Incremental K-means Clustering Algorithms: A Review

Amit Yadav
Department of Computer Science Engineering

Prof. Gambhir Singh
H.R.Institute of Engineering and Technology, Ghaziabad

Abstract: Clustering is the process of grouping the object based on their attributes and features such that the data objects that are similar or closer to each other are put in the same cluster. K-means is most popular clustering algorithm which partitioned the data but When the amount of data to be clustered is large and/or when data becomes available incrementally then incremental clustering procedures are used. This paper reviews how to handle incremental data efficiently and also remove the drawback of K-means Clustering Algorithm by using different approaches. In this review paper, analysis of different methodology of Incremental Clustering algorithm is done which improve the quality and accuracy of clusters.

Keywords : Clustering, Data object, K-means clustering, Incremental Clustering

I. INTRODUCTION

Data Mining is the Process of analyzing data from different perspective and summarizing it into useful meanings which extracts some useful information, Patterns and Relationships from data sources such as DB, text and the web. Data mining finds valuable information hidden in large volumes of data. Variety of tools and algorithms are used for the mining of the data. Data Clustering is very valuable field of data mining to group the similar data into cluster and dissimilar data into different clusters and it is form of unsupervised learning in which no class labels are provided. K-means is most popular clustering algorithm which partitioned the data but there are many limitations of this algorithm such as number of clusters needs to be defined beforehand. A major problem of modern data clustering algorithm is that continuous dumping of new data sets into an existing bulky DB and it's not viable to perform data clustering from scrape every time new data instances get added up in database. It requires the design of new Clustering algorithms which handle this problem is to integrate clustering algorithm that functions incrementally and for that incremental K-means clustering is used.

II. K-MEANS CLUSTERING

K-Means is partition based clustering technique employing Euclidean distance. The commonly used distance measure is the *Euclidean metric* which defines the distance between two points P= (x1(P), x2(P),…) and Q = (x1(Q), x2(Q),…) is given by :

$$d(\mathrm{P},\mathrm{Q}) = ((\mathrm{x}1(\mathrm{P}) - \mathrm{x}1(\mathrm{Q})^2 + (\mathrm{x}2(\mathrm{P}) - \mathrm{x}2(\mathrm{Q})^2 + ...)^{1/2}$$

$$= (\sum_{j=1}^{p} (\mathrm{xj}(\mathrm{P}) - \mathrm{xj}(\mathrm{Q})^2))^{1/2}$$

K-means is very popular because of its simplicity and speed of classifying massive data very efficiently. However, the output of K-Means algorithm mainly depends upon the selection of initial cluster centers because the initial cluster centers are chosen randomly[5]. The other limitation is to input the required number of clusters which requires some sort of intuitive knowledge about appropriate value of K which sometimes difficult to predict as it requires domain knowledge.

III. INCREMENTAL CLUSTERING

Today most of the databases are very dynamic in nature because of fast growth of World Wide Web so new data sets are dynamically added into an existing DB and updated data is not perform clustering every time very easily. The dataset is dynamic, so it is impossible to collect all data objects before starting clustering. When new data comes, non-incremental clustering will have to re-cluster all the data, which certainly decreases efficiency and wastes

computing resource but incremental clustering group the new data and update new clusters to previous clustering results. The strategy optimizes clustering process and mainly adapts to those applications where time is a critical factor for usability. Incremental document clustering is one the most effective techniques to organize documents in an unsupervised manner for many Web applications[7]. This algorithm is used to handle incremental data in existing database very efficiently. Some shortcomings of basic K-means clustering algorithm can be overcome very easily by using Incremental clustering algorithm. Initial clustering and handling of incremental data points are two important steps of the incremental clustering Approaches[3].

IV. CLASSIFICATION OF INCREMENTAL K-MEANS CLUSTERING APPROACHES

The original K-means algorithm is computationally very expensive because each iteration computes the distances between data points and all the centroids so for that different incremental K-means clustering approaches are proposed.

