Evaluate the performance of the K-means method when increasing the sample Size

(Applied to a Dataset package for diabetics)

Omer A.Khalid*, Mubarak H. Elhafian*, **, Afra H. Abdulatif*

* Department of Statistics, Sudan University of Science and Technology,
Khartoum, Sudan omer5bob@hotmail.com,hafian10@yahoo.com
Afra_hashim@sustech.edu

** Department of Mathematics,
Faculty of Science and Arts, King Abdul-Aziz University,
Jeddah, Kingdom of Saudi Arabiamhelhafian@kau.edu.sa

DOI: 10.29322/IJSRP.8.9.2018.p8134
http://dx.doi.org/10.29322/IJSRP.8.9.2018.p8134

ABSTRACT— *In this paper, weexamined K-means clustering algorithm as one of the methods of data clustering and analysis process. The proposed study investigated the k-means algorithm with different size of samples. This study was tested the diabetes UCI standard dataset as case study. We start our experiments with 1015(Random number) sample size, and then we increase the sample size for each run by 1015 units. The elapsed time for each tested was calculated and reported for investigation process. Our investigation procedure was conducted using the R software. The One-Sample Kolmogorov-Smirnov Test used for examined the normality of the sample size and clustering processing time. The test proofed that the samples that we used is not distributed normally. In order to test the effect of the increase in sample size, we divided the samples into two groups: small samples (1015-15225) and large samples (17255-30450). Using a Mann-Whitney U test to test for a difference between the two groups, the median was found to be equal in the two groups. The correlation coefficient between time and sample size was also calculated and found to be weak. Our investigation shows that the efficiency of the k-means cannot influences with the sample size.*

Keywords— K-means; cluster analysis; Big Data; sample size, elapsed time

1. INTRODUCTION

Data represent the important aspect of any applied study.In our area there are a number of things such as Internet networks and other sources of information that contain data of various kinds images, videos, texts, numbers and other types of data .Diffrent files that contain these data and different programs deals with them. The term large data (Big Data) appeared with the development of means of communication and information networks. The main problems of large data are data storage ,data retrieval and statistical analysis. In our study, we will address the efficiency of statistical programs used in ordinary computers in data analysis when increasing sample size , this study we will be interested in the time it takes to implement the algorithms of cluster analysis. Cluster analysis is considered an important statistical analysis and has many uses in different fields. Many studies have been concerned with cluster analysis and the different ways of doing it. Most of the studies dealt with the use of algorithms, the different methods of cluster analysis, development of new algorithms and comparison between methods [1] [2] [3].A few studies concerned the time and the improvement of the time of completion according to different method [4][5].

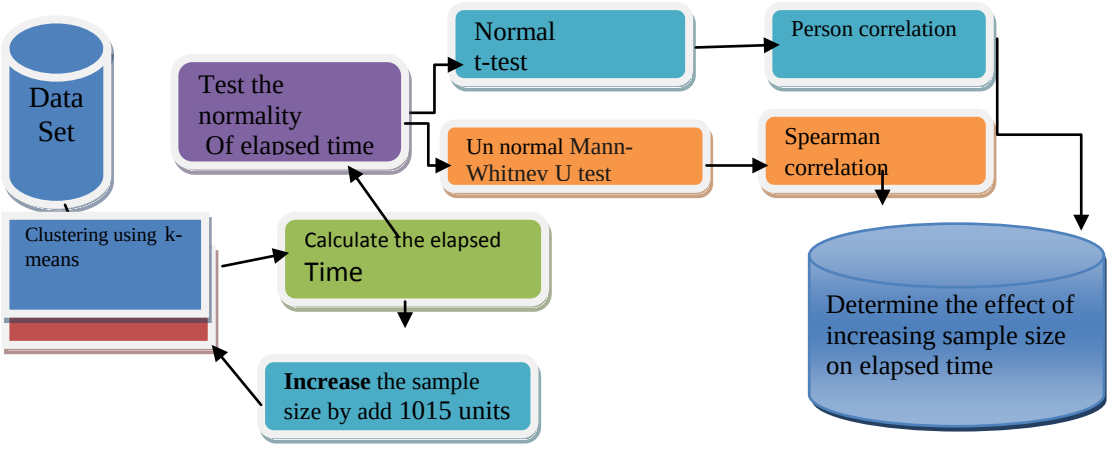
2.Basic Idea of Cluster analysis: "Cluster analysis" is the generic name for a wide variety of procedures that can be used to make a classification. These procedures empirically form "clusters" or groups of highly similar structure. More specifically, a clustering method is a multivariate statistical way that starts with a data set containing information about a sample of entities and attempts to reorganize these entities into relatively homogeneous groups [6][7]. There are many methods used in cluster analysis and the most important of these methods are the method of the K-means which is the most commonly used unsupervised machine learning algorithm for partitioning a given data set into a set of different grouped (k) groups, where k represents the number of clusters pre-specified by the researcher. It classifies units in

