IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Performance Evaluation Of K-Means and K-Medoids Clustering Techniques on a Server Log File

Charu Sharma
Himachal Pradesh University, Shimla, India.

Abstract

In an era where the use of technology has evolved manifolds has resulted in the availability of information globally. Due to the multiple availability of data over the internet, retrieval of desirable information has become a tedious and a time consuming task. Hence, use of various Data Mining techniques has become necessary for all the websites to allow the needed information to be accessed by the user. This resulted in the concept of Website Personalization, for which various Data Mining techniques are used, which further means to customize the website as per the need of almost every user visiting it. One of the commonly used technique in Website Personalization is Clustering. Clustering is an unsupervised learning technique and is the process of grouping similar objects into different groups based on their centroid distance in a data set. The two most widely used techniques of clustering are K-Means and K-Medoids. In this paper, the concept of clustering is studied and the performance of both the techniques is analyzed graphically. The paper analyzes the performance using Average Within Centroid Distance and Davies Bouldin on the Server Log Data.

Keywords: Data Mining, Website Personalization, Cluster Analysis, Euclidean Distance, K-Means, K-Medoids.

1. Introduction

Data Mining is the process of mining the data to retrieve desirable information from the larger data set. Hence, data mining is considered as a synonym to “knowledge discovery in databases”[1]. There various techniques to perform data mining, amongst them the most widely used technique is Clustering. Clustering is an unsupervised learning technique which groups similar objects into clusters, based on their centroid distance and which show common characteristics. Hence, the objects belonging to two different clusters are not similar to each other. There are three broad techniques of Clustering which are: Partitioning, Hierarchal and Model Based[9].

In this paper, the two most widely used partitioning techniques of Clustering are studied i.e. K-Means and K-Medoids which are Centroid Based Clustering Algorithms in which clusters are formed on the basis of least centroid distance measured from a central vector[2,3]. This leads to the concept of Optimization in which the K value of the algorithms is optimized. Each step of the algorithms is reiterated (n) number of times so as to form the clusters which have least centroid distance, where (n) is the number of iterations done to form a cluster centroid.

There are various distance measures used to compute the distance of the centroid such as Euclidean Distance, Manhattan Distance, Chebyshev Distance, etc.[2]. In this study, Euclidean Distance is used as distance measure to identify the performance of the algorithms for different values of K.

In this learning, to analyze the performance of the Clustering Algorithms, Cluster Analysis is performed on a Server Log data which is a single server data having access of multiple users of a Himalayan Rider Website. In order to perform clustering and thereby analyzing the performance of the Centroid Based Clustering Algorithms, a tool is used naming Rapid Miner.

2. Problem Definition

With the advent of the use of the technology and the Internet for multiple purposes, there has become an urgent need to fulfill the information requirement of almost of every user accessing a website, thereby resulting in the use of various Data Mining Techniques on the large data set to provide the needed data to its users. There are many Mining techniques which may outright the purpose, amongst all, the most widely used is the Clustering Technique which is an Unsupervised Partitioning Technique. This method is used because the information is huge enough to identify the classes of such large dataset. Therefore, there is an urgent need to understand the performance of the widely used Partitioning Techniques i.e. K-Means and K-Medoids so that the efficiency of the information accessed is improvised. Hence, this paper studies the performance of these two algorithms for different values of K, by calculating the distance from the centroid, on the Server Log data of the website so as to enhance the availability of information on the website.

3. K-Means Clustering

K-Means Clustering is one of the widely used Unsupervised Partitioning Clustering Technique which was developed in 1967 by MacQueen and further modified in 1975 by Hartigan and Wang. The approach of K-Means is to take an input parameter, value of ‘K’ and the Distance Measure. The algorithm will be executed forming K number of clusters in a way that the intra cluster similarity is high and inter cluster similarity is low thereby having the least Centroid Distance from the object taken as a centroid[2]. To measure the Centroid Distance, this paper uses Euclidean Distance as the measure. The algorithm is as follows:

i. Partition the data set into K subsets.

ii. Compute the Centroid Points which is the mean distance among the points in the K number of clusters.

iii. Assign each object to a subset such that the object has least distance from the chosen Centroid point. This will separate the objects which are similar with the objects to that which are dissimilar due to certain characteristics.

iv. Repeat (ii) till the centroid selected is the most appropriate and cluster formed has high intra cluster similarity[2,3,4].

