International Journal of Latest Trends in Engineering and Technology
Special Issue SACAIM 2017, pp. 434-436
e-ISSN:2278-621X

COMPARATIVE STUDY OF CLUSTERING TECHNIQUES

Clifton Avil D'Souza[1] & Sandesh Mestha[2]

Abstract- Clustering is widely used in Data mining. Clustering is a process of partitioning a set of data into a set of meaningful sub-classes, called clusters. It is used to group a set of similar data objects within the same group based on similarity criteria. There are various clustering algorithms. The objective of this paper is to perform a comparative analysis of two clustering algorithms namely K-means algorithm and K-medoids. These algorithms are compared in terms of efficiency and accuracy.
Keywords – Clustering techniques, K-means, K-Medoids, Data mining.

1. INTRODUCTION

In any problem choosing the suitable approach to analysis of data without proper understating of the data on specific domain and the expected result.in this paper the comparative analysis among two clustering algorithms are reviewed. Clustering techniques is one powerful tool used in Data Mining to explore knowledge. Grouping based on parameters usually varied based on their data and problem domain. There is no exact solution for all data, this research focus on the two techniques which are K-means algorithm and K-Medoids algorithm. Clustering techniques is also known as unsupervised data classification, is an important subject in data mining.it aims to partition a collection of patterns into clusters of similar data points K-Medoids used to obtain each centers to be one-point. K-Means takes random points.in below sections both the algorithms are described and the result has been presented.

2. CLUSTERING ALGORITHM

2.1 K-means Algorithm:

K-means clustering technique is also called as Lloyd-Forgy method, and was developed by James MacQueen in 1967. [2] The k-means algorithm is well known for efficient clustering algorithms. K-means algorithm is sensitive to the dataset that has been taken, but K-Means algorithm is not suitable for large datasets. [3] This algorithm applies a standard distance formula to calculate the similarity of the data in order to get the high inter-cluster distance among clusters. K-means algorithm is used to recognize the hidden patterns that exists in the dataset. Because of that k-means is widely used technique for clustering. [5]

- Determine hard-NP and soft-NP to sort it based on the priority of which attribute being set as main attribute to cluster. Take the first sample as data for analysis.
- Weighted for each attributes and parameters being analyzed. Weight value given after thorough analysis in each parameter being adjusted to its optimal points.
- Centroid is determined for K-means by Euclidian Distance formula.
- The result of Euclidean Distance and repeat until it reaches optimal points.
- The K-Means analysis repeated by using different value of centroid. Then the distance between the data points and Euclidean distance is calculated.
- Once result is obtained, accuracy of the K-Means cluster is calculated. compare it with K-Medoids cluster analysis.

2.2 Limitations of K-Means:

K-Means has the ability to distribute extremely large and extremely small value of dataset. This algorithm fails to handle outlier values. And also K-Means can be used only when the mean values are declared initially. along with number of clusters in advance before partition of points. We use K-Medoids to overcome these issues.

2.3 K-Medoids:

K-medoids or Partitioning Around Medoid (PAM) method was proposed by Kaufman and Rousseeuw. [7] K-Medoids is similar to K-means both are partition algorithms. but K-Medoids algorithm is harder than K-means and because of computing the medoids uses the frequency of occurrence. [4] K-medoids centers are located in the data point.

- Determine the probability of frequency by using the probability occurrences calculation method. The probability of keyword occurrences in a questions being computed. Take the sample of training data first to perform the analysis.
- weight for each attributes and parameters being analyzed. Weight value given after thorough analysis in each parameter being adjusted to its optimal points.
- Cluster the data based on association to the data point with nearest medoid. Calculate the distance measure.

[1] Department of Information Technology and Bioinformatics, AIMIT, St. Aloysius College Beeri, Karnataka, India
[2] Department of Information Technology and Bioinformatics, AIMIT, St. Aloysius College Beeri, Karnataka, India

- Swap the data point for each medoid and non-medoid data. If the distance measure increased, recalculate the value by swap to nearest point to it.
- Once the model of the outcome is generated, rules for data must be set.
- The accuracy of the K-Medoids cluster is calculated to compare with K-Means cluster analysis

3. EXPERIMENT AND RESULT

K-Means and K-Medoids results to reduction in distance between the data points. Each data point represents the parameters. selecting the best technique is impossible when data is dynamic.

3.1 Correlation analysis

Correlation analysis is done in order to find the exact pattern and relationship between one another and determine the effects that has been involved in the two formulas (1) namely Chi-square test and regression analysis.

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad (1)$$

Where,
O is observed attribute.
E is expected attribute.

3.2 K-means analysis

The K-Means analysis which was performed using a random dataset. The figure.1 represents the regression analysis that has been determined. And it has four clusters as optimal clusters in 6 iteration loops. This proves that the best performance that has been done. Moreover, K-means performance is based on the distance between the data points. It is a faster clustering method comparing when used along with its variations. [2] That has to be compared between the K-means and K-Medoid algorithms.