multiple clusters, such that data units within the same cluster are homogeneous. When we used the k-means clustering method, each cluster is represented by its center which corresponds to the mean of points assigned to the cluster [8][9].
When constructing partitions with a fixed number k of clusters, it is often assumed that there exists a function which measures the quality of different clustering of the same data set [10] [11]. For further information on clustering and clustering algorithms, see [12] [13] [14].
There are several k-means algorithms available. The standard algorithm is the Hartigan-Wong algorithm (1979), which defines the total within-cluster variation as the sum of squared distances Euclidean distances between items and the corresponding centroid [8] :
This method consists of its three simplest steps:
1. Divide values into primary groups.
2. Place each value in the group with the nearest Arithmetic Mean (usually using the Euclidean distance to calculate the distance either by actual observations or standard observations). We recalculate the arithmetic mean of the group to which the new item was added and the group from which the individual was lost.
3. Repeat the second step until the values distribution process stops. Instead of starting with the first step by dividing all the values into K from the averages of the initial groups, we can begin by specifying K (base points) and then implementing the second step.
The final distribution of data values depends, to a certain extent, on the groups on the first partition used or on the first choice of base points. Usually one does not know the number of clusters present in a data set. Because most partitioning methods provide a fixed number k of clusters, one must apply them for several values of k in order to find the most meaningful clustering [15].
In this study we will take care of the time of clustering .The following algorithm were used to calculate the time "Elapsed Time" variable as the total time from the beginning of the command until the end of it, to obtain the results we use the function proc.time ( )where this function stores the timing in a variable immediately before the beginning of the implementation and then calculate the time after implementation and subtract the initial value of it to obtain the time taken where the output appears in the form of three values .As described in the following steps[10]

Start.point <- proc.time()
"Type your code here"
proc.time () – Start.point
Start.point ≡ "the time on Start".

Proc.time () ≡ "function retrieve the time from Windows".
Note: The three steps must be shaded at execution until the R program performs the steps sequentially [10][16].

3. **Formulation of the Problem and proposed method of solution.**
Previous studies have been concerned with cluster analysis and the algorithms used in the different methods. There are few studies that have studied this analysis in other sides. This study deals with the time taken to make the clustering for data set in case of increasing the size of the sample. The important question is whether the size of the sample affects the time taken to clustering data using the k-means method? Does analyzing big data in this cluster analysis by k-means method require a special type of computer or not?



The diagram above illustrates the steps that will be taken to study and analyze the study problem

4. Numerical results. To study our problem, we took data from the site of the center of learning machine and intelligent systems University of California .Data were collected during the period1999-2008 from 130 hospitals in the United

States. These data were prepared for the purpose of diabetes study. The total number of patients whose data are available is (101766) patients. The data included variables such as strain, age, sex, duration of stay in hospital, number of laboratory tests, Number of procedures (non - laboratory) performed by the patient during his stay in the hospital, Number of medical drugs, number of outpatient clinics, emergency visits per year before entering the hospital, Number of visits to the hospital [17].

Table (1):The variables

variable	Variable label	Percentage of missing values%
Strain	Strain patient	2
Sex	Type of patient as well as definition of non-identification type as an unknown value	0
Age	The age of the patient is divided into 10 categories of 10 years in each category from 0 to 100	0
Duration	The period of the patient's stay until he is out in the range between (1-14) days	0
Laboratory procedures	Number of laboratory tests performed for the patient	0
procedures	Number of procedures (non - laboratory) performed by the patient during his stay in the hospital	0
pharmaceutical	Number of medicines taken	0
Outpatient clinics	Number of visits to outpatient clinics within a year	0
Emergency visits	The number of emergency visits carried out by the patient during the year	0
Internal visits	Number of visits to the hospital	0

Table (4) illustrates the variables introduced in the research and their definition.

-To analyze the data the R Program (R-Programming) used to measure the time taken to cluster data set. In the beginning, a random sample of 1015 patients was selected and a cluster analysis was performed using the k- mean method. Through the algorithm proc.time() in the program R the time of implementation of this process was counted. Then we add the same number each time and repeat the previous steps each time. This process was carried out 30 times and the data obtained in the table (2).

