, we map each type of variable to the [0,1] interval, and then use the weighted linear combination method to calculate the final distance matrix. The daisy function in the R package cluster can directly calculate the "gower" distance of Byar.

Table 5. The most(least) similar sample pair

fields	the most similar		the least similar	
Observation	320	11	346	279
Cardiovascular Disease History	0	0	1	0
Bone Metastases	0	0	0	1
Electrocardiogram Code	0	0	4	0
Stage	3	3	3	4
Survival Status	0	0	0	1
Performance Rating	0	0	0	3
Age	71	77	73	59
Weight	87	89	129	99
Systolic Blood Pressure	15	15	21	13
Diastolic Blood Pressure	8	8	11	8
Serum Hemoglobin	156	156	167	127
Size of Primary Tumor	8	3	10	17
Index of Tumor Stage and Historic Grade	8	8	8	13
Serum Prostatic Acid Phosphatase	6	6	7	9999

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202105_7(5).0009

In order to verify the accuracy of the “gower” distance, we extracted the most similar and the least similar sample pairs, they were the two pairs of samples with the shortest distance and the longest distance in the “gower” distance matrix. The extraction results are reported in Table 5. The results showed that patients 11 and 320 had the shortest distance, and the value of each indicator is the same or close. Therefore, the “gower” distance is reasonable.

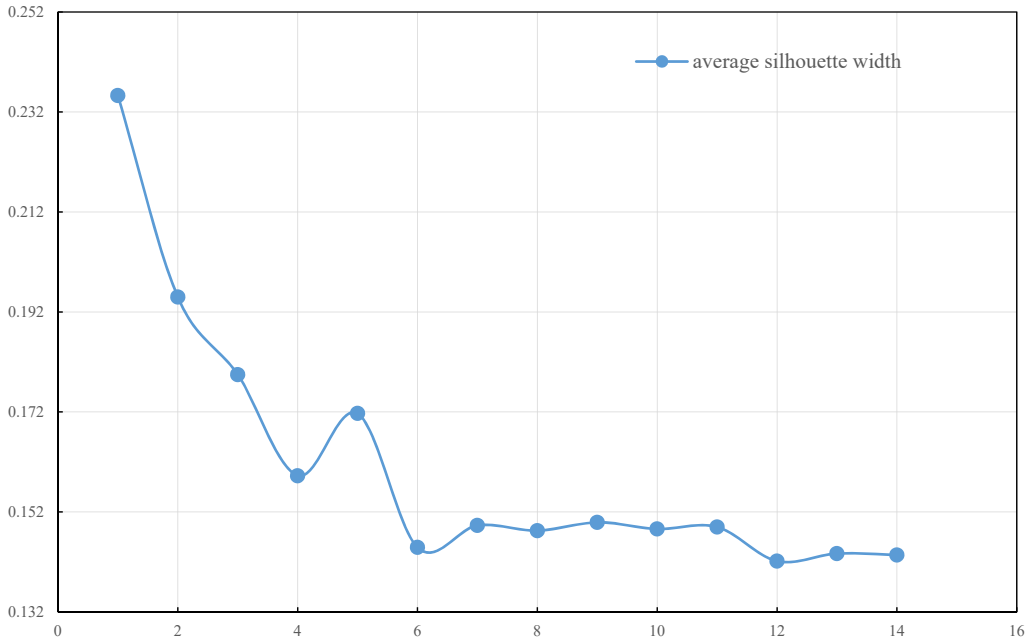
“Observation” is used as a marked field and does not participate in distance calculation. The values of “Serum Prostatic Acid Phosphatase” were quite different, we calculated the distance after taking the logarithm of it. For the binary asymmetric categorical variables “Cardiovascular Disease History” and “Bone Metastases”, we were marked with "asymm". The descriptive statistics of the “gower” distance are reported in Table 4.

The indicators of patients 279 and 346 are the least similar, and the distance between these two samples is the furthest. The observed values of the 14 indicators of the two patients in this pair of samples are not the same, and there were large differences. The distance between these two samples was so large that they were bound to be divided into different clusters.

