Journal of the Nigerian Association of Mathematical Physics
Volume 55 (**February** *2020 Issue*), **pp83 –90**
© J. of NAMP

COMPARISON OF TWO-STEPS AND K-MEANS CLUSTERING ALGORITHMS FOR THE CLASSIFICATION OF DISEASES RECORD DATA

Dodo B.[1], *Alhaji B. B.*[2] *and Doguwa S.I.S.*[3]

[1,3]**Department of Statistics, Ahmadu Bello University, Zaria, Nigeria**
[2]**Nigeria Defense Academy, Kaduna**

Abstract

In this paper, two clustering algorithm have been apply to large medical record dataset. Various clustering algorithm have been developed to group data into clusters in many domains especially in medical field. Some of these algorithm considered the size of the data set and worked either on continuous, categorical or mixed data type. The Two-steps and K-means are evaluated in order to identify the best fit of the data. The experimental result revealed five clusters for the K-means and six clusters for the Two-steps. The test of independence from chi-square showed that the number of cluster differ from one clustering method to another. Furthermore, K-means clustering showed best fit with RMSSTD value of 3.7847, which is lower than that of the two-steps clustering with a value of 3.9652.

Keywords**:** Two-steps cluster,K-means clustering, Root Mean Squared Standard Deviation.

Introduction

With the availability of massive data stored in many hospital (files, databases, and diverse repositories), it is crucial to develop some means for analyzing and interpreting these types of data which can help the decision-makers.

Morrison [1] said that multivariate statistical analysis is concerned with the collection of data on several dimensions of the same individual; such observations are common in the social, behavioral life and medical science. It therefore helps the researcher to summarize the data and reduce the number of variables necessary to describe it. Morison [1] further stated that, as a result of the complexity of some variables in some datasets, multivariate techniques such as cluster analysis will be used for classification.

Chaudhary and Sharma [2] stated that cluster analysis depends on, among other things, the size of the data file. Methods commonly used for small data sets are impracticable for data files with thousands of cases. They continue by saying that in presence of large data file (even 1000 are large for clustering) or mixture of continuous and categorical variables, you should use the two-step procedure.

A common application of cluster analysis particularly in medicine is to categorize patients into subgroups or diagnostic categories based upon patterns of clinical signs and symptoms that empirically goes together [3]. There are many clustering algorithms that have been developed for various fields. But most of these classical clustering algorithms for example (hierarchical and k-means) focus either on numerical data or on categorical data type [4]. They have some difficulties of being applied on a data set that contain both types of data (numerical and categorical). Also, a hierarchical clustering method is mostly used for a small size of data set while the k-means clustering approach is used for a large size data set of numerical types [5]. In spite of these weaknesses; this work proposed to investigate which of these methods: two-steps clustering analysis approach and the k-means clustering approach best fit the large data set in hand.

The k-mean cluster tries to form k reasonably homogeneous clusters based on the variables and supposed that the right number of clusters have been chosen. The aim of k-mean clustering consists of partition n observation into k clusters in such a way that each observation belongs to the cluster with closest mean. It can analyse data set that are extremely large. In k-means clustering, k is the number of selected clusters, and a case is allocated to the cluster for which its distance to the cluster mean is the minimum. So, the analysis starts with an initial set of means and grouped cases based on their square distance to the centers. Then the cluster mean is calculated using the cases that are assigned to the cluster. Next, all cases are reclassified based on the different set of means. This stage is repeated up to the stage where clusters means are nearly the same among consecutive steps. At the end, clusters means are calculated again and cases are allocated to their permanent clusters.