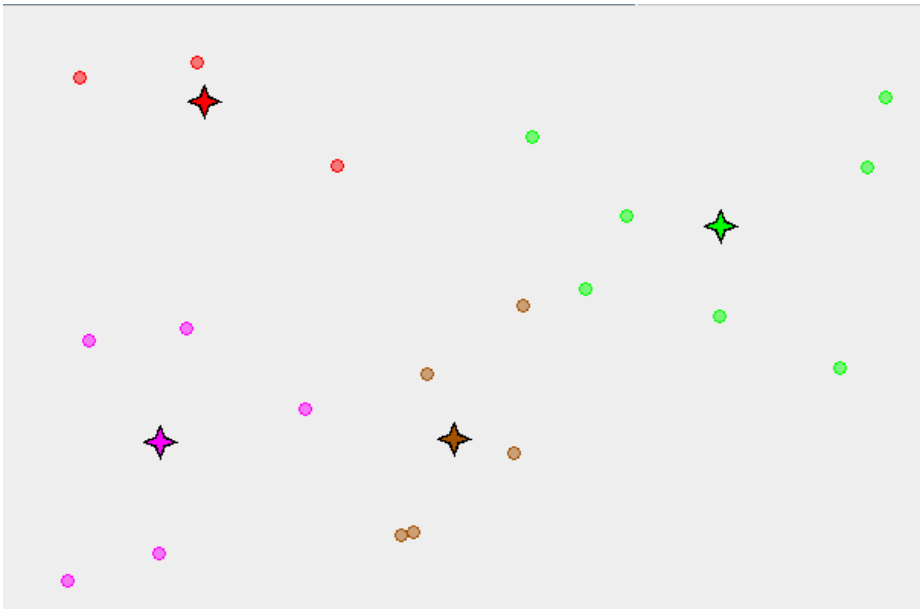

Figure 1 Cluster by using the K-Means

3.3 K-Medoids Analysis

Using the result from correlation analysis-Medoids algorithm performed the same. The results are shown in the figure .2 for each different values there where three different values of K. It decreased the value of swapping points being reversed K-Medoids has the 7 iteration loops. In order to do the efficiency of the difference between K-Means and K-Medoids this was done.

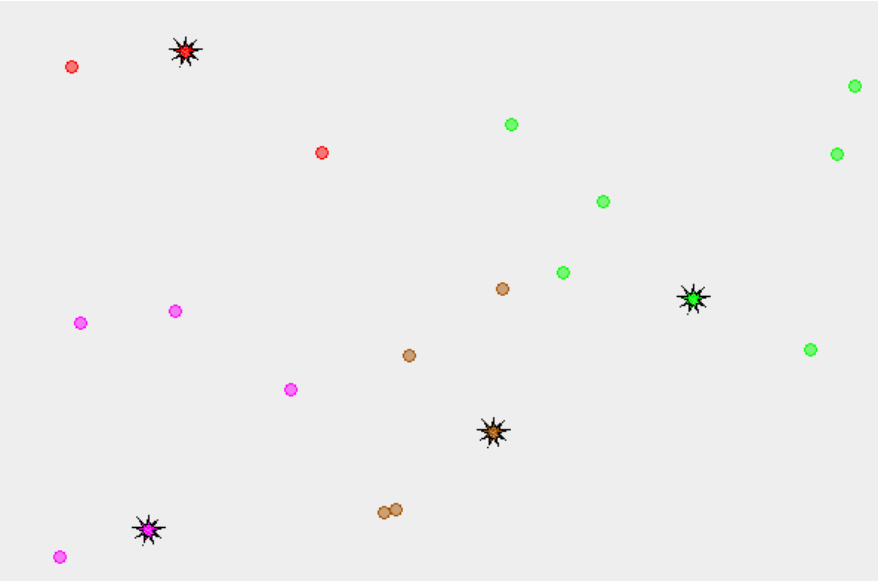

Figure.2 Clusters of K-Medoids

3.4 Comparative analysis

Comparison of Clustering Techniques		
Parameters	**K-MEANS**	**K-MEDOIDS**
Inter-cluster mean distance	LESS DISTRIBUTED	HIGHLY DISTRIBUTED
Iteration to reach optimal	6	7
Time Consumed	HIGH	LOW

Table 1 Comparison of K-Means and K-Medoids

From table 1, for two different approach. K-means shows good result compared to K-medoids with lower iterations loops. However, K-medoids gives best cluster values of inter-cluster mean distance is lower compared to K-means with higher inter-cluster distance. Despite K-medoids lacks performance in inter-cluster. For this dataset K-medoids performance is better than K-means

4. CONCLUSION

K-Means and K-Medoids analysis has been done using random datasets. Both were approaches implemented with Java and the result is good with minimal errors. Based on this datasets, K-Medoids was found best as per our analysis and it consumes less time. For further analysis, dataset of different domain must be used to test the efficiency of algorithms. Different types of dataset might give different result.

5. REFERENCES

[1] Kahkashan Kouser and Sunita,``A comparative study of K Means Algorithm by Different Distance Measures``In IJIRCCE Vol. 1, Issue 9, November 2013.

[2] Kalpit G. Soni and Dr. Atul Patel," Comparative Analysis of K-means and K-medoids Algorithm on IRIS Data "In International Journal of Computational Intelligence Research, ISSN 0973-1873 Volume 13, Number 5 (2017), pp. 899-906.

[3] Shivani Vishwakarma, Dr. Pramod S Nair, D. Srinivasa Rao," A Comparative Study of K-means and K-medoid Clustering for Social Media Text Mining",In Internamtional Journal of Advance Scientific Reseach and Engineering Trends,Vol 2,2015.

[4] Norazam Arbin, Nur Suhailayani Suhaimi, Nurul Zafirah Mokhtar, and Zalinda Othman," Comparative Analysis between K-Means and K-Medoids for Statistical Clustering",In Third International Conference on Artificial Intelligence, Modelling and Simulation,2015.

[5] Bashar Aubaidan, Masnizah Mohd and Mohammed Albared," Comparative study of K-means and K-mean++ Clustering Algorithms on Crime Domain", In Journal of Computer Science, ISSN: 1549-3636,2014.

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

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