Nidhi Gupta and R.L Ujjwal[1] proposed an algorithm which uses threshold T which denotes dissimilarity between data objects and give a value of Tth initially then choose an object randomly from the given datasets and it become the center of a cluster, and choose another object from the given datasets again and compute distance between the selected data object and the existing cluster center. If this distance is larger than Tth then form a new cluster and selected object will be the center of the cluster otherwise group the object into existing cluster and update its centroid. Proposed algorithm remove the disadvantage of K-means algorithm. In this algorithm we do not need to specify the value of K And produces clusters in less computation time so we can say proposed algorithm is better than the K-means algorithm.

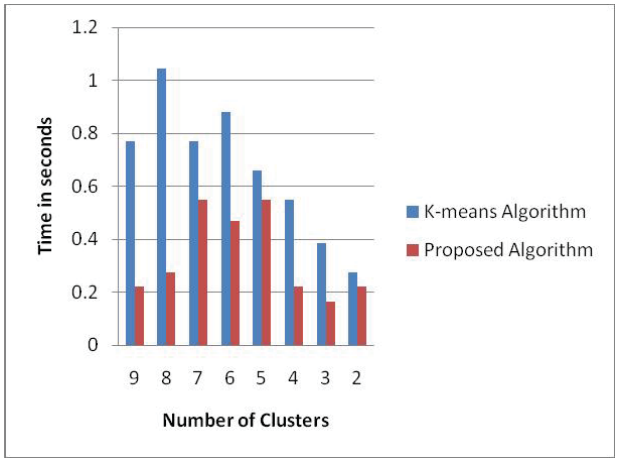

Fig.1. Variation of number of clusters w.r.t. time for K-means algorithm

A.M.Sowjanya and M.Shashi[2] proposed a cluster feature based incremental clustering approach for handling incremental data efficiently. Initial clustering and handling of incremental data points are two important steps of the incremental clustering approaches. For initial clustering, K-means clustering algorithm is applied and then for clustering the incremental data points, CFICA has been applied. Finally, by making use of the mean value, closest pair of clusters has been merged after processing the set of data points. After the database is updated, there are three possibilities for clustering the updated points:

1. Adding with the existing cluster
2. Formation of a new cluster
3. Possibility to merge the existing clusters when updated points are in between the existing two clusters. This incremental clustering approach is very efficient in terms of clustering accuracy.

Sanjay Chakraborty, N.K. Nagwani[3] proposed an algorithm which define and evaluate that the particular point of change in the database ("Threshold value" or % delta) upto which incremental K-means clustering performs much better than the existing K-means clustering. Nowadays databases are dynamic so incremental algorithm is used to handle incremental data. The proposed algorithm identifies the value of percentage of size of original DB x, which can be added to original DB. There might be two cases:

1. If original DB is change up to x% then use previous result.
2. If origional database is change up to more than x % then rerun the algorithm again.

Threshold value = (New data - old data)/old data *100 [3]

As a result, Incremented K-means clustering algorithm is applied on incremental data after collecting necessary information from the result DB so upcoming data is directly inserted into the existing DB without running the K-means algorithm again and again. This algorithm provide faster execution because the number of scans for the database will be decreased. So basically it provide better performance than K-means Clustering Algorithm.

