Performance Analysis of KMeans and KMedoids Algorithms in Air Pollution Prediction

S. Suganya, T. Meyyappan, S. Santhosh Kumar

Abstract: Air pollution is a major part of human health problems in many cities. Air pollution can cause many negative effects on the environment. The most basic solution for air pollution is human should have responsible habits and also have to use more efficient devices to predict and control the atmosphere pollution. The nearly everyone important objective of this effort is to analyze and predict the atmosphere smog using data mining techniques. And help to take inevitable carriage steps or decision for protect the future generation from the rapid increase of air pollution. To turn raw data into useful information, Data mining technique is used by many companies. Data mining extract the hidden useful information from the air pollution dataset. It helps and supporting human making decision human, being as responsible, and control avoids the air pollution as claimed by the severe level of pollution in the airspace city based. To analyze and predict simple and efficient clustering techniques such as K-Means and K-Medoids has been used.

Keywords: Air pollution, Clustering, Algorithm K-Means, K-mean Medoids.

I. INTRODUCTION

Air pollution is a mix of gases in the air due to vehicle emission, and emission of harmful chemicals and gases from factory. The impact of health effects of air pollution causes complication in breathing, out of breath, coughing, asthma and aggravation of presented respiratory and cardiac surroundings. The human health effects of unfortunate air quality are far getting, but mainly affect the body's respiratory organization and the cardiovascular arrangement. Mouthful of polluted air leads to a senior complication for asthma and different serious disease. High level of the uses of particle greenhouse gases can cause higher impact of heart problems. The glowing of impression fuel and the leave go of carbon dioxide within the impression are causing the Earth to become warmer. Pollution is the beginning of dangerous contaminants added to air, water or soil. These contaminants can have diverse special effects on entire ecosystems, destruction of living things life cycle of humans, plant and animals.

Data mining is used by companies to turn raw data into useful information. By means of software for extract pattern in large batch of data, business can learn more in relation to their customers to expand more effective advertising strategy add to sales and reduce costs. Researchers make use of data mining approaches like multi-dimensional database, machine learning, soft compute Data mining, the development of discover insists and meaningful patterns and their relations in large volume of data., data revelation and information. It is defined as a development used to take out usable information from a larger volume of raw data information. It insinuated analyzing data patterns in large batches of data by means of much different software. Data mining is also acknowledged as Knowledge Discovery in Data (KDD). Data mining is the development of determine a hidden information from laves data sets. It is incorporates with machine leaning and statistical techniques involving method at the connection of data mining plays significant role in knowledge discovery process in real time application.

Data methods divide data points to segments based on the similarity in data points. The segmentation is a have five process when the data points are allowed to reach its closest data points. The quality of clustering is build on the bury cluster comparison between the data elements has intra cluster relationship among related clusters. A good clustering method will bring into being high quality clusters in which. The value of a clustering technique is also evaluated by its ability to learn the concealed patterns. In simple words, the aim is to set apart segments with similar qualities and allocate them into clusters. Clustering is a process of unsupervised learning and is a common technique for numerical data analysis used in a lot of fields.

In Data Science, we can use clustering analysis to increase some valuable insight from our data by bearing in mind what segments the data point fall into when we be relevant a clustering algorithm. Clustering can be measured the most significant unsupervised knowledge problem; so, as every other difficulty of this kind type, it agreement with judgment a composition in a collected works of unlabeled data. A movable meaning of clustering could be “the development of organize objects into segments whose member are similar in a little way”. K-means clustering is one of the simplest and fashionable unsupervised machine knowledge algorithms. In supplementary words, the K-means algorithm identify k number of centroids, and next allocates each one data spike to the nearby cluster, while maintenance the centroids as little as achievable. Clustering is useful for search data. If there are many belongings and no understandable groupings, clustering algorithms can be nearly new to find accepted groupings. Clustering can also serve as a useful data-preprocessing step to recognize consistent segments on which to manufacture supervised models.
There stay alive a large figure of clustering algorithms in the text. The preference of clustering algorithm depends both on the classification of data obtainable and on the exacting purpose and application.