4. K-Medoids Clustering

K-Medoids (also called as Partitioning Around Medoid) Clustering Algorithm was proposed by Kaufman and Rousseeuw in 1990[2]. It is categorized as a supervised learning algorithm as it allows the user to explicitly select the value of ‘K’. The main aim of the algorithm is to minimize the dissimilarities with all the other points in the cluster, thereby increasing the intra cluster similarity. In this technique, the medoid is the squared distance calculated among all the other points in the cluster unlike K-Means where the centroid is the mean or median distance among all the points in a cluster. The distance measure used in this study is Squared Euclidean Distance. The algorithm is as follows:

i. Select K initials medoids which are assumed to be the centre of all the ‘n’ data objects within a cluster.

ii. Compute the squared distance from the selected medoid to all the other non-selected points in a cluster.

iii. Select the distance which has a minimum value and consider that as a medoid.

iv. Repeat (ii) until a minimum distance medoid is selected which has least dissimilarity with all the others points of a cluster[6].

5. Distance Measure Methodology

Distance Measure is a measure of similarity or dissimilarity among the data points in a cluster. This defines the points which have a least value for the corresponding distance measure which will further have more similarity and will be chosen as a centroid or medoid point for a cluster[6,7]. The most widely used distance measures are:

i. **Euclidean Distance** : This is a most widely used distance measure which is based on Pythagorean Formula. This distance is computed by taking the root of squared differences of distance between the two points in a cluster. The distance is computed as follows:

$$d(p, q)^2 = (q_1 - p_1)^2 + (q_2 - p_2)^2 + \ldots\ldots + (q_n - p_n)^2$$

where n are the number of data points in a cluster.

ii. **Squared Euclidean Distance**: The distance metric is metric is the same equation as the Euclidean distance metric, but the distance is not computed by taking the square root of the squared differences between the two data points in a cluster. The formula for computing the distance using this measure is as follows:

$$d\ (p, q) = (q_1 - p_1)^2 + (q_2 - p_2)^2 + \ldots\ldots + (q_n - p_n)^2$$

where n are the number of data points in a cluster.

iii. **Manhattan Distance**: It is also called as a City Block Distance and is less sensitive to outliers than the Euclidean Distance measure. This metric simply calculates perpendicular distance between the two corelated pair of data points in a cluster. Suppose we have two points P and Q in X-axis and Y-axis. The data points are: P (x1, y1) and Q (x2, y2), the formula to compute the distance is as follows:

$$P \text{ and } Q = |x1 - x2| + |y1 - y2|$$

iv. **Chebyshev Distance**: It is also known as Tchebyshev Distance is a maximum measure metric, where the distance between two vectors is the greatest of their differences along any coordinate dimension. The

Chebyshev distance between two vectors or points x and y, with standard coordinates $x_i$and y_i , respectively, is computed as follows:

$$d(x,y) = \max_{i=1,2,\ldots n} |x_i - y_i|$$

where n are the number of data points in a cluster[3,6,7].

6. Performance Measure

In order to measure the Performance of K-Means and K-Medoids on the Server Log Data, the paper uses the Rapid Miner Tool and the following measures are considered for the evaluation:

1. **Average Within Centroid Distance**: The average within cluster distance is calculated by averaging the distance calculated using the Distance Measure between the centroid and all data points of a Cluster.

2. **Davies Bouldin:** This measure will compute the ratio of low intra cluster similarity and high intra cluster similarity. The algorithm which produces clusters with low intra-cluster distances (high intra-cluster similarity) and high inter-cluster distances (low inter-cluster similarity) will have least value of the measure and will be used for considering the number of cluster. In this paper, the value is optimized for different values of K to get least value of the measure for K number of clusters.

