Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 10, Issue, 08, pp.72224-72226, August, 2018

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

ISSN: 0975-833X

RESEARCH ARTICLE

PRIVACY PRESERVATION IN WASTE WATER TREATMENT PLANT USING K-MEANS AND BIRCH ALGORITHM

***Prabhjeet Kaur and Dr. Rekha Bhatia**

Department of Computer Science, PURCITM Mohali, Punjab, India

ARTICLE INFO

Article History:
Received 10th May, 2018
Received in revised form
24th June, 2018
Accepted 05th July, 2018
Published online 30th August, 2018

Key Words:
Data mining, K-means Algorithm, BIRCH Algorithm.

ABSTRACT

K-MEANS is a partitioning clustering algorithm in data mining which is very useful technique to find the nearest clusters in data. K-MEANS is an unsupervised learning, which use for divide the data into K-clusters. BIRCH is a Balanced Iterative reduction and clustering Hierarchies' algorithm. It is a hierarchical based clustering in which is used to divide a large data in small clusters. To improve the quality to large dataset in which some values are not present we used a combination of K-MEANS and BIRCH algorithm to solve this problem. In this paper, we discussed the data set in which we do not get the exact problem in data set and how we solved it. To solve this type of problem we use represent a combination of K-MENAS and BIRCH algorithm.

Citation: ***Prabhjeet Kaur and Dr. Rekha Bhatia.*** **2018.** "Privacy preservation in waste water treatment plant using k-means and birch algorithm, *International Journal of Current Research*, 10, (08), 72224-72226.

INTRODUCTION

In Today's world, data mining is very popular. It is use to reduce the large data into small parts. Clustering is sub class of data mining which used to form clusters of large data. Sometimes this large data values do not relate with each other. Clustering used to handle all types of dissimilar data. To handle various types of data different algorithms are used. To solve different type of problems like distortion in data, accuracy, missing value in data, large data in which we do not find out what type of data is used. To solve all these problems partitioning and hierarchical clustering used which solve these problems and provide privacy of data that no one misuse this data. In this paper, we solve the problem in water treatment data set, which is very large and cannot examine what type of problem is and it solved. In partitioning based clustering, we form the clusters of small data set and we only examine the same type of cluster problem .In hierarchical clustering we solve the large data which form the small clusters of big data as compare to K-means. However, some difficulties are not solving by birch alone like distortion in large data at what point data value increase. To solve the problems of distortion and accuracy in large data and provide the privacy of the data that it cannot be by other to harm the other things.

Corresponding author:* **Prabhjeet Kaur
Department of Computer Science, PURCITM Mohali, Punjab, India
DOI: https://doi.org/10.24941/ijcr.32005.08.2018

Because water treatment data is very large but when it openly solve in environment it causes many issues so data privacy is also useful in this paper. In section I we only introduce how we work. In section II, we discuss about how clustering algorithms used. In section III, we tell the methodology. In section IV we discuss how it done their work ands conclusions of the research. In last section V, we discuss the future scope of technique, which we used in our research.

Summary of related work: Clustering algorithms used to get a better result for unknown data or a large data, which is used. Kaur *et al*. (2014) concluded that k-means data set is better in find out the noise reduction as compare to K-mediods. BIRCH algorithm is better for large data set as compare to CURE and CLARANS. BIRCH algorithm is also better for when data values missed. Hendrik Fichtenberger *et al*. (2013) found that BIRCH algorithm is better for time. Sachin Shinde *et al*. examined that K-MEANS algorithm also improve the time complexity to find same type of research paper in limiting time. Tian Zhang *et al* analyzed that BIRCH algorithm is deal with real problems. For building a pixel classification tool is solve with BIRCH algorithm. Raghu Ramakrishnam *et al*. (1997) analyzed that BIRCH algorithm is used which deal with the multi-dimensional metric data points to produce the best quality clustering with the available resources. Raj bala (Raghu Ramakrishnam Zhang, 1997) analyzed that K-MEANS is efficient and mining data easily as compare to density based algorithm.

B.S. Sangeethas proposed For searching a research, paper easily in less time with K-MEANS algorithm.