The two-step clustering developed by [6] is used to cluster data into different clusters and allocate classes based on variables. The two-step clustering is a combination of the two classical methods (Hierarchical clustering and k-mean clustering). The Two-Step cluster technique is considered as accessible cluster analysis algorithm that is designed to handle very large datasets. It is suitable to handle both regular, categorical variables and attributes [7]. It is achieved in two steps in which the first step consists of pre-cluster the cases or

Corresponding Author**:** Dodo B., Email: dboubakar6@gmail.com, Tel: +2347037627212

records into several small sub-clusters and the second step consists of pulling together the sub-clusters that are the output of pre-cluster step into the desired number of clusters. It can also naturally select the number of groups. This clustering technique is very effective in classification of huge data sets and it has the ability to form clusters by using categorical and continuous variables and it is provided with natural selection of number of clusters.

In practice, a cluster analysis is the end product of a series of analytical decisions. The analytical decision made at each point in the series can significantly affect subsequent decisions, as well as the overall results of a cluster analysis [3]. The analytical decision usually involve choice about what object to cluster, what unit of measurement to use for the variables, what proximity measure to use as an index of similarity or dissimilarity among the object, what type of clustering algorithm to use, and what criteria to use for determining the number and quality of clusters in the data.

Ravinder and Sharma [8] used two-step clustering approach using back propagation for tuberculosis data. The study showed two-steps clustering is efficient clustering algorithm, which it describe as best in clustering and classification. The authors compared their model with existing models and showed that proposed model performed better in terms of accuracy.

Vijayarani and Sudha [9] predicted five types of diseases using hemogram blood samples. The diseases are Leukemia, Inflammatory disease, bacterial or viral infection, HIV infection and Pernicious anaemia. In this research, K-means clustering algorithm, fuzzy c-means clustering and the weight-based k means clustering algorithms were compared by using the performance factors namely time, clustering accuracy and error rate. The proposed weight-based K-means clustering algorithm has performed well when compared to other algorithms. Rujasiri and Chomtee [10] compared the effectiveness of five clustering techniques with multivariate data. The techniques were hierarchical clustering method; K-means clustering algorithm; Kohonen's Self Organizing Maps method (SOM); K-medoids methods with Dynamic Time Warping distance measure (DTM). Root mean square standard deviation (RMSSTD) and RS results were used to evaluate these five techniques. At the end of their research, they observed that for both real and simulated datasets provided the same result with the k-means clustering method having the lowest RMSSTD and the SOM method having highest r^2 (RS) in the simulation studies. Hence both k-means and SOM were considered to be the most suitable techniques for cluster analysis.

Source of data

The data for this research was collected from patient medical diagnosis record book of the CSI's of Maboyé amaré, Gueban zogui, Madaou, Illéla, keita. The data set contain the following variables age, the gender of the patient, the provenance of the patient, the diagnosis of the patient and finally the date and years of visiting the CSI. The research study covers a period of four years, going from 2014 to 2017 and a total of 28,717 patients were considered. Among which we have 11,656 men and 17,061 females.

The Study Area

The region of Tahoua is one of the eight (08) regions that make up Niger Republic. The population of the region as at 1st January 2015 was 3,598,280 (INS). The region of Tahoua has a total area of about 113,371 km^2 (8.95% of national territory). The region comprises 12 (départements) local governments that are Abalak, Bagaroua, Birni N'Konni, Bouza, Illela, Keita, Madaoua, Malbaza, Tahoua, Tassara, Tchintabaraden and Tillia. Tahoua is limited in the North by the region of Agadez, in the South by the Federal republic of Nigeria, in the East by the region of Maradi and in the West by the regions of Dosso, Tillabéry and the republic of Mali.The geographic coordinates at the zero point of the region, situate it between parallels 13°42'and 18°30' North latitude and meridians 3°53'and 6°42' East longitude.

