International Journal of Computational Intelligence Research
ISSN 0973-1873 Volume 13, Number 5 (2017), pp. 899-906

http://www.ripublication.com

Comparative Analysis of K-means and K-medoids Algorithm on IRIS Data

Kalpit G. Soni*[1,2] and Dr. Atul Patel[3]

[1]Ph.D. Research Scholar, CMPICA, Charotar University of Science and Technology, Changa, 388421, India.
[2]Assistant Professor, Shri Alpesh N.Patel P.G. Institute, Anand., 388001, India.
[3]Dean and Principal, CMPICA, Charotar University of Science and Technology, Changa, 388421, India.
**Corresponding author*

Abstract

Clustering techniques are important methods for the examination of data, predictions based on the examinations and for eliminating the discrepancies observed in them. Iterative techniques are used to group dataset which forms part of a cluster as per collateral and identical characteristics. Clustering is a very useful technique for identifying and grouping the ever growing amount of data generated on daily basis and to generate the patterns and knowledge that can be exploited further. In this paper, we strived to compare K-means and K-medoids algorithms using the dataset of Iris plants from UCI Machine Learning Repository. The results obtained were in favour of K-medoids algorithm owing to its ability to be better at scalability for the larger dataset and also due to it being more efficient than K-means. K-medoids showed its superiority over k means in execution time, sensitivity towards outlier data and to reduce the noise since it employs the method of minimization of the sum of dissimilarities of datasets.

Keywords: Clustering Method, K-means, K-medoids, IRIS Dataset.

INTRODUCTION

The process of grouping objects into clusters such that the similar ones occupy same group and the dissimilar ones into other group is called as Clustering. It's an important method of segregating various objects in a way that the homogenous data occupies the

same cluster while the heterogeneous are placed in another cluster. [1]Clustering has gained wide usage and its importance has grown proportionally because of the ever-growing amount of data and exponential increase in computer's processing speeds. The importance of clustering can be understood from the fact that it has a wide variety of application whether in education or industries or agriculture or economics or even in ecological sciences. In the modern technologies of artificial intelligence and pattern recognition too this technique has found its applications.[2] Clustering Techniques have become very useful for large datasets even in social media such as face book and twitter.[3] The clustering techniques are categorised as follows-[4]