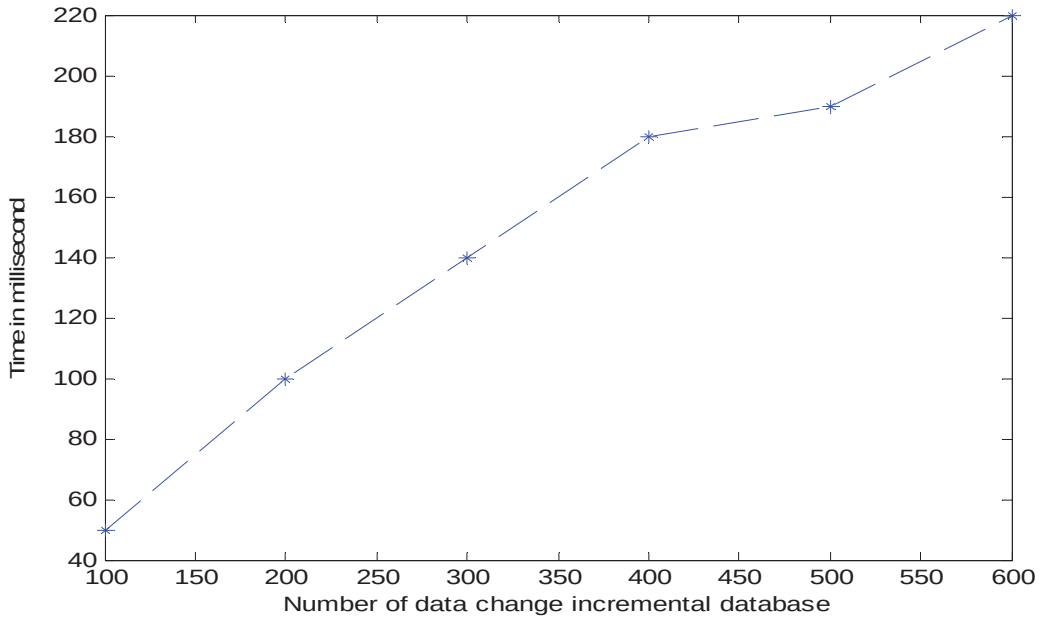

Fig.2.Graph for K-means algorithm[3]

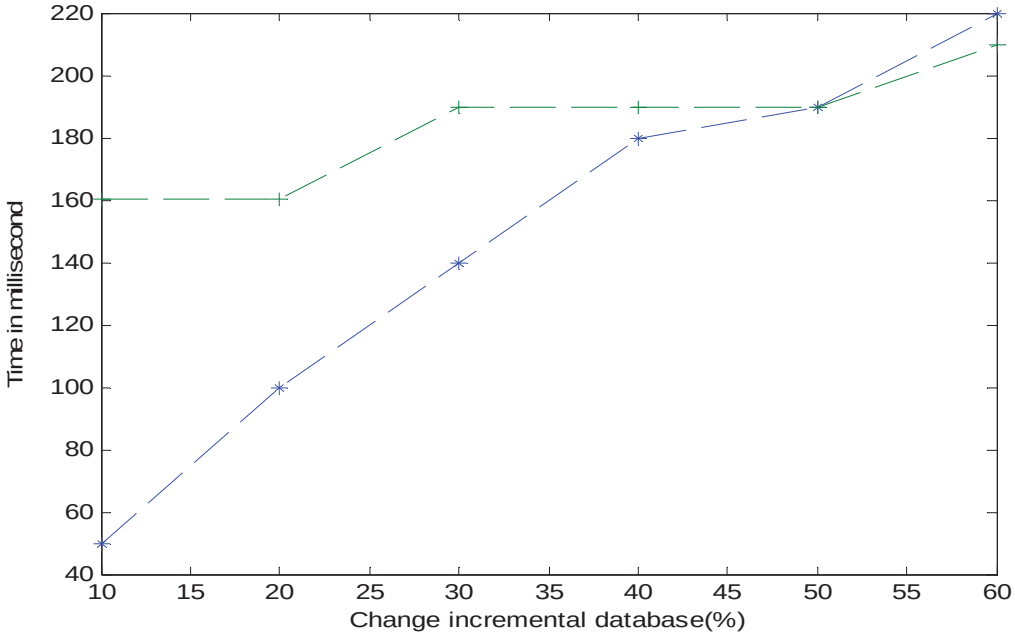

Fig.3.Graph for actual K-means Vs incremental K-means algorithm[3]

Xiaoping Qing and Shijue Zheng[4] proposed a method to compute initial cluster centers for K-means clustering and based on an efficient technique for estimating the modes of a distribution. K-means clustering algorithm mainly depends on initial cluster centers which are selected randomly, so the algorithm could not lead to the unique result[4]. A Proposed method avoid the empty clusters problem which re-assign all empty clusters to points farthest from their respective centers and also implemented new method on Gaussian Data and compare the proposed initialization method with the random choice of cluster centers which is more effective but it is limited to small dataset.