Methods of partition

In the common usage, clustering is employed to partition a set of objects or individuals into classes such that any of this objects or individuals belongs to one and only one class (and there will not be an empty classes). So, partition is defined as such system of classes. Mathematically it can be expressed as:

Let I be a set of elements. We said that $Q(I)$ is a partition of I when there exists a set $(q_1, q_2, \ldots, q_k)$ a sub set of I non-empty such that:

$$\begin{cases} \bigcup_{i=1}^{k} q_i = I \qquad \text{with } q_i \neq \varnothing \\ q_i \bigcap q_j = \varnothing \qquad for\ i \neq j \end{cases}$$

Finding the best partition of a set I through a given criteria consist of examining all possible partitions of the set I . It is a difficult combinatory problem for some sets of big dimension.The possible number of partitioning I into 2 clusters is given by:

$$P_{n,2} = 2^{n-1} - 1$$

Standardize the data matrix

In many situations, the variables have different units that can arbitrary affect the similarities between the observations. By standardizing the variables and changing them into dimensionless units will first of all, remove the arbitrary affect and secondly makes variables contributions equal in the similarities among observations.

To standardize the dataset, we first choose an equation called standardizing function applied to the dataset. The dataset has n variables that have subscripts $i = 1, 2, ..., n$ and t observations subscripted $j = 1, 2, ..., t$; in the data set the value of data for any i^{th} variable and j^{th} observation is noted by X_{ij} and the corresponding standardized data matrix value is given as Z_{ij} .

The standardizing function is denoted by:

$$Z_{ij}=\frac{X_{ij}-\bar{X}_i}{S_i} \tag{3.1}$$

where $\bar{X}_i=\frac{\sum_{j=1}^{t}X_{ij}}{t}$ and $S_i=\left[\frac{\sum_{j=1}^{t}\left(X_{ij}-\bar{X}_i\right)^2}{t-1}\right]^{\frac{1}{2}}$

Measure of Similarity or Dissimilarity

There are different measures of similarity that are frequently used for continuous variables and qualitative variables in cluster analysis. For high dimensional data the popular measure is the Euclidean distance and the log-likelihood distance.

Minkowski Metric

This is the p root of the sum of the differences between the measurements for each variable.

$$d\left(i_r,i_k\right)=\left[\sum_{i=1}^{t}\left(X_{rj}-X_{kj}\right)^t\right]^{\frac{1}{t}} \tag{3.2}$$

Where

t is the dimension of the data.

When $t=2$ we have the Euclidean

When $t=1$ we have the Manhattan distance

Euclidean Distance

It is the most commonly used. It is expressed by the following equation:

$$d\left(i_r,i_k\right)=\left[\sum_{j=1}^{t}\left(X_{rj}-X_{kj}\right)^2\right]^{\frac{1}{2}} \tag{3.3a}$$

"Distance**"** between element i_r and element i_k in a set I of n dimension. But with standardization we obtained:

$$d\left(i_r,i_k\right)=\left[\sum_{j=1}^{t}\left(z_{rj}-z_{kj}\right)^2\right]^{\frac{1}{2}} \tag{3.3b}$$

"Standardized distance**"** between element i_r and element i_k in a set I of n dimension. Variable by variable, the difference and the squares of these differences are taken. After the squaring none are negative, then sum the squared difference and take the square root of their sum. It can be used only if all variables are continuous [5].

Log-likelihood distance

The Log-likelihood distance can be used for both continuous and categorical variables. The distance between two clusters is related with the decrease of the natural logarithm of likelihood function, since they are grouped in one cluster to compute the likelihood distance it is assumed that the continuous variables have normal distributions and the categorical variables have multinomial distributions, and also the variables are independent of each other [5].The distance between cluster i and j is defined as:

$$d\left(i,j\right)=X_i+X_j-X_{<i,j>} \tag{3.4}$$

Where $X_s=-N_s\left\{\sum_{k=1}^{q}\frac{1}{2}\log\left(\hat{\delta}_k^2+\hat{\delta}_{sk}^2\right)+\sum_{k=1}^{p}\hat{E}_{sk}\right\}$; $q=k^A$; $p=k^{\mathbf{B}}$

And $\hat{E}_{sk}=-\sum_{i=1}^{L_k}\frac{N_{skl}}{N_s}\log\frac{N_{skl}}{N_s}$

With notations:

$d\left(i,j\right)$ is the distance between cluster i and j ; $<i,j>$ index that represents the cluster obtained by combining clusters i and j ; k^A is the total number of continuous variables; $k^{\mathbf{B}}$ total number of categorical variable; L_k is the number of categories for the $k-th$ categorical variable; N_s is the total number of data records in cluster s ; N_{skl} is the number of records in cluster s whose categorical variables k takes l category; N_{kl} is the number of records in categorical variable that take the l category; $\hat{\delta}_k^2$ the estimated variance (dispersion) of the continuous variable k for the entire data set; $\hat{\delta}_{sk}^2$ the estimated variance of the continuous variable k in cluster j .

k-Means Clustering Algorithm

The algorithm has the following steps:

Step 1:

i. Randomly chose k centers, these centers can be either any points in the variables spaces or elements of I . these k centers are temporary and let them be: $C_1^{(0)}, C_2^{(0)}, C_3^{(0)}, ..., C_k^{(0)}$

ii. Affect every individual of I to a class and only one class which center is one of the k centers earlier established through the following rule:

$$\begin{cases} i \text{ belongs to the class named } P_l^{(0)} \\ \text{of center } C_l^{(0)} \end{cases} \Rightarrow \begin{cases} \left\| i - C_l^{(0)} \right\| \text{is minimum to go through} \\ \text{all the centers } C_1^{(0)}, C_2^{(0)}, C_3^{(0)}, ..., C_k^{(0)} \end{cases}$$

At the end of the step 1, there will be k classes named $P_1^{(0)}, P_2^{(0)}, P_3^{(0)}, ..., P_k^{(0)}$ with centers respectively $C_1^{(0)}, C_2^{(0)}, C_3^{(0)}, ..., C_k^{(0)}$ and every element i of I belongs to one of the classes $P_l^{(0)}$.

Step 2:

i. Determine k new centers in the following way: computes the centroid of the classes $P_1^{(0)}, P_2^{(0)}, P_3^{(0)}, ..., P_k^{(0)}$ and note them $C_1^{(1)}, C_2^{(1)}, C_3^{(1)}, ..., C_k^{(1)}$ the new centers.

ii. Affect every individual i of I to a class and only one class whose center is one of the k centers $C_1^{(1)}, C_2^{(1)}, C_3^{(1)}, ..., C_k^{(1)}$ through the following rule:

$$\begin{cases} i \text{ belongs to the class named } P_l^{(1)} \\ \text{of center } C_l^{(1)} \end{cases} \Rightarrow \begin{cases} \left\| i - C_l^{(1)} \right\| \text{is minimum to go through} \\ \text{all the centers } C_1^{(1)}, C_2^{(1)}, C_3^{(1)}, ..., C_k^{(1)} \end{cases}$$

At the end of this step, a new system of k classes were generated and noted by $P_1^{(1)}, P_2^{(1)}, P_3^{(1)}, ..., P_k^{(1)}$ with respective centers $C_1^{(1)}, C_2^{(1)}, C_3^{(1)}, ..., C_k^{(1)}$ and every element i of I belongs to one of the classes $P_l^{(1)}$.

Step h:

i. The new K centers are noted; $C_1^{(h)}, C_2^{(h)}, C_3^{(h)}, ..., C_k^{(h)}$ and determined from the classes obtained at the step $(h-1)$: $P_1^{(h-1)}, P_2^{(h-1)}, P_3^{(h-1)}, ..., P_k^{(h-1)}$ by computing the centroid of the classes.

ii. Every individual i of I is affected again to a class and only one whose center is one of the k centers $C_1^{(h)}, C_2^{(h)}, \dots, C_i^{(h)}, \dots, C_k^{(h)}$ through the following rule:

$$\begin{cases} i \text{ belongs to the class named } P_l^{(h)} \\ \text{of center } C_l^{(h)} \end{cases} \Rightarrow \begin{cases} \left\| i - C_l^{(h)} \right\| \text{is minimum to go through} \\ \text{all the centers } C_1^{(h)}, C_2^{(h)}, C_3^{(h)}, ..., C_k^{(h)} \end{cases}$$