II. RELATED WORK

C. Kondal raj contributed a study about the difference between the three algorithms K–Means, K-Medoids and DBSCAN. The timing chart denotes that the time taken to execute the result of algorithms, it is noted that the DBSCAN algorithm require very minimum time compared with k means and k Medoids without noise. Finally, this research work conclude that the calculation time of DBSCAN algorithm is with a reduction of the k-Means along with k-Medoids algorithm from now, the organization of DBSCAN algorithm is improved than the others and it is one of powerful tools for discovering arbitrary-shaped clusters in big spatial databases. [1]

Dr. Aishwarya Batra given that measuring similarity connecting data items is effortless than mapping data objects to data point in feature breathing space, this pair wise comparison based clustering algorithms can to a great extent reduce the struggling in increasing clustering based pattern acknowledgment applications. The limitation in k-means algorithm is its execution time and cluster efficiency with another limitation is it works less efficient with large data sets. [2]

Santosh Nirmal proposed that the k-medoids is much costlier method than K-Means. The author also stated that the k-means attains better accuracy tolerance than k-medoids the study is base on the detachment measures applier in the two clustering algorithms. They suggested to use other two distance metrics called Euclidean and Manhattan for better performance. [3]

Clustering, Fuzzy C-means Clustering Algorithm, Hidden Markov Models, Hierarchical methods, Partitioning method, Density-based method, Grid-based method, and Model-based method Observation Sequence Generation based on methods. In our work, we have Air Quality Data Set. From this data set only two attributes has taken for this work. Such as Air Quality parameters are Air Quality (AQ) values of NO2, SO2. For the analytics purpose the most popular clustering technique such as k-means and k-medoids are used in this work. Centroid-Based Method: The K-Means technique Basic Algorithm the K-means Clustering Technique is very simple and we right away begin with a description of the essential algorithm.

III. METHODOLOGY

3.1. Data set

In the Proposed research work experimented using Air Quality data set which is collected from GOVERNMENT OF INDIA, National Centre for Medium Range Weather Forecasting, Ministry of Earth Science (NCMRWF), Delhi. This data set contains totally seven attributes; these are collected during the year 2011. From this data set only two attributes has taken for this work. Such as Air Quality parameters are Air Quality (AQ) values of NO2, SO2. Remaining attributes are having so many null values. But NO2 and SO2 only present in this collect data set. Which means NO2 and SO2 are more present in the air. And NO2 and SO2 are more dangerous for living things.

3.2 The Proposed Work for Implementation

Collected data set is preprocessed and analyzed using Big Data Tools and Techniques. For the analytics purpose the famous clustering technique such as k-medoids and k-means are used. The famous R Tools is used for implementation of the clustering techniques adopted. Sample data is given away in Table 1.

Table 1. Sample Data Set

| Initial Point | NO2 | SO2 |
|---------------|-----|-----|
| A1            | 9   | 4   |
| A2            | 10  | 4   |
| A3            | 29  | 5   |
| A4            | 9   | 4   |
| A5            | 18  | 4   |
| A6            | 10  | 4   |
| A7            | 22  | 6   |
| A8            | 11  | 6   |
| A9            | 9   | 4   |
| A10           | 11  | 6   |
| A11           | 20  | 13  |
| A12           | 13  | 6   |
| A13           | 14  | 6   |
| A14           | 15  | 6   |
| A15           | 20  | 9   |

Data Set- 15 Objects

3.3. The proposed work:

Collected data set has been pre-processed by removing the null values using R Tools and Techniques. Pre-processed data has been clustered by K-Means and also K-Medoids techniques. Clustering technique is used to group the similar objects. The main center of this occupation is to extract the hidden information from the collected Air Pollution dataset. And evaluate the presentation of the K-Medoids and K-Means clustering techniques.
3.4 K-Means

K-Means is the individual of the majority famous unsupervised learning algorithms. K-Means clustering system is group the similar objects or data points based on the selected centroid point of the data. In this proposed work 3 different clusters generated using R Tool. The Euclidean disinterest between the centroid and all the nearest data points are calculated by the formula

\[ j = \sum_{i=1}^{n} \sum_{j=1}^{n} \| x_i^{(j)} - c \|^2 \]

Where \( \| x_i \| \) is the distance between the Centroid Cj and a data point xij.

Data points in the K-means are grouped into K clusters. Here K- indicates the centroid value of each cluster. Position of the Centroid is place the vital role since it can give different results. To switch the data points belong to one cluster into another cluster, again new k centroids calculated more times. This operation is repeated until there is no possibility of switching over of dataset remains.

**Table2. K-Value (Elements)**

| No.of Clustering | K-Means | K-Medoids |
|------------------|---------|-----------|
| 1                | 36.66   | 17        |
| 2                | 50.75   | 19        |
| 3                | 110.83  | 15        |

It’s K-Means clustered and K-Medoids clustered Computational of values.

**Table3. K-Value and Grade Value**

| S.No | K-Means | Grade | K-Medoids | Grade |
|------|---------|-------|-----------|-------|
| 1    | 36.66   | 17%   | 17        | 33%   |
| 2    | 50.75   | 33%   | 19        | 37%   |
| 3    | 69.2    | 50%   | 15        | 30%   |

K-Means clustered and K-Medoids clustered Computational of K-values and G-Value.