3.3. PAM and CLARA Cluster Analysis

The second step of PAM and CLARA cluster analysis is to determine the number of clusters k. The number of clusters k is often determined by the average silhouette width, which is an index to measure the dispersion of clusters. Its value is between -1 and 1, the larger the value, the greater the difference between clusters, the better the clustering effect. Determine the best k value by comparing the size of the average silhouette width under different k values.

3.3.1 PAM Cluster Analysis



Figures 4. Silhouette plot of the PAM algorithm

The contour coefficient map drawn by the PAM algorithm is reported in Figure 4. When divided into two categories, the contour coefficient takes the maximum value of 0.235. With the increase of clusters, the contour coefficient gradually decreases and reaches a stable state after k=6. Mentioned earlier, the PAM algorithm selects the best number of clusters according to the maximum average contour width. Therefore, the PAM algorithm should divide these 475 patients into two clusters.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202105_7(5).0009

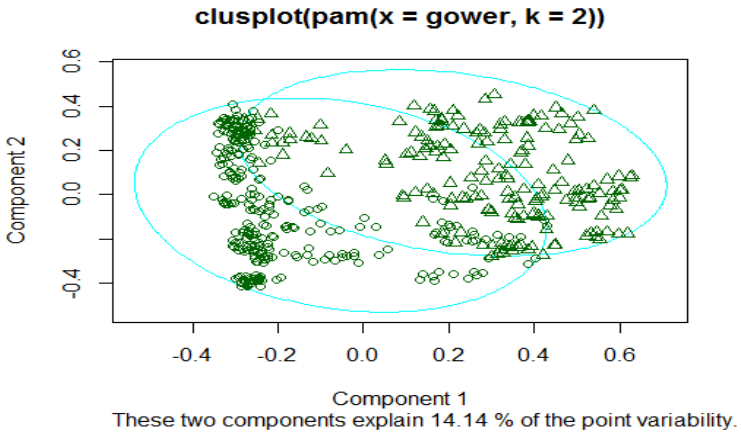


Figure 5. PAM clustering renderings

The clustering results of the PAM algorithm are reported in Figure 5. The first category has 292 samples, and the second category has 183 samples. These two components only explain 14.14% of the total variance. We found that the clustering effect of the PAM algorithm was poor, and the ability to interpret the samples was weak.

Table 6. Center point extracted by PAM algorithm

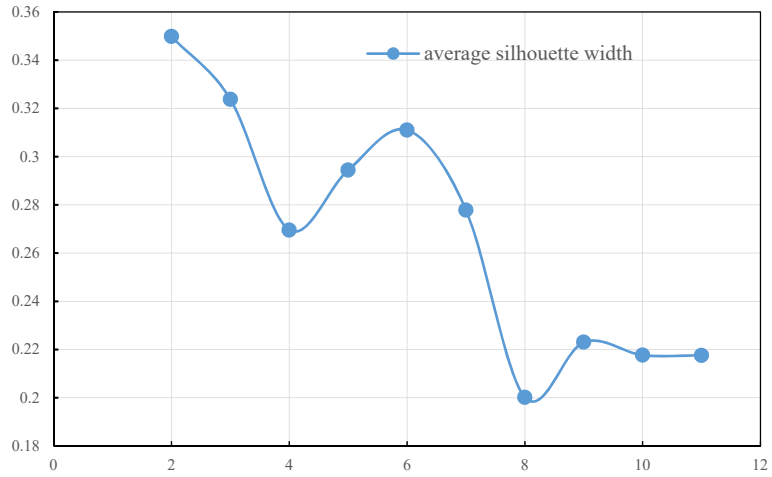
fields	cluster 1	cluster 2
Observation	497	281
Cardiovascular Disease History	0	1
Bone Metastases	0	0
Electrocardiogram Code	0	0
Stage	3	4
Survival Status	0	1
Performance Rating	0	0
Age	68	74
Weight	98	107
Systolic Blood Pressure	14	15
Diastolic Blood Pressure	8	8
Serum Hemoglobin	134	146
Size of Primary Tumor	6	18
Index of Tumor Stage and Historic Grade	9	11
Serum Prostatic Acid Phosphatase	7	84