At the step h, determined the partition $P_1^{(h)}, P_2^{(h)}, P_3^{(h)}, ..., P_k^{(h)}$ into k classes whose centers are $C_1^{(h)}, C_2^{(h)}, C_3^{(h)}, ..., C_k^{(h)}$.

Repeat these steps until none of the cluster assignment change. That means that the clusters are stable (convergent).

Two-step clustering algorithm

The algorithm of two-step clustering method has some characteristics that differentiate it from other clustering methods.

i. It has the capacity of creating cluster based on continuous and categorical variables.

ii. It selects the number of clusters automatically.

iii. It has the capacity of analyzing efficiently large dataset.

The algorithm can be summarized to two steps:

Step 1:

The algorithm starts with the building of Cluster Features Tree. The tree starts by placing the first case or object at the root of the tree in a leaf node that contains information of the variable concerning that case. Each case is one after the other added to an existing node or form a new node, based upon its similarity to the node that exist and use distance measure criterion for similarity. A node that have multiple cases contains a summary of variables information concerning those cases. Then, the Cluster features tree provide a summary of the dataset.

Step 2:

The leaf node of the Cluster feature tree is then merged using agglomerative clustering algorithm. A range of solution is provided by the agglomerative clustering methods. To determine the best number of clusters, the Bayesian information criterion (BIC) or the Akaike information criterion (AIC) can be used [6].

For each number of clusters, the AIC, AIC change, Ratio of AIC changes and Ratio of distance measure are computed. A good solution for the number of cluster is obtained where a large Ratio of Distance Measures is obtained SPSS 2001; [6]; or computing the ratio of the two largest values of the ratio of distances measures. Let these values be k_1 and k_2. When k_1/k_2 is greater than 1.15 then our number of clusters is $k = k_1$ if not k will be the maximum of $k_1 \& k_2$ [11].

Test of independence

Here we want to test whether number of clusters obtained from the two methods are independent. Using test for independence.

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}} \qquad (3.8)$$

Which is distributed approximately as chi-square statistics with $(r-1)(c-1)$ degrees of freedom. O_{ij} is the observed frequencies and E_{ij} is the expected frequencies.

The hypothesis are:

H_0 : The number of clusters obtained from the two methods of clustering do not differ

H_a: The number of clusters obtained from the two methods of clustering differ.

The test statistics reject H_0 when $\chi^2_{cal} > \chi^2_{(1-\alpha)(I-1)(J-1)}$ at $\alpha\%$ level of significance otherwise reject**.**

Criteria for evaluation RMSSTD

The root mean square standard deviation is an evaluation method used to measure the quality of the clustering algorithm. The lower its value the better the separation of cluster.

$$RMSSTD = \sqrt{\frac{\sum_{j=1...p}^{i=1...k}\sum_{a=1}^{n_{ij}}\left(x_a-\bar{x}_{ij}\right)^2}{\sum_{j=1...p}^{i=1...k}\left(n_{ij}-1\right)}} \qquad (3.10)$$

Where $k$the number of clusters is, p is the number of independent variables in the dataset, $\bar{x}_{ij}$ is the mean of the data in variable j and cluster$i$, and n_{ij} is the number of data which are in variable p and cluster$k$.