MATERIALS AND METODS

We build a combination of two clustering algorithms to improve the accuracy and distortion in a data. Some problems are solved by K-MEANS algorithm and some other problems solved by BIRCH algorithm. BIRCH algorithm used to handle large data sets. Firstly, perform the K-MEANS algorithm for finding the random clusters centers of the data. Calculate the distance between clusters. Recalculate the new cluster centre. Then start using BIRCH algorithm on K-MEANS algorithm to find new clusters number of cluster (N). Load the clusters into memory then condense the data. Then use the previous algorithm for global clustering and refine the clusters.

To implement combined Algorithm, following steps formed:

- We found the dataset of water treatment and first deal with missing value in it.
- For the solution of missing values in data, then observed the average values of the whole table and form a new table and for this, we used vector quantization method to change the dataset.
- Apply K-MEAN ALGORITHM to find the clusters because this algorithm is scalable. Number of clusters to be found from original dataset and transformed dataset was taken same as number of cluster for quantization.
- Apply the Hierarchical-clustering algorithm BIRCH on the original and transform quantized dataset. CF(clustering feature) TREE formed form both dataset and found the distance between clusters with the use of F-measure to found that how closely clusters formed as compare to original data-set. Global clustering done on the data set. Then cluster refinement performed in which formed which clusters are actually useful for data.
- We compared the distortion formed due to change from original to average data with original segments and quantized data and observed the F-MEASURE between segments and quantized data.

RESULTS

In this, we observed the results between the distortion and segments of the data and between distortion and quantized data. In figure, one shows graph between distortion and segment size (L).

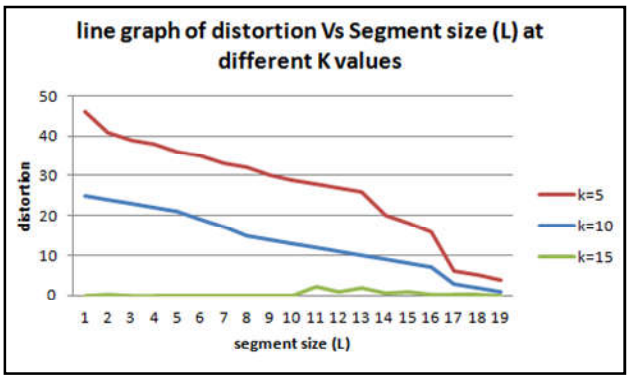

Fig. 1. Line graph of distortion Vs Segment Size at different L

It shows that distortion value decreases with increase in L. and it is fact more the value of L more the attribute of quantization affected and distortion increase but in our study distortion decreases with increases the segment size (L). Distortion decreses because in when the number of clusters decresed information about the data also decresed then privacy formed of the data with decrease in distortion. In figure, two showed the graph between distortion and quantized data (K). In this distortion decreases with increases the number of clusters. In this distortion, leads to loss of information of the original segment so it decreased with various segment size data on quantized data.

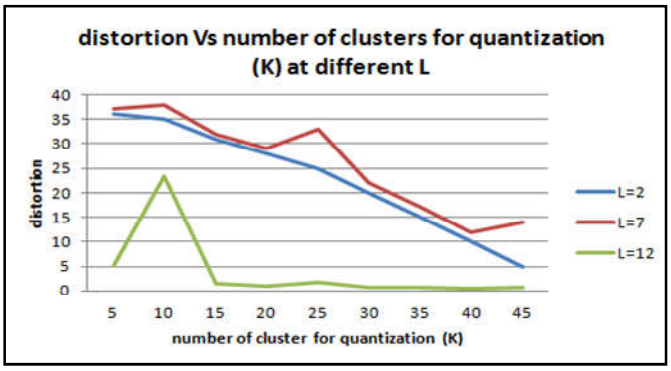

Fig. 2. Distortion Vs quantization (K) at different L

In figure 3 showed the minimum difference between clusters i.e. F-MEASURE and segment size (L).In starting value the where K value is low information is loss is also more and give privacy at point 30 where the distortion is not highest and after this value of F-MEASURE starts decreasing which gives a low effecting privacy to the data.