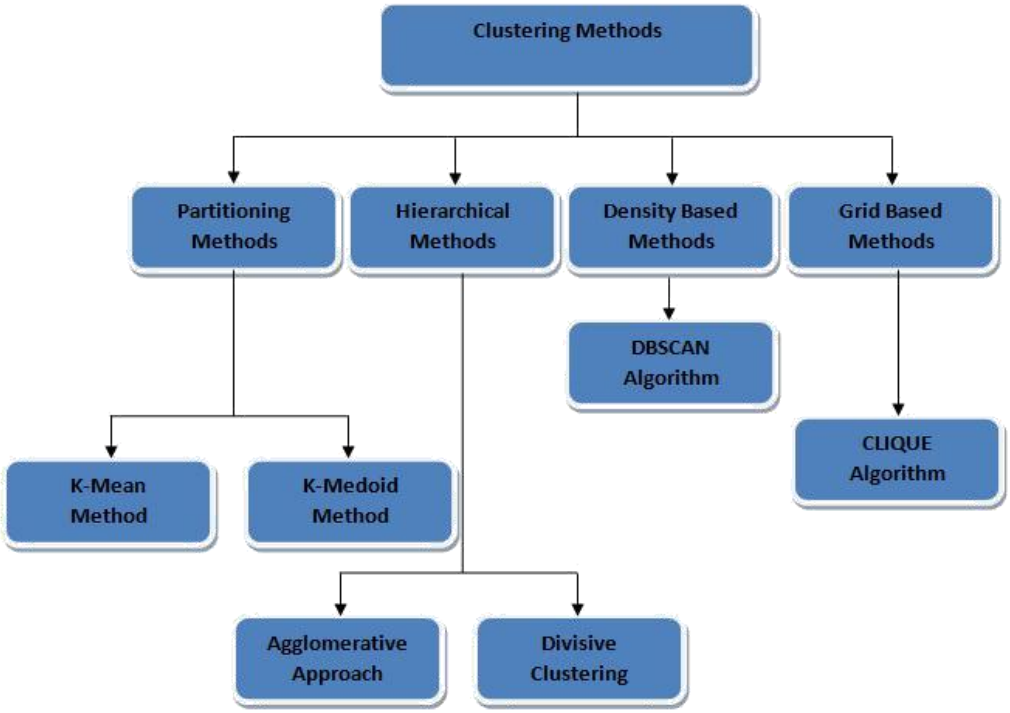


Figure 1: Clustering Methods

In this study, we are focusing on the applicability of Partitioning method of clustering techniques to decide which category of it; K-means or K-medoids method acquires better accuracy in partitioning the data.

PARTITIONING METHODS:

During this method, the large objects are grouped into a cluster with each cluster having at least one element. Partitioning is an iterative process whereupon the objects may be relocated into other groups based on their similarity or relevance. Partitioning is effective when the size of data set is smaller or in mid-size segments. The two

majorly used partitioning techniques are- K-means and K-medoids methods, including some of their variations.[5]

This study strived to compare the quality of results obtained when both K-means and K-medoids methods are used to segregate and cluster the data.

K-Means Method:

K-means clustering technique (or sometimes called Lloyd-Forgy method) was developed by James MacQueen in 1967[6] as a simple Centroid-based method. It is still one of the most widely used algorithms for clustering.[7] In K-means algorithm, the 'n' number of observations is divided into 'k' clusters such that the observations in a cluster are nearest to each other in reference value like cluster mean and the distance of the object. When used in conjunction with other algorithms like Lloyd's algorithm etc, the K-means methods can be applied to large data sets also.[8]

In K-means, the higher value of the distance between clusters is calculated using a standard formula to measure distance which basically gives the similarity of repetitive data. It is a faster clustering method comparing when used along with its variations.

K-Means Methodology:

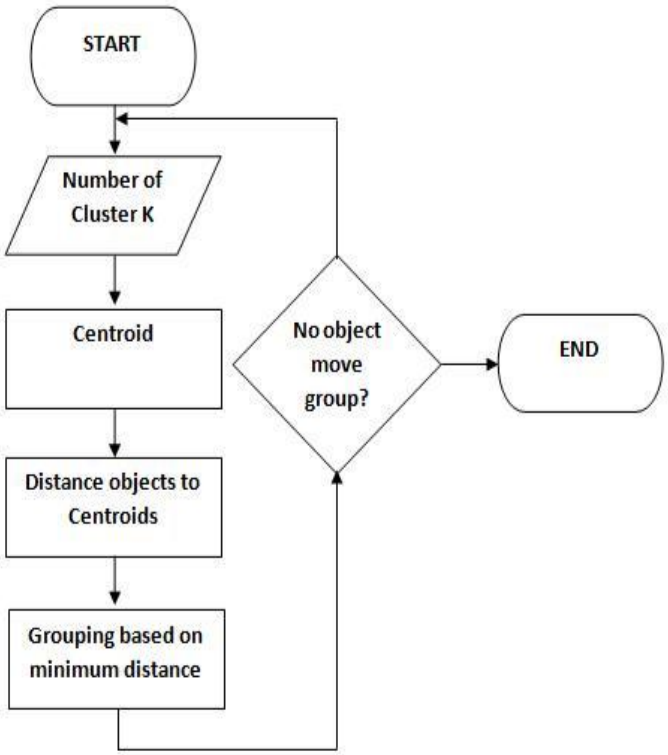
The data in K-means are classified in advance into K clusters to define the k-centroid value of each cluster. The location of Centroid is of paramount importance since it may give different results when the farther they are the better it is. In the subsequent steps, the data points that belong to a set are moved towards the nearest centroid so that no point remains unmoved.[9] The new k centroids are recalculated many times over so that the dataset belonging to one cluster may switch into another cluster at the time of new clustering. This process is repeated until no possibility of switching over of dataset remains.

The Euclidean distance between an object and all the nearby centroid is calculated as per the formula-[10]

$$j = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

Where $||Xi^{(j)} – Cj ||^2$ is the nearest distance measure between a data point xij and the Centroid Cj, and it indicates the distance between data points from their Centroid. The time complexity of the K-means algorithm is subjected to the formula; $O(n^{dk+1})$.