Table 1: K-Mean Clustering: Iteration History[a]

	Change in Cluster Centers				
Iteration	1	2	3	4	5
1	2.949	2.977	3.095	2.646	2.837
2	.310	.165	.050	1.106	1.322
3	.225	.138	.055	.937	.635
4	.179	.082	.180	.858	.368
5	.091	.185	.219	1.101	.211
6	.041	.095	.236	.765	.110
7	.029	.293	.568	.435	.040
8	.150	.157	.550	.476	.057
9	.179	.108	.152	.314	.067
10	.086	.139	.015	.180	.025
11	.236	.177	.009	.361	.028
12	.115	.061	.012	.177	.043
13	.030	.016	.008	.039	.020
14	.015	.008	.002	.009	.008
15	.005	.003	.002	.002	.004
16	.001	.001	.002	.001	.003
17	.000	.000	.001	.000	.001
18	.000	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is 0.000. The current iteration is 18. The minimum distance between initial centers is 6.416.

Table 1 shows the progress of the clustering process starting from the first iteration to the twelve iterations. We observed that the cluster centers shift quite a lot up to the 13th iteration. From the 14th to the 18th iteration the cluster centers converge and the algorithm stops due to convergence at the 18th iteration.

Table 2: K-Mean Clustering: Distances between Final Cluster Centers

Cluster	1	2	3	4	5
1	-	2.257	2.898	1.798	3.148
2	2.257	-	2.670	2.104	3.185
3	2.898	2.670	-	2.693	3.562
4	1.798	2.104	2.693	-	3.123
5	3.148	3.185	3.562	3.123	-

The table 2 shows the Euclidean distances between the final clusters centers in which, large number of distances between clusters correspond to greater dissimilarities. Cluster 5 is different from clusters 1, 2, 3 and 4; but cluster 1 is almost the same as cluster 4. Cluster 2 is approximately similar to clusters 3 and 4.

Table 3:the distribution of the observations among the clusters

Cluster	Number of cases
1	5342.000
2	9808.000
3	4634.000
4	6471.000
5	2462.000
Valid	28717.000

Table 3 above shows the distribution of the observations among the clusters. Cluster 2 having the highest number of observations and cluster 5 the smallest number of observations. The clusters 1, 3 and 4 having almost the same number of observations.

Table 4:Two-step auto clustering

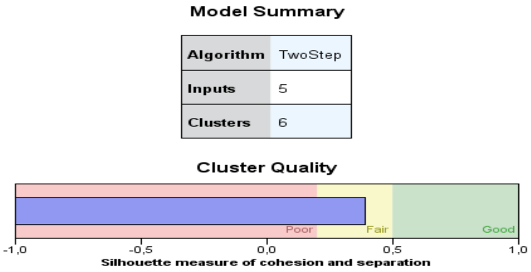
Number of Clusters	Akaike's Information Criterion (AIC)	AIC Change	Ratio of AIC Changes	Ratio of Distance Measures
1	99543.038			
2	84079.761	-15463.277	1.000	1.341
3	72556.559	-11523.202	.745	1.171
4	62721.285	-9835.274	.636	1.361
5	55499.463	-7221.822	.467	1.316
6	50015.743	-5483.720	.355	1.833
7	47032.685	-2983.057	.193	1.183
8	44513.647	-2519.038	.163	1.064
9	42147.481	-2366.166	.153	1.089
10	39975.821	-2171.660	.140	1.070
11	37946.957	-2028.864	.131	1.179
12	36229.613	-1717.344	.111	1.048
13	34592.126	-1637.486	.106	1.317
14	33353.348	-1238.779	.080	1.043
15	32166.130	-1187.218	.077	1.096

The Table 4 gives the process of selecting the optimum number of clusters through the two-steps auto clustering table. For each number of clusters, the AIC, AIC change, ratio of AIC changes and Ratio of distance measure are computed. A good solution for the number of cluster is obtained where a large ratio of distance measures is obtained (SPSS 2001; [6]); or computing the ratio of the two largest values of the ratio of distances measures. Let these values be k_1 and k_2. When k_1/k_2 is greater than 1.15 then our number of clusters is $k = k_1$ if not k will be the maximum of k_1 & k_2 [11]. So, in our case we have $k_1 = 1.833$ and $k_2 = 1.361$; the ratio of $k_1/k_2 = 1.3468$ which is greater than 1.15, so our optimum number of clusters will be 6 for the two-steps clustering.