7. Related Works

Usage of Data Mining Techniques for exploring the pattern of data and identifying the correct information needed, has become a necessity to identify which clustering technique should be used to get the correct correlation in an information. Today, Data Mining is not only applied in static data but also in a Server Log Data so as to identify the pattern of data which the user is accessing on any website. This has resulted an increase of research field in which the performance of Data Mining Techniques is studied i.e. the content of a website. Hence, Santosh Nirmal[2] performed a comparative study of K-Means and K-Medoids Clustering Techniques on a sample dataset of different sizes. In the study, the author compares the performance using Java and C++ language on larger data set. The paper concluded that the K-Medoids algorithm is better in performance to K-Means in handling of outliers and noisy data. Another study was done by Rani Patel, Ashish Tiwari[1], in which the authors proposed an algorithm for efficient Centroid Selection for K-Means Clustering Technique. In the paper, a method for K-Means clustering is proposed in which radial basis function (RBF) kernel also known as Gaussian kernel is added to the original K-Means algorithm. The paper summarizes that the modified version of the algorithm provides efficient centroid distance calculation and handles the data accurately consuming less time. There are various Clustering Techniques to unearthen the information, thus, another research which was done by Santosh Kumar Uppada[3], which compared and studied the Centroid Based Clustering Algorithm on the basis of various parameters such as Sensitive to Noise, Outlier Structure, Centric, Minimize Intra-cluster Variance and Complexity. The paper concluded that in order to select any Clustering Technique for performing clustering on any type of dataset, the Distance parameter is the most vital which should be calculated.

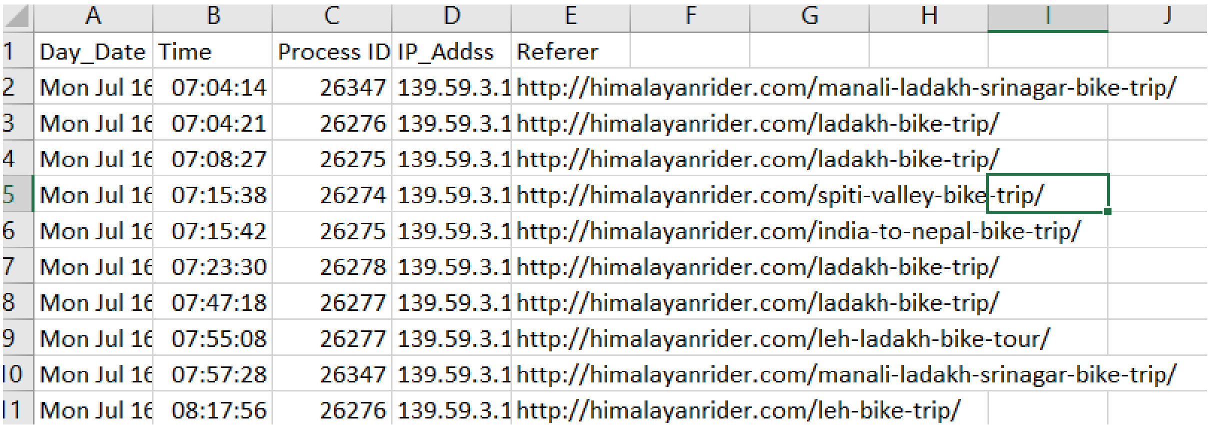
8. Analysis and Results

8.1 Analysis

The paper performs Data Pre-processing on the Server Log Data which is a Server Log File of a Himalayan Rider Website. The data is used to analyse the performances of K-Means and K-Medoids in order to select appropriate technique to identify user access pattern on the website.. The Clustering Technique which has the better performance for the respective value of K can further be employed for Website Personalization[10]. The paper uses Euclidean Distance and Squared Euclidean Measure as Distance Measure for different values of K in the range of (3 – 10) given as an input using the Rapid Miner Tool for performing Clustering. The two measures used to analyse the performance of the algorithms are Average Within Centroid Distance and Davies Bouldin which calculates a centroid in a cluster which has high intra cluster similarity and low inter cluster similarity.

8.1.1 Data Set

The data is extracted from a Server Log File of a single website which is transformed into a structured format with distinct attributes using PHP code implement in XAMP Package to construct Excel File of 150KB having 5000 instances. The data is of one month having user access on the website.