We define the following variables

User: User times of the current R process

* System: System times of the current R process

* Elapsed: Time since the process was started

Table (5) shows the measurement of the time taken by the computer for the steps of the k-means method. The time was obtained in seconds with the sample size being increase by 1015 units each time

Table (2):The time of the K-means method in seconds with sample size

no	Sample size	K-means time				no	Sample size	K-means time		
		elapsed	system	user				elapsed	system	user
1	1015	0.05	0	0		16	16240	0.85	0.01	0.14
2	2030	0.01	0	0.02		17	17255	0.64	0	0.06
3	3045	0.03	0.01	0.01		18	18270	0.3	0	0.03
4	4060	0.02	0	0		19	19285	2.24	0.02	0.08

.

5	5075	0.01	0	0.01		20	20300	0.73	0	0.13
6	6090	0.01	0	0.01		21	21315	0.5	0.01	0.07
7	7105	0.02	0	0.02		22	22330	0.82	0.02	0.13
8	8120	0.02	0	0.02		23	23345	0.38	0	0.04
9	9135	0.05	0	0.01		24	24360	0.28	0.03	0.06
10	10150	0.34	0	0.05		25	25375	0.35	0	0.06
11	11165	1.09	0.01	0.06		26	26390	0.98	0.2	0.28
12	12180	1.29	0	0.03		27	27405	0.31	0	0.1
13	13195	0.64	0	0.06		28	28420	0.23	0.01	0.09
14	14210	0.89	0	0.06		29	29435	0.22	0.02	0.1
15	15225	5.03	0.02	0.06		30	30450	0.85	0.01	0.14

-To determine which test we were used we must test the normality of the data values .we use One-Sample Kolmogorov-Smirnov Test

Graph no (1):Goodness of fit for normal distribution

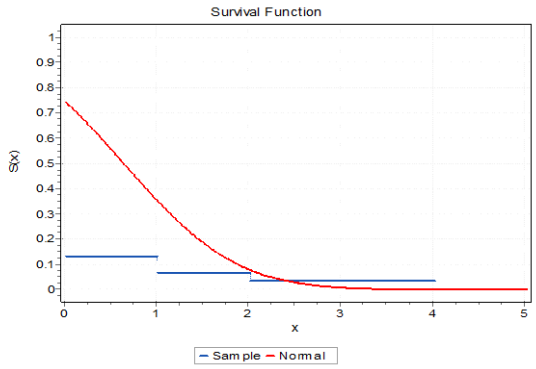


Table (3):One-Sample Kolmogorov-Smirnov Test

Hypothesis	Test value	P-value	Decision
Data are not distributed normally	0.25766	0.03035	Accept the hypothesis

Table (6) shows the Normality Test. the Sig for One-Sample Kolmogorov-Smirnov Test is equal to (0.03035) is less than 0.05 that lead us to accept the alternative hypothesis (the data are not distributed normally).

-To illustrate the relationship between the sample size and the elapsed time we compute the spearman rank correlation coefficient

Graph no (2)
Relationship between elapsed time and sample size

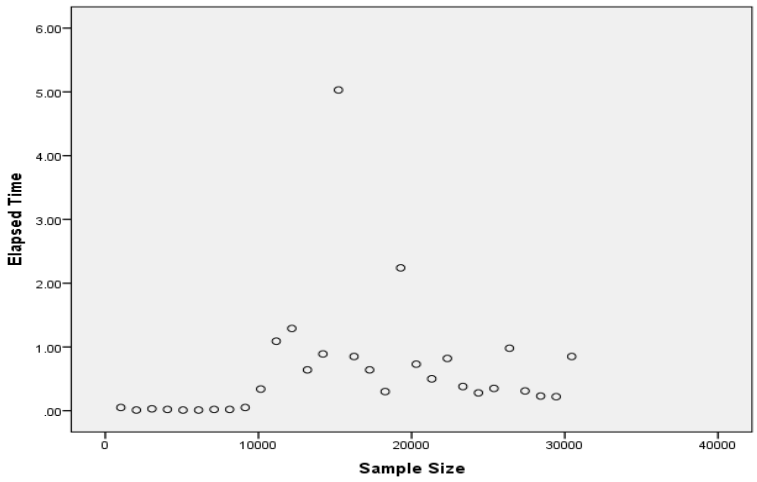


Table (4):Correlation coefficient

hypothesis	Correlation coefficient	p-value	Decision
There is no statistically significant relationship between time and sample size	0.477	0.359	Weak positive linear relationship

Through the linear correlation coefficient value, we observe that the correlation between the elapsed time and sample size is weak. And through the p- value (0.359) we note that it's greater than 0.05, which leads to the acceptance of null hypothesis and therefore there is no statistically significant relationship between time and sample size

-In order to test the difference between the median of elapsed time in small samples (1-15) and the median of elapsed time in large samples (16-30), the Mann-Whitney U test will be used because elapsed time does not follow normal distribution

Table (5):Statistics

The Measures	Mean	Median	Standard deviation	Lowest value	Highs value
Small Samples	0.6333	0.0500	1.29453	0.1	5.03
Large Samples	0.36543	0.5000	0.51226	0.22	2.24

Table 7 shows descriptive measurements of small and large samples

Table (6):*Nonparametric Tests: Mann-Whitney U test.