Further, we extracted the center point of each cluster, they were samples No. 497 and No. 281 respectively. However, we still have no way of knowing the characteristics of these two clusters. Fortunately, the approximate characteristics of each cluster can be inferred by looking at the attributes of the final center sample. The field characteristics of each central sample are reported in Table 6. The first center is patient No. 497, currently in Stage 3, no history of cardiovascular disease, no bone metastases, normal electrocardiogram, normal activity, between 60-70 years old, about 80-100 pounds, the size of the primary tumor is 6cm, and finally survived. The second center is patient No. 281, currently in Stage 4, have a history of cardiovascular disease, no bone metastasis, normal activity, between 70-80 years old, about

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202105_7(5).0009

100-120 pounds, the size of the primary tumor is close to 20cm, and finally died of prostate cancer. We speculated that the classification of these two clusters is based on whether there is a history of cardiovascular disease, Stage, the size of the primary tumor, and the final survival status.

3.3.2 CLARA Clustering



Figures 6. Silhouette plot of the CLARA algorithm

The contour coefficient map drawn by the CLARA algorithm is reported in Figure 6. The maximum value of the average silhouette width is 0.35, and the corresponding number of clusters is k=2. Therefore, like the PAM algorithm, the CLARA algorithm also recommends dividing patients into 2 clusters.

The clustering renderings drawn by the CLARA algorithm is reported in Figure 7. Although it is divided into two clusters, the clustering effects of PAM algorithm and CLARA algorithm are quite different. Because the two components of the CLARA algorithm explain 65.83% of the total variance, 51.69% higher interpretation ability than the PAM algorithm. In other words, the clustering effect of CLARA algorithm is better than that of PAM algorithm. The CLARA algorithm extracts 44 samples from the Byar as the best samples, then uses the PAM algorithm on these 44 samples. Finally, the first cluster divided has 276 samples, and the second cluster has 199 samples. This is similar to the result of dividing into two clusters according to Stage, because Stage 3 has 273 samples and Stage 4 has 202 samples.

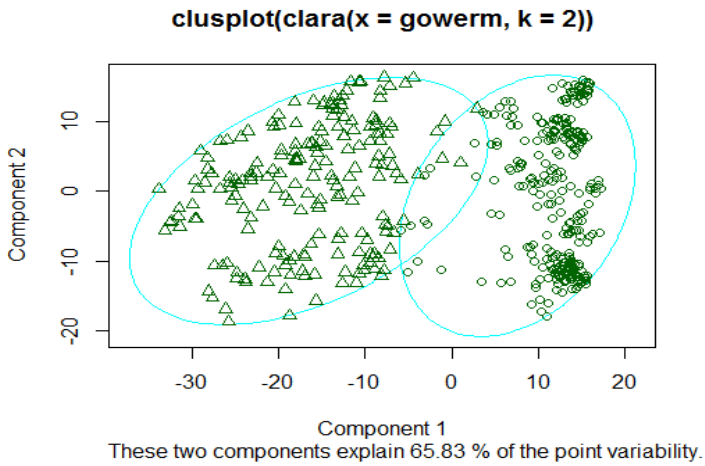


Figure 7. CLARA clustering renderings

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202105_7(5).0009

Next, cross-analyze the clustering results of the PAM algorithm and CLARA algorithm vs fields. The field we care about is "Stage". The results of the cross-contingency table are reported in Table 7. We found that the result of CLARA algorithm is based on "Stage", 98.5% of patients in the first cluster belong to Stage 3, and 97.5% of patients in the second cluster belong to Stage 4. In contrast, the results of the PAM algorithm are not so obvious. In the PAM algorithm, only 87.7% of patients in the first cluster are in Stage 3, and 90.7% of patients are in Stage 4. In other words, the result of CLARA algorithm is better than PAM algorithm.

Table 7. PAM, CLARA results and Stage

	CLARA		PAM	
cluster	3	4	3	4
1	268	8	256	36
2	5	194	17	166

3.3.3 Comparison

Extracted the samples whose PAM clustering results are different from the CLARA clustering results, and reported in Table 8. We found that out of 475 patients, there were 66 patients whose classification results were inconsistent under the two clustering algorithms, with a difference rate of 13.9%. Due to space limitations, we only listed the top 5 inconsistent samples.