	A	B	C	D	E
1	Day_Date	Time	Process ID	IP_Addss	Referer
2	Mon Jul 16	07:04:14	26347	139.59.3.1	http://himalayanrider.com/manali-ladakh-srinagar-bike-trip/
3	Mon Jul 16	07:04:21	26276	139.59.3.1	http://himalayanrider.com/ladakh-bike-trip/
4	Mon Jul 16	07:08:27	26275	139.59.3.1	http://himalayanrider.com/ladakh-bike-trip/
5	Mon Jul 16	07:15:38	26274	139.59.3.1	http://himalayanrider.com/spiti-valley-bike-trip/
6	Mon Jul 16	07:15:42	26275	139.59.3.1	http://himalayanrider.com/india-to-nepal-bike-trip/
7	Mon Jul 16	07:23:30	26278	139.59.3.1	http://himalayanrider.com/ladakh-bike-trip/
8	Mon Jul 16	07:47:18	26277	139.59.3.1	http://himalayanrider.com/ladakh-bike-trip/
9	Mon Jul 16	07:55:08	26277	139.59.3.1	http://himalayanrider.com/leh-ladakh-bike-tour/
10	Mon Jul 16	07:57:28	26347	139.59.3.1	http://himalayanrider.com/manali-ladakh-srinagar-bike-trip/
11	Mon Jul 16	08:17:56	26276	139.59.3.1	http://himalayanrider.com/leh-bike-trip/

Image 1: Structured Data Set

In the above image, the data has been converted to the structured Excel format resulted by performing data pre-processing on the log file and thereby, corresponds to one month user access of the data on the website having attributes as Day, Date, Year, IP Address, Time, Process ID, and Referrer.

8.1.2 Rapid Miner Clustering

In order to evaluate the performance of the Clustering Techniques the data is loaded as input in the tool and the Technique is executed for different values of K in the range of 3-10 clusters. In the tool, model is made using modules of the tool to perform the clustering and identify centroid values for corresponding value of K. The table below shows distance values of Clustering Techniques of centroids for the measures i.e. DB means Davies Bouldin and the Avg. Distance means Average Within Cluster Distance.

K Values	K-MEANS Avg. Distance	K-MEDOIDS Avg. Distance	DB K-MEDOIDS	DB K-MEANS
3	-9117334.7	-17593506.65	-0.676	-0.438
4	-4438660.126	-6035041.266	-0.431	-0.451
5	-3176243.568	-4621218.031	-0.454	-0.468
6	-2019145.642	-2569259.046	-0.464	-0.453
7	-1303202.938	-2090971.585	-0.507	-0.448
8	-891412.849	-1399943.505	-0.457	-0.424
9	-709542.2	-1007301.375	-1.081	-0.423
10	-536634.172	-931869.096	-1.207	-0.376

Table 1: Distance Values

The above table 1, describes the values, for 'K' number of clusters. The value of the distance measure is negative as the Tool has set the value as 'False' at default to perform optimization of the result.

8.2 Result

The results are analysed graphically using MS-Excel which is as follows:

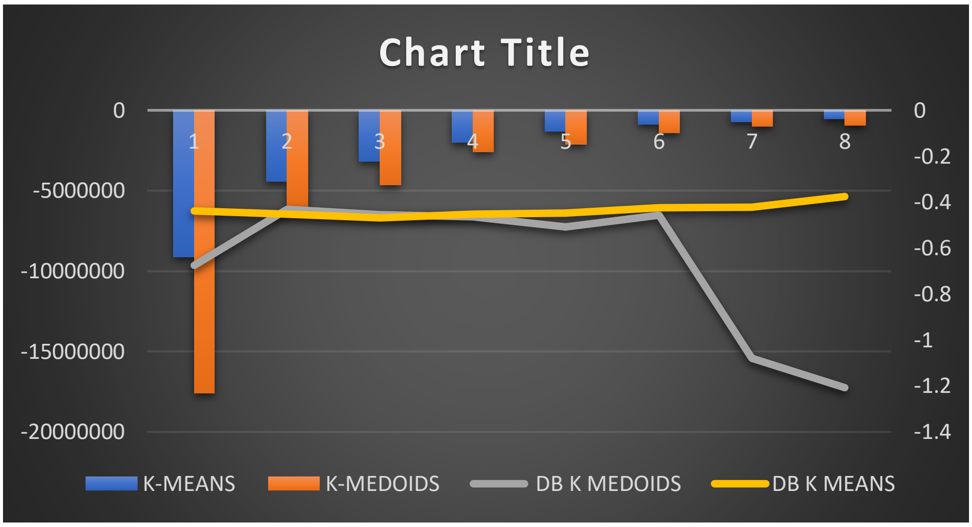


Image 2: Distance Values Comparison

The above image describes, left axis values for Average Within Cluster Distance and right axis values for Davies Bouldin for K-Means and K-Medoids Clustering Algorithms. The values are negative, as the tool has set the default value to 'False' for optimization of the result. The image explains that the least value is of K-Medoids Technique in comparison to K-Means in respect of the Server Log Data for K = 3 in case of Average Within Cluster Distance Measure and for K=10 in case of Davies Bouldin.

9. Conclusion

As the area of usage of Data Mining has increased with the increase in availability of information, there has become a need to explore the performance of various Data Mining Techniques on different data sets so that the correct information is made available to the user. Hence, Data Mining Techniques is applied on the web to enhance the availability of information and make the access of it easy for the user. The paper studies two Centroid based Partitioning Techniques of Data Mining i.e. K-Means and K-Medoids. Also, the paper analyses the performance of the two above mentioned techniques using Average Within Cluster Distance and Davies Bouldin performance measures using the Rapid Miner Tool on the Server Log Data. The study concluded after using the two performance measures that the K-Medoids is better than the K-Means, as the value of K-Medoids Clustering Technique is the most negative for the data set. Hence, this concludes that the K-Medoids is better in performance than K-Means for K=3 as the algorithm handles larger data set well and is not susceptible to the outliers and the noise in the data set. The only drawback of the K-Medoids is the time it takes to perform the clustering. Hence, the paper also concludes that for the Server Log Data, K-Medoids is most appropriate among the two to analyse the user access patterns in order to enhance the availability of information on the website. Thus, for performing Web Usage Mining K – Medoids Clustering Technique should be used as it would provide better clusters and the user access pattern can be evaluated in an enhanced manner.

References

1. Rani Patel, Ashish Tiwari, “Efficient Centroid Selection to Improve k-Means Clustering Performances”, International Journal of Engineering Science and Computing (IJESC), Vol 7 Issue I, 2017.
2. Santosh Nirmal, “Comparative Study between K-Means and K-Medoids Clustering Algorithms”, International Research Journal of Engineering and Technology (IRJET), Volume: 06 Issue: 03, Mar 2019.
3. Santosh Kumar Uppada, “Centroid Based Clustering Algorithms-A Clarion Study”, International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 5 (6), 2014.
4. T.Velmurugan & T.Santhanam, “Performance Analysis Of K-Means And K-Medoids Clustering Algorithms For A Randomly Generated Data Set”, International Conference on Systemics, Cybernetics and Informatics, pp. 578- 583.
5. Kalpit G. Soni and Atul Patel, “Comparative Analysis of K-means and K-medoids Algorithm on IRIS Data”, International Journal of Computational Intelligence Research, Volume 13, 2017, pp. 899-906.
6. Swayal Gultom, S. Sriadhi, M.Martiano, Janner Simarmata, “Comparison analysis of K-Means and K-Medoid with Euclidian Distance Algorithm, Chanberra Distance, and Chebyshev Distance for Big Data Clustering”, 2nd Nommensen International Conference on Technology and Engineering, 2018.
7. "https://www.geeksforgeeks.org/measures-of-distance-in-data-mining/”, Accessed: 8th April, 2020 at 5:00 p.m.
8. “https://en.wikipedia.org/wiki/Cluster_analysis”, Accessed: 8th April, 2020 at 7:00 p.m.
9. “https://shodhganga.inflibnet.ac.in/bitstream/10603/90817/12/12_chapter8.pdf”, Accessed: 8th April, 2020 at 3:00 p.m.
10. S.Kannan1, Dr.G.N.K. Suresh Babu2, Web Personalization Techniques in Data Mining, “Iaetsd Journal For Advanced Research In Applied Sciences”, Volume 5, Issue 1, Jan 2018.