Table 5: Two-steps cluster distribution

Cluster	N	% of Total
1	7779	27.1%
2	3794	13.2%
3	2089	7.3%
4	4750	16.5%
5	2705	9.4%
6	7600	26.5%
Total	28717	100.0%

The Two-steps cluster distribution table above shows the frequency of each cluster. Out of the 28717 total observations, 7779 were assigned to the first cluster, 3794 to the second, 2089 to the third, 4750 to the fourth, 2705 to the fifth and 7600 to the last.



Model Summary

Algorithm	TwoStep
Inputs	5
Clusters	6

Figure 1: The two-steps output model viewer object

The model viewer output indicates the model summary and the quality of the cluster. The table of the model summary shows that six clusters were found based on the five variables and the silhouette measure of cohesion and separation chart (cluster quality) indicates that the overall model quality is Fair.

The Test of Independence

The chi-square and the likelihood ratio test statistics are used to test whether the number of clusters obtained from the two clustering methods differ or not. The results of the two analyses are shown below.

Table 6: Chi-Square Tests of Independence

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	73730.781[a]	20	.000
Likelihood Ratio	59560.993	20	.000
Linear-by-Linear Association	5669.317	1	.000
N of Valid Cases	28717		

Result from Table 6 indicates that the null hypothesis is rejected at $\alpha = 0.05$ level since the p values are less than 0.05; meaning that the number of clusters obtained from the two clustering methods differ.

The Root Mean Square Standard Deviation (RMSSTD)

After running the two algorithm (the k-means and two-steps clustering) the Root Mean Square Standard Deviation [12] is used to evaluate the experimental results. This statistic shows the Root Mean Square Standard Deviation for individual cluster. It is the grouped standard deviation of the entire variables that create the cluster. Its value should be the smallest as possible to conclude that the clusters formed are homogeneous and that is because the main objectives of any clustering methods is to get homogeneous groups. The values of the RMSSTD obtained are computed using the formula (3.10).

The results of RMSSTD obtained from the two methods, shows that the RMSSTD of the two-steps clustering methods is slightly greater than that of K-mean clustering methods. Based on that, we came to a conclusion that the K-mean clustering methods is the one that best fit the dataset.

Conclusion

In this paper, an attempt to study two different clustering algorithms namely K-means and Two-steps to group a large dataset obtained from medical records which comprises of five variables and 28717 observations. After applying the two methods of clustering a one-way ANOVA was applied to see the variable that contribute much in the separation of the clusters and a test of independence (chi-square) is performed to see whether the variables with respect to the number of clusters obtained from each method differ. Also, the Root Mean Square Standard deviation (RMSSTD) is used to evaluate the fitted model. All the analysis was conducted using SPSS version 21.

The result of the K-means clustering algorithm ended with five clusters. A maximum number of twenty (20) iteration is considered and the initial cluster centers show that the variable values is well spaced among the clusters for each of the five variables. The final cluster centers computed as the mean for each variable within each final cluster reflects the characteristics of the case (observation) for each cluster. The distribution of the observations among the clusters shows that cluster 2 has the highest number of observations (9808) and cluster 5 the smallest number of observations (2462). The clusters 1, 3 and 4 have respectively (5342; 4634 and 6471).

The output of the Two-steps gives six clusters. The model summary shows that six clusters were found based on the five variables and the silhouette measure of cohesion and separation chart (cluster quality) indicates that the overall model quality is fair.The predictor importance diagram shows each clustering variables relative importance. We observed that all the five variables are important predictors.

The chi-square test and the likelihood ratio test statistics was used to test whether the number of clusters obtained from the two clustering methods differ. The result shows that H_0 is rejected and conclude that the number of clusters differ from one method to another.