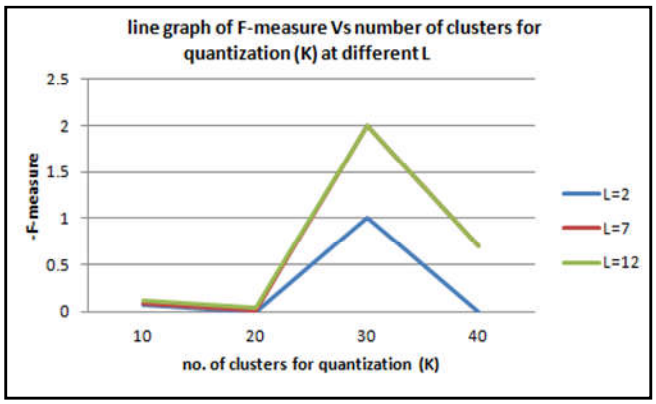

Fig. 3. Line graph between F-MEASURE Vs K at different L

In Figure 4 we formed a graph between F-MEASURE and segment size L at various K values where value of F-MEASURE is increased upto 1 at segment value 3 at various k-values. We found that for water treatment dataset for privacy preservation its giving best result at value 1 when segment size is three.

Future scope: As future work, new and effective work done on large data, which is very difficult to handle. Combination of both K-MEAN and BIRCH algorithm gives a better result on complex data. Easily handle the large data with missing values. It will give privacy preservation for large data and form clusters easily. It will also decrease the distortion between the data.

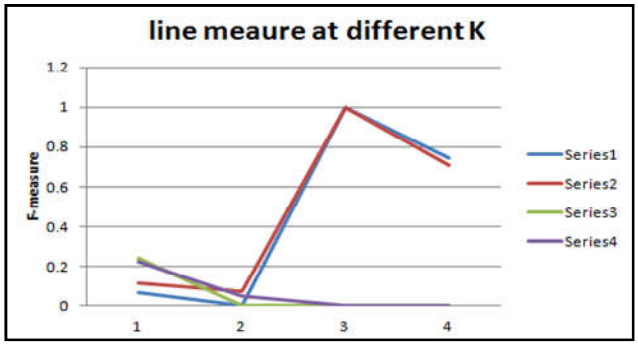

Fig. 4. Line graph between F-MEASURE and segment (L) at different L

REFERENCES

Fichtenberger H., Gillé M., Schmidt M., Schwiegelshohn C., Sohler C. 2013. "BICO: BIRCH Meets Coresets for *k*-Means Clustering. In: Bodlaender H.L., Italiano G.F. (eds) Algorithms". ESA 2013. Lecture Notes in Computer Science, vol 8125. Springer, Berlin, Heidelberg

Kaur, S., Chaudhary S., Bishnoi N. 2015. "A Survey: Clustering Algorithms in Data Mining." *International Journal of Computer Applications,* (0975- 8887), 12-14

Kedar B. Sawant Shree Rayeshwar 2015. "*International journals of advances in management and enginerring sciences,* volume 4, issue 6(1) RJanuary, PP 22-27 ISSN 2349-4395(Print) & ISSN2349-4409(Online)

Raghu Ramakrishnam Zhang, 1997. "BIRCH: A New (6) Data Clustering Algorithm and Its Applications" Data Mining and Knowledge Discovery, Volume 1, pp 141–182

Raj bala. "A Comparative Analysis of Clustering Algorithms" Research Scholar (M.Tech) Amit University Haryana, India

Sachin Shinde *et al*. "Improved K-means Algorithm for Searching Research Papers". International Journal of Computer Science & Communication Networks,Vol 4(6),197-202

Shraddha Shukla and Naganna 2014. "A Review ON K-means DATA Clustering APPROACH" *International Journal of Information & Computation Technology*. ISSN 0974-2239 Volume 4, Number 17 (2014), pp. 1847-1860

UCI Repository of machine learning databases, University of California, Irvine. http://archive.ics.uci.edu/ml/

Wikipedia. Data mining. http://en.wikipedia.org/ wiki/ Data_mining

Zhang, T., Ramakrishnan, R. and Livny, M. 1997. "BIRCH: A New Data Clustering Algorithm and Its Applications" Data Mining and Knowledge Discovery, Volume 1, Issue 2, pp 141–182

****