Anupama Chadha and Suresh Kumar[5] proposed an algorithm which does not require the number of clusters K as input. It simply remove the limitation of k-means clustering is to input the required number of clusters K. Initially Proposed algorithm create two clusters are by choosing two initial centroids which are farthest apart in the data set so that in the initial step itself we can create two clusters with the data members, which are the most dissimilar ones. If the accuracy of the clusters is to be increased then the selection of the initial centroids should be good. In order to improve the selection of the initial centroids we increase the number of trials which in tum affects the computation time and it depends mainly on size and number of dimensions in the dataset. Experimental result shows that the quality and accuracy of clusters are not compromised but it's limited to numeric dataset.

Above all approaches improve quality and accuracy of the clusters and we get the effective performance than the K-means clustering algorithm so we can say incremental K-means clustering approaches are work better than the K-means and it also reduces time complexity and review of how they are useful to handle updated data efficiently.

Table 1 : Analysis of Incremental K-means Approaches

PAPER	REVIEW
An Efficient Incremental Clustering Algorithm[1]	It remove the drawback of K-means algorithm so we don't need to specify the value of K which takes less time than K-means algorithm & it's better than K-means
CFICA For Numerical Data[2]	CFICA is more efficient approach for clustering incremental DB. It has two modules: initial(K-means) & incremental clustering(CFICA). It's more efficient in terms of clustering accuracy.
Performance Evaluation of Incremental K-means Clustering Algo[3]	It evaluates the particular point of change in the DB upto which incremental K-means clustering performs much better & also measure and compare the performance with the existing K-means clustering. It gives better performance than existing k-means clustering.
A new method for initialising the K-means clustering algorithm[4]	It makes improvement with the clustering problem which avoids the empty cluster problems and also implement the new method on Gaussian data and compare the proposed initialization method with random choices of cluster centers which is more efficient and effective than K-means.
An Improved K-Means Clustering: A StepForward for Removal of Dependency on K[5]	It removes major limitation of basic K-means is to require K as input. It improve the time complexity, quality of the clusters and accuracy of the clusters but it's limited to numeric dataset.

V. Conclusion

Clustering is a challenging task in periodically incremental data and bulk of updates so perform data clustering every time for updated data in DB it is simply waste of time. The purpose of Incremental K-means clustering is to handle incremental data in DB very efficiently. It removes the limitation of K-means clustering algorithm and gives accurate result in less time so we can say it's very efficient than standard K-means clustering algorithm and quality of cluster is also improved. From Our analysis of incremental K-means approaches, we conclude that it's better than K-means clustering algorithm.

REFERENCES

[1] Nidhi Gupta, R.L Ujjwal."An Efficient Incremental Clustering Algorithm" in World Of Computer Science and Information Technolgy Journal (WCSIT) ISSN: 2221-0741 Vol. 3, No. 5, 97-99,2013.

[2] A.M.Sowjanya and M.Shashi." Cluster Feature-Based Incremental Clustering Approach(CFICA) For Numerical Data" in IJCSNS International Journal of Computer Science and Network Security, VOL.10 No.9, September 2010.

[3] Sanjay Chakraborty , N.K. Nagwani ." Performance Evaluation of Incremental K-means Clustering Algorithm " .IFRSA International Journal of Data Warehousing & Mining ,2011.

[4] Xiaoping Qing and Shijue Zheng." A new method for initialising the K-means clustering algorithm" in 978-0-7695-3888-4/09 $25.00 © 2009 IEEE.

[5] Anupama Chadha, Suresh Kumar. "An Improved K-Means Clustering Algorithm: A Step Forward for Removal of Dependency on K" in 978-1-4799-2995-5/14/$31.00©2014 IEEE.

[6] Bryant Aaron, Dan E. Tamir, Naphtali D. Rishe, and Abraham Kandel." Dynamic Incremental K-means Clustering" in 978-1-4799-3010-4/14 $31.00 © 2014 IEEE

[7] Yongli Liu, Qianqian Guo, Lishen Yang, Yingying Li," Research on Incremental Clustering", in 978-1-4577-1415-3/12/$26.00 ©2012 IEEE.