The RMSSTD was used to evaluate the quality of the methods. The analysis shows that the k-mean method has a RMSSTD value of 3.7847 and the two-steps a value of 3.9652. From these values we observed that both methods fit our dataset, but the best fit is the k-means clustering methods since it has the lowest value of RMSSTD.

The cluster 1 has 5342 observations among which we have 5162 females and 180 males. The distribution of the diseases shows that the cluster 1 is dominated mainly by infection buccale, infection dermatologique, infection respiratoire and paludisme as shows.

The second cluster is formed of 9808 observation contain the following diseases Infection buccale, infection dermatologique, infection respiratoire, paludisme, galireuse, infection urinaire and anemia. The third cluster contain 4634 observation and is dominated mainly by the following diseases infection, cephale, kyste, Conjonctivite, traumatisme and Octure. The fourth cluster is formed of 6471 patients among which most of them suffers from infection buccale, infection dermatologique, Paludisme, infection respiratoire infection urinaire and Glaireuse. The fifth cluster has 2462 patients that mostly suffers from infection buccale, paludisme, infection dermatologique, Glaaireuse and infection respiratoire.

The provenance that need more attentions are Koweit, M/A, Madaoua, Malala, S/G, Wadata and Bilbis in cluster I. In cluster II we Bilbis, G/Z, Illela, Madaoua, M/A, Koweit, Malala, S/G and Wadata. We have also in cluster 3, Bilbis, G/Z, Illela, Koweit, M/A, Madaoua, Malala, S/G and Wadata.

We observed that Bilbis, G/Z, Illela, Koweit, M/A, Madaoua and Malala need more attention cluster 4. Lastly, in cluster 5 we have Bilbis, Illela, Koweit, M/A, Madaoua, S/G and Wadata as observed.

References

[1] Morrison, D. F. (1990): “Multivariate statistical method.” (MC Craw Hill).

[2] Chaudhary, K.,& Sharma, A. (2014). Implementation of two steps clustering using telcommunication system. *International Journal of IT and Knowledge Management*, 7(2), 42-48.

[3] Everitt, B. S., Landau, S.,& Leese, M. (2001). *Cluster analysis* (4th ed.). NY: Oxford University Press.

[4] Shih, M., Jheng, W. J., Lai, F. L., (2010). A Two-Step Method for Clustering Mixed Categorical and Numeric Data. *Tamkang Journal of Science and Engineering*. 13(1). 11-19.

[5] Schiopu, D. (2010). Applying Two-Step Cluster analysis for identifying Bank Customers’ Profile, *BULETINUL*, 62(3), 66-75.

[6] Chiu, T., Fang, D., Chen, J., and Wang, Y. (2001): A robust and scalable clustering algorithm for mixed type attributes in large database environment. *Proc. Int. Conf. On Knowledge Discovery and Data Mining*, San Fransico ACM, 263 - 268.

[7] Richa S., Shailendra, N. and Sujata, K. (2016). Medical Data Mining Using Different Classification and Clustering Techniques: A Critical Survey. *IEEE*, 978-1-5090-0210.

[8] Ravinder, K. & Sharma, A. (2017). “Two Step Clustering Appoach using Back Propagation for Tuberculosis Data” *International Journal o Engineering Applied Sciences and Technology* 2(3). 73-78.

[9] Vijayarani, S. & Sudha, S., (2015). “An efficient clustering algorithm for predicting diseases from hemogram blood test samples” *Indian Journal of Science and Technology*, 8(17), 1. doi: 10.17485/ijst/ 2015/v8i17/52123.

[10] Rujasiri, P. & Chomtee, B. (2009) Comparison of Clustering Techniques for cluster Analysis, Kasetsart J. (Nat. Sci.). 43:378 – 388.

[11] Rakotomalala, R. (2016). Tanagra Data mining.

[12] Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2002). Cluster validity methods: Part 1. *SIGMOD Rec*. 31(3), 40-45.