The reason for this difference is that the PAM algorithm is divided incorrectly. Some patients in Stage 3 is divided into the second cluster, and other patients in Stage 4 is divided into the first cluster. This is contrary to the division of most samples. Maybe it is because these difference samples happen to be in the transition stage from Stage 3 to Stage 4, and their various index values are in a critical state, and the PAM algorithm cannot easily distinguish whether they are in Stage 3 or Stage 4.

Table 8. Samples with inconsistent results (top 5)

Observation	Stage	PAM result	CLARA result
14	3	2	1
27	4	1	2
44	4	1	2
45	4	1	2
55	3	2	1

3.4. Empirical Analysis Results

3.4.1 Results of Exploratory Analysis

(a) The age of the patients is mainly 70-80 years old, accounting for 63%. The probability of getting the disease before the age of 50 is extremely small, but after 50 years old, as the age increases, the probability of getting the disease increases. Therefore, the elderly should prevent the disease in time.

(b) The weight of patients is between 80-120 pounds, accounting for 84%. Therefore, the older and lighter ones should pay attention to prevention.

(c) 93% of patients in Stage 3 are able to move normally, 86% of patients in Stage 4 are able to move normally. It is no need to worry about the patient's restricted mobility.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202105_7(5).0009

(d) The survival rate of patients in Stage 3 is 33%, which is 10% higher than that of Stage 4. In addition, patients in Stage 3 have a 23% probability of dying from heart and vascular disease, which is 12% higher than the probability of dying from prostate cancer. However, nearly half of patients in Stage 4 die of prostate cancer. In other words, the probability of cure is small in advanced cancer.

(e) The size of the primary tumor has an obvious right-skewed distribution, about half of the patients can be found when the tumor size is 0-10cm.

3.4.2 Results of Cluster Analysis

(a) Both the PAM algorithm and the CLARA algorithm divide the samples into two clusters, but the clustering effects are obviously different. The interpretation ability of CLARA algorithm's clustering results is 51.69% higher than PAM algorithm.

(b) The clustering effect of CLARA algorithm is better than that of PAM algorithm.

(c) The CLARA algorithm is mainly divided according to the value of Stage. 98.5% of patients in the first cluster are in Stage 3, and 97.5% of patients in the second cluster are in Stage 4.

(d) There are 66 samples with inconsistent classification results between the PAM algorithm and the CLARA algorithm, with a difference rate of 13.9%. The reason for the difference is that the PAM algorithm cannot determine which cluster the patients in the transition stage belong to.

4. CONCLUSION

We mainly focus on the mixed data clustering analysis of the PAM algorithm and CLARA algorithm. First, we elaborated on the basic ideas of the two clustering algorithms, the calculation of the comprehensive distance and the algorithm process. Second, we selected the Byar data set as the sample input and conducted a preliminary exploratory data analysis. Finally, we use PAM algorithm and CLARA algorithm to cluster the samples. We found that the PAM algorithm and the CLARA algorithm have significantly different results for the same samples. The variance contribution rate of the PAM algorithm was only 14.14%, while the CLARA algorithm was 65.83%. In addition, 13.9% of samples with inconsistent results. The clustering effect of CLARA algorithm is significantly better than PAM algorithm.

REFERENCES

[1] Binhui Wang. Multivariate statistical analysis and R language modeling[M]. Jinan University Press, 2016.

[2] Park H S, Jun C H. A simple and fast algorithm for K-medoids clustering[J]. Expert Systems with Applications, 2009, 36(2):3336-3341.

[3] Dodge Y. Statistical Data Analysis Based on the L1-Norm and Related Methods[M]. North-Holland, 1987.

[4] Kaufman L, Rousseeuw P J. 3. Clustering Large Applications (Program CLARA) [M]// Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons, Inc. 2008:126-163.

[5] Ng R T, Han J. Efficient and Effective Clustering Methods for Spatial Data Mining [M]. University of British Columbia, 1994.

[6] Wei Chen, Lei Wang, Ziyun Jiang. Hybrid attribute data clustering algorithm based on K-prototypes[J]. Journal of Computer Applications, 2010, 30(8):2003-2005.