K-Means Algorithm:



Flowchart of K-means Algorithm

During this technique, a dataset D containing an object and k number of clusters are taken for partitioning and the result obtained is stored in the K-clusters present in A set.

This algorithm is performed in following steps-[11]

Step 1: The initial centroids are prepared by placing k number of points containing the objects that are to be clustered.

Step 2: The nearest Centroid is the group of each object moved.

Step 3: The recalculation of k- centroids is performed in the case of all the objects that has been allotted a group.

Step 4: All the procedure of allocation of centroids are repeated until there remains to movable centroids. This will result in the formation of groups from which the reference metric can be minimised.

The K-means technique pays emphasis to the initial centroid taken as reference points. Hence in order to avoid chances of taking a wrong centroid, the process is

repeated many times by taking different centroids. This property makes it a very good method for working with random data points.

Limitation of K-means:

Because of an object's ability to disorient the distribution in the case of an extremely large or extremely small value of a dataset, the K-means algorithm is very prone to the effects of outliers.[12] This method assumes all the clusters have an equal number of observations which may not be the case always especially if there is more outlier with extreme values. Since the choice of a centroid is random in this method hence it may result into different centroids when performed many times even in similar conditions restricting the repeatability of results.

To get away with is an issue another clustering technique called K-medoids method is used which is basically a method of representative objects.

K-medoids Method:

K-medoids or Partitioning Around Medoid (PAM) method was proposed by Kaufman and Rousseeuw[13], as a better alternative to K-means algorithm. In this method, before calculating the distance of a data object to a clustering centroid, k clustering centroids are randomly selected from n data objects such that initial partition is made on the basis of closeness of each object to the clustering centroid to begin the partitioning of data. Then, iteration methods are employed continuously till the most appropriate partition value is obtained. In this method, after every iteration, the object from each clustering samples are chosen based on the improvement of clustering quality. The most centrally located object in a cluster is taken as a reference point here which is actually a medoid and not a mean value of elements in a cluster. The basic principle of K-medoids method is that the minimization of the total sum of the distance of dissimilar points from a reference point should be done for partitioning. By empirically taking a representative data from each cluster, total k clusters are taken such that each of the remaining data points is clustered with medoid. This algorithm works effectively for a small dataset but does not scale well for large dataset.

K-medoids Algorithm:

This algorithm is performed in following steps-[14]

Step 1: From a given dataset of n, total K random points are selected as Medoids.

Step 2: Use of any of distance finding metrics, each data point is clustered with its closest medoid.

Step 3: Total Swapping Cost (TCih) is calculated for each data pair having 'i' selected and 'h' not selected objects such that if TCih is less than zero than 'i' is replaced by 'h'

Step 4: Steps 2 & 3 are repeated unless there occurs a point where no more change in medoids can be further done.

The Manhattan distance can be calculated as per the formula:[15]

$$j = \sum_{i=1}^{k} \sum_{p \Omega c_j}^{n} \left\| P - O_j \right\|$$

The time complexity for the K-medoids algorithm is subjected to the formula;

$O(k(n-2)^2)$. The efficiency and performance of the results in the cluster are directly dependent on clustering centre chosen. Hence all efforts to improve this algorithm depend on the which k cluster points are chosen as reference.

EXPERIMENT EVALUATION:

Data Sets:

For the purpose of experimental evaluation of superiority of K-medoids over K-means algorithm, UCI Machine Learning Repository was used which is also a collection of the database that if very often employed by the researchers in the domain of Machine Learning.[16] This database is of special use in the empirical algorithm analysis. From UCI depository the dataset of Iris plant parameters was taken. In all, five attributes were taken from Iris dataset of plants; four were quantitative viz. Sepal width, Sepal length, Petal Width and Petal length while one of the qualitative i.e. class name. One hundred fifty instances, 50 each in three classes, are taken. The three classes were Iris Versicolour, Iris Setosa and Iris Virginica. Although one class was linearly separable from the other two classes, the later two were not linearly separable.

K-means Vs K-medoids Algorithms:

For comparison of both algorithms, Front end Java has implemented the reason being Java (Sun Microsystems) is a modern, robust but simple, portable and object-oriented programming language which is eventually based on C and C++ programming languages. Both K-means and K-medoids are assessed on their approach towards large data set.