Hypothesis	p-value	Decision
The median of elapsed time in small samples equals the median time in large samples	0.143	Accept the null hypothesis

Table(6) shows that the p- value (0.143) is greater than 0.05. That means accept the null hypothesis (The median of elapsed time in small samples equals the median time in large samples or there are no significant differences between the median elapsed time in the small samples and the median elapsed time in the large samples).

5.Conclusion. To verify the hypothesis of this study that elapsed time is not affected by increasing the size of the sample, the samples were divided in to two groups large sample and small samples. Using the hypothesis tests it was found that the mean time for the small samples is not significantly different from the medium of the large samples. Also, using the correlation coefficient method to measure the degree and direction of the relationship betweenelapsed time and sample size, it was found that the relationship between elapsed time and sample size is a weak positive relationship. From the above, it can be said that when using the K-means method in cluster analysis, the sample size does not affect the elapsed time, which means that the efficiency of the k- means method is not decrease in the case of large samples.

6.REFERENCES

[1] Ghosh, Soumi, and Sanjay Kumar Dubey. "Comparative analysis of k-means and fuzzy c-means algorithms." International Journal of Advanced Computer Science and Applications 4, no. 4 (2013).
[2] Sawant, Kedar B. "Efficient determination of clusters in k-mean algorithm using neighborhood distance." The International Journal of Emerging Engineering Research and Technology 3, no. 1 (2015): 22-27.
[3] Velmurugan, T. "Efficiency of k-means and k-medoids algorithms for clustering arbitrary data points." Int. J. Computer Technology & Applications 3, no. 5 (2012): 1758-1764.
[4] Napoleon, D., and P. Ganga Lakshmi. "An efficient K-Means clustering algorithm for reducing time complexity using uniform distribution data points." In Trendz in Information Sciences & Computing (TISC), 2010, pp. 42-45. IEEE, 2010.
[5] Rauf, Azhar, Saeed Mahfooz Sheeba, Shah Khusro, and Huma Javed. "Enhanced k-mean clustering algorithm to reduce number of iterations and time complexity." Middle-East Journal of Scientific Research 12, no. 7 (2012): 959-963.
[6] Aldenderfer, Mark S., and Roger K. Blashfield. "Cluster analysis: Quantitative applications in the social sciences." Beverly Hills: Sage Publication (1984).
[7] Bernstein, Ira H. Applied multivariate analysis. Springer Science & Business Media, 2012.
[8] Everitt, Brian. "Cluster Analysis. Second." (1980).
[9] Kaufman, Leonard, and Peter J. Rousseeuw. Finding groups in data: an introduction to cluster analysis. Vol. 344. John Wiley & Sons, 2009.
[10] Everitt, Brian, and Torsten Hothorn. An introduction to applied multivariate analysis with R. Springer Science & Business Media, 2011.
[11] Krajewski, Grzegorz, and Danielle Matthews. "RH Baayen, Analyzing linguistic data: A practical introduction to statistics using R. Cambridge: Cambridge University Press, 2008. Pp. 368. ISBN-13: 978-0-521-70918-7." Journal of Child Language 37, no. 2 (2010): 465-470.
[12] Capoyleas, Vasilis, Günter Rote, and Gerhard J. Woeginger. "Geometric clusterings." J. Algorithms 12, no. 2 (1991): 341-356.
[13] Kaufman, Leonard, and Peter J. Rousseeuw. Finding groups in data: an introduction to cluster analysis. Vol. 344. John Wiley & Sons, 2009.
[14] Inaba, Mary, Naoki Katoh, and Hiroshi Imai. "Applications of weighted Voronoi diagrams and randomization to variance-based k-clustering." In Proceedings of the tenth annual symposium on Computational geometry, pp. 332-339. ACM, 1994.
[15] Kaufman, Leonard, and Peter J. Rousseeuw. "Partitioning around medoids (program pam)." Finding groups in data: an introduction to cluster analysis (1990): 68-125.
[16] Duin, Robert PW, and D. M. J. Tax. "Statistical pattern recognition." In Handbook of Pattern Recognition and Computer Vision, pp. 3-24. 2005.
[17] https://aci.info/2014/07/12/the-data-explosion-in-2014-minute-by-minute-infographic/