3.5 K-Mean Algorithm

| Initial Point | Dataset | Cluster 1 (75.5) Dist Mean1 | Cluster 2 (34.5) Dist Mean 2 | Cluster 3 (42.25) Dist Mean 3 | Cluster |
|---------------|---------|-------------------------------|-------------------------------|-------------------------------|---------|
| A1            | 09      | 66.5                          | 25.5                          | 33.25                         | 2       |
| A2            | 10      | 65.5                          | 24.5                          | 32.25                         | 2       |
| A3            | 29      | 46.5                          | 5.5                           | 13.25                         | 2       |
| A4            | 09      | 66.5                          | 25.5                          | 33.25                         | 2       |
| A5            | 18      | 57.5                          | 16.5                          | 24.25                         | 2       |
| A6            | 10      | 65.5                          | 24.5                          | 32.25                         | 2       |
| A7            | 22      | 53.5                          | 12.5                          | 20.25                         | 2       |
| A8            | 11      | 64.5                          | 23.5                          | 31.25                         | 2       |
| A9            | 09      | 66.5                          | 25.5                          | 33.25                         | 2       |
| A10           | 11      | 64.5                          | 23.5                          | 31.25                         | 2       |

**Algorithmic steps for k-Means Clustering**

Let \( X = \{ a_1, a_2, a_3, \ldots, a_n \} \) be the collected works of data point and \( Y = \{ D_1, D_2, \ldots, D_k \} \) beginning point to the near centroid.

1. Go for K point as the first centroids.
2. Assign each and every one point to the neighboring centroid.
3. Recomputed the centroid of everyone come together.
4. Do again splitting distance 2 and 3 till the centroids don’t change.

From the Table, 4. Can understand that, in k-means

| S.No | K-Means | Grade |
|------|---------|-------|
| 1    | 36.66   | 17%   |
| 2    | 50.75   | 33%   |
| 3    | 69.2    | 50%   |

In this K-Means values and Grade graphing

**Fig2. Working of k-Means**

3.6. K-Medoids Algorithm

It is also the most well-known clustering algorithm. K-Medoids keep down the separation between the selected centroid and all other data points. In case of K-Means it is minimize the square distance between the centroid and all other data points. In K-Medoids, Medoids indicates the centroid point. Like K-means, in K-Medoids for every clustering new Medoids recalculated until there is no Medoids to move.
The Manhattan distance can be computed using the formula:

\[ d = \sum_{i=1}^{k} \sum_{p \in \Omega_c} || p - o || \]

| Initial Point | Data Set | Distance from C1=(9,4) | Distance from C2=(18,4) | Cluster |
|---------------|----------|------------------------|------------------------|---------|
| A1            | 09       | 4                      | 0                      | 1       |
| A2            | 10       | 4                      | 1                      | 8       | 1       |
| A3            | 29       | 5                      | 21                     | 12      | 2       |
| A4            | 09       | 4                      | 0                      | 9       | 1       |
| A5            | 18       | 4                      | 9                      | 0       | 2       |
| A6            | 10       | 4                      | 1                      | 8       | 1       |
| A7            | 22       | 6                      | 15                     | 6       | 2       |
| A8            | 11       | 6                      | 4                      | 9       | 1       |
| A9            | 09       | 4                      | 0                      | 9       | 1       |
| A10           | 11       | 6                      | 4                      | 9       | 1       |
| A11           | 20       | 13                     | 20                     | 11      | 2       |
| A12           | 13       | 6                      | 6                      | 7       | 1       |
| A13           | 14       | 6                      | 7                      | 6       | 2       |
| A14           | 15       | 6                      | 8                      | 5       | 2       |
| A15           | 20       | 9                      | 16                     | 7       | 2       |

Cost Calculation

3.7 Algorithmic steps for k-Medoids clustering:

1. Select K beginning points. These point medoids and are calculated to be extremity of their clusters.

2. Think about the effect of replace one of objects (Medioids) with individual of objects. This is finished in the coldness of every non-selected position from the medoid is considered, and this detachment is points.

Table 5. k-Medoids

| S.No | K-Medoids | Grade |
|------|-----------|-------|
| 1    | 17        | 33%   |
| 2    | 19        | 37%   |
| 3    | 15        | 30%   |

In this K-Medoids values and Grade graphing

Table 6. k=2 K-Mean and K-Medoids Clustering k=2

IV. RESULT AND DISCUSSION

After doing in the analysis three different clusters received for the k-Mean techniques. For the K-Medodies also three different clusters received. In K-Means Technique, Accuracy is calculated by finding the average of the three clusters. For the K-Medoids also Accuracy is calculated by the same method followed in k-Means technique.

Table 7. Comparison connecting K-Means and K-Medoids

| S.No | Clustering   | Accuracy |
|------|--------------|----------|
| 1    | K-Means      | 52.20%   |
| 2    | K-Medoids    | 17%      |

For the K-Medoids also Accuracy is considered by the equivalent process follow in k-Means method.