The resulting output using k-means and k-medoid algorithm is as follows-

Table 1: Cluster result of iris data by K-means

	Setosa	***Versicolor***	***Virginica***
Setosa	50	0	0
Versicolor	0	47	3
Virginica	0	14	36

Table 2: Cluster result of iris data by K-medoids

	Setosa	***Versicolor***	***Virginica***
Setosa	50	0	0
Versicolor	0	41	9
Virginica	0	3	47

RESULTS:

K-medoids fared better than k means for the clustering accuracy checked against true classes such that the value of K-means is 88.7% and that for K-medoids, 92%. K-means is erratic in the grouping as it mixes objects in Virginica classes with that in Versicolor classes.

CONCLUSION:

The present work aimed to compared K-medoids algorithm and K-means algorithm to check the improved efficiency and scalability of each of these. The results obtained after performing clustering a number of times prove K-medoids superiority of K-means in the execution time, quality clustered classes and also the number of records. The data obtained using K-medoids was compared with K-means using real samples obtained from the reliable repository.

FUTURE SCOPE:

We plan to compare K-medoids algorithm with other established algorithms also to assess chances of further improvement in the study. That will help to improve the efficiency and scalability options by reduction the time for execution of the algorithm.

REFERENCES

[1] H. Jiawei,M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, (San Francisco California, Morgan Kaufmann Publishers, 2012).

[2] C. Zhang, and Z. Fang, An improved k-means clustering algorithm, Journal of Information & Computational Science, 10(1), 2013, 193-199

[3] Madhuri V. Joseph, ―Significance of Data Warehousing and Data Mining in Business Applications, International journal of Soft Computing and Engineering, Vol No:3,Issue no:, March 2013.

[4] Fahad, A, Alshatri, N., Tari, Z., AlAmri, A., Zomaya, Y., Khalil, I., Foufou, S., Bouras, A, "A Survey of Clustering Algorithms for Big Data: Taxonomy & Empirical Analysis," Emerging Topics in Computing, IEEE Transactions on ,vol.PP, no.99, pp.1,1. 2014

[5] Shalini S Singh & N C Chauhan, "K- means v/s K- medoids: A Comparative Study", National Conference on Recent Trends in Engineering & Technology, 2011.

[6] MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. 1. University of California Press. pp. 281–297. MR 0214227. Zbl 0214.46201. Retrieved 2009-04-07.

[7] Anil K. Jain, "Data clustering: 50 years beyond Kmeans", 19th International Conference in Pattern Recognition,2009.

[8] Agrawal, R., Gehrke, J., Gunopulos, D. and Raghavan, P. 1998 Automatic subspace clustering of high dimensional data for data mining applications

[9] T. Soni Madhulatha, An overview on clustering methods, IOSR Journal of engineering, 2(4), 2012, 719-725.

[10] K. S. Kadam and S. B. Bagal, ―Fuzzy Hyperline Segment Neural Network Pattern Classifier with Different Distance Metrics‖, International Journal of Computer Applications 95(8):6-11, June 2014.

[11] Arora, Deepali, Varshney, Analysis of K-Means and K-Medoids Algorithm For Big Data, *International Conference on Information Security & Privacy* (ICISP2015), 2015.

[12] Abhishek Patel,"New Approach for K-mean and K-medoids algorithm", International Journal of Computer Applications Technology and Research,2013.

[13] Kaufman, L. and Rousseeuw, P.J.(1987), Clustering by means of Medoids, in Statistical Data Analysis Based on the –Norm and Related Methods, edited by Y.Dodge, North-Holland, 405-416.

[14] "Data Mining Concept and Techniques", 2nd Edition, by Jiawei Han, Han Kamber.

[15] K. S. Kadam, S. B. Bagal, Y. S. Thakare, N. P. Sonawane, ‖Canberra Distance Metric Based Hyperline Segment Pattern Classifier Using Hybrid Approach of Fuzzy Logic and Neural Network‖, 3rd International Conference on Recent Trends in Engineering & Technology (ICRTET'2014), India, March 28-30, 2014.

[16] Gupta, A.: Classification Of Complex UCI Datasets Using Machine Learning And Evolutionary Algorithms. 4, 85–94 (2015).