V. CONCLUSION

The current work aimed to compared K-Medoids algorithm and K-means algorithm to check the improved good organization 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 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 dependable repository.

ACKNOWLEDGMENT

This article has been written with the financial Support of RUSA-Phase 2.0 grant sanctioned vide Letter NO.F.24-51/2014-U,Policy (TN Multi-Gen),Dept of Edn. Govt of India, Dt. 09.10.2018

REFERENCES

1. C.Kondal Raj, “Comparison of K- Means, K Medoids, Dbscan Algorithms Using Dna Microarray Dataset “International Journal of Computational and Applied Mathematics. ISSN 1819-4966 Volume 12, Number 1 (2017) © Research India Publications http://www.ripublication.com
2. Dr.Aishwarya Batra, “Analysis and Approach: K-Means and K-Medoids Data Mining Algorithms”, Ahmedabad, India.E-mail:batra.aishwarya@gmail.com

3. Santosh Nirmal, “Comparative Study between K-Means and K-Medoids Clustering Algorithms” SMaharastra, India. International Research Journal of Engineering and Technology (IRJET) www.irjet.net.

4. C .Kondal raj, “Comparison of K means and K medoids, DBSCAN, algorithms using DNA Microarray dataset, International Journal of Computational and Applied Mathematics, ISSN 1519-4966, Volume 12, Number 1, 2017.

5. Clustering 1: K-means, K-medoids, Ryan Tibshirani, Data Mining: 36-462/36-662, January 24 2013, Optional reading: ISL 10.3, ESL 14.31

6. Selection of K in K-means clustering, D T Pham , S S Dimov, and C D Nguyen Manufacturing Engineering Centre, Cardiff University, Cardiff, UK The manuscript was received on 26 May 2004 and was accepted after revision for publication on 27 September 2004. DOI: 10.1243/095440605X8298.

7. Oyelade, “Application of k-Means Clustering algorithm for prediction of Students’ Academic Performance” (IJCSSS) International Journal of Computer Science and Information Security, Vol. 7, o. 1, 2010.

8. [8] Sudhir Singh and Nasib Singh Gill, “Analysis And Study Of K-Means Clustering Algorithm” International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 7, July – 2013.

9. [9] Huda Hamdan Ali , Lubna Emad Kadium, “K-Means Clustering Algorithm Applications in Data Mining and Pattern Recognition” International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

10. Tapas Kanungo “An Efficient k-Means Clustering Algorithm: Analysis and Implementation” IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 24, NO. 7, JULY 2002.

11. Chunhui Yuan and Haitao Yang, “Research on K-Value Selection Method of K-Means Clustering Algorithm” Received: 21 May 2019; Accepted: 15 June 2019; Published: 18 June 2019.

12. K-medoids Clustering Algorithm.

13. James Newling “K-Medoids for K-Means Seeding.

14. Raghuvira Pratap A “An Efficient Density based Improved K-Medoids Clustering algorithm” (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011

15. Preeti Arora , Dr. Deepali, Shipra Varshney, “Analysis of K-Means and K-Medoids Algorithm For Big Data” International Conference on Information Security & Privacy (ICISP2015), 11-12 December 2015, Nagpur, INDIA

AUTHORS PROFILE

S. Suganya, M.sc., B.Ed., PhD Research scholar, Department of computer science, Alagappa University, Karaikudi, Tamilnadu. India. Her research area are Big Data Analysis and Data Mining.

Dr. T. Meyyappan, M.Sc., M.Tech., M.Phil., Ph.D. currently, Professor, Department of Computer Science, Alagappa University, Karaikudi, Tamilnadu. India. He has organized conferences, workshops at national and international levels. He has published 90 numbers of research papers in National and International journals and conferences. He has developed Software packages for Examination, Admission Processing and official Website of Alagappa University. As a Co-Investigator, he has completed Rs.1 crore project on smart and secure environment funded by NIRO, New Delhi. As principal Investigator, he has completed Rs. 4 lakhs project on Privacy Preserving Data Mining funded by U.G.C. New Delhi. He has been honoured with Best Citizens of India Award 2012. His research areas include Operational Research, Digital Image Processing, Fault Tolerant computing, Network security and Data Mining.

Dr. S. Santhosh Kumar

M.Sc.,M.Phil.,M.Tech.,Ph.D. currently, Assistant Professor, Department of Computer Science, Department Coordinator, NET/SET Training classes, Department Co-ordinator, Mentor Mentee program. Department, Co-ordinator Students Blog & Forum, Department Coordinator, Go Green Committee, Event Co-coordinator, ALUTES Event, Alagappa University, Karaikudi, Tamilnadu, India.