A Modified Approach for Incremental K-Means Clustering Algorithm

Nidhi S. Shah
M.E (Information Technology)
L.D College Of Engineering, Ahmedabad

Abstract **- Clustering is 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 and it is form of unsupervised learning so 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, number of iterations are unknown etc. Incremental data can be handled very efficiently by incremental clustering algorithm. It tries to generate new clusters at the end for each updated data, which cannot be merged with existing cluster so it increases computation time and also accuracy of clusters is reduced. This report presents most efficient modified approach for incremental K-means Clustering algorithm where clusters are generated dynamically without rerun the K-means on whole dataset which reduces computation time and gives more accurate result. For that, Initial Clustering is performed on static Database by using K-means clustering. Then for upcoming points, major distance between centroid to farthest point and upcoming point which is used to define the upcoming point is in existing cluster or not based on some criteria which is defined in proposed scheme. If it's not in exiting cluster then recompute K-means for the outside points only.**

Keywords **- Clustering, Data object, Centroid, K-means clustering, Incremental Clustering**

I. INTRODUCTION

Data Mining is the Process of analyzing data from different perspective and summarizing it into useful meanings. It extracts some useful information , Patterns and Relationships from data sources such as databases, text and the web which 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 computational statistics and data mining. Clustering is a data mining technique to group the similar data into cluster and dissimilar data into different clusters[10]. 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. A Solution to handle this problem is to integrate clustering algorithm that functions incrementally[3]. Incremental Clustering algorithms permit a single or few passes over the whole dataset to put updated item into the cluster[3]. With respect to size of the set of objects, algorithms and number of attributes, incremental clustering algorithm are of scalable nature[3]. In this work modified approach for incremental K-means clustering algorithm is used where clusters can be generated dynamically without rerun the K-means on whole dataset and it reduces computation time and gives more accurate result.

II.BASIC IDEA OF K-MEANS AND INCREMENTAL K-MEANS CLUSTERING ALGORITHM

K-means clustering is a popular clustering algorithm based on the partition of data. The main merits are its simplicity, memory efficient and speed which allows it to run on large datasets. It is an unsupervised learning technique that solve the very known clustering problem But there are some limitations in it such as number of cluster needs to be defined beforehand, Number of iterations are unknown, very sensitive to outliers and for selection of initial seed etc. Its demerit is that it does not yield the same result with each run because the resulting clusters depend on the initial seed assignments which is assigned randomly.

Incremental Clustering 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]. Initial clustering can be performed by basic K-means for static database then for upcoming data incremental approaches can be used such as Threshold which is the lowest possible input value of similarity required to join two objects in one cluster. The existing K-means clustering algorithm which is developed in Java is applied on the original dataset and the result is stored in a result database using mysql. Then after new data is inserted into that existing database which is called as incremental database. The incremental K-means clustering algorithm is applied on data after collecting necessary information from the result database. This way incremental data is directly inserted into the existing database without running the K-means algorithm again and again. Finally the results of these two are compared and also evaluate the performance as well as its correct threshold value.

So we can observe that when K-means clustering is used for incremental data then it can be rerun for whole dataset always but when we used Incremental K-means then it can't be rerun for whole dataset but only rerun for outside points which are not put in existing clusters. Therefore reducing computation time and give better accuracy proposed scheme is published.

III.PROPOSED ALGORITHM

In this paper, an incremental K-means clustering approach is used. A Modified approach for Incremental K-means clustering algorithm and it is used where Initial clustering is performed on static DB by using k-means clustering Algorithm . Then for upcoming points, major distance between farthest point and centroid (D_{FC}) and also calculate the distance between upcoming point and centroid of given cluster (D_{UC}) which is used to define the upcoming point is in existing cluster or not. If it follows below condition then put this point into existing cluster

$$\mathbf{D_{UC} <= D_{FC} + (D_{FC} * (1/3))}$$

And if it not follow above condition then it create the list of that points which are not accommodate in existing clusters upto **|D| / 4** points and then after it recompute K-means for that outside points. For making a list of not accommodate points in existing cluster is very useful to reduce computation time as well as improve the accuracy of clusters which can be measure by confusion matrix.

Proposed Algorithm

Input: Dataset, Number of clusters(K), upcoming data points
Output: Clusters of points

Working:

Step1: Run K-means clustering Algorithm with given number of clusters(K) for given dataset.

Step2: For upcoming datapoints, Calculate distance between farthest point and centroid suppose it is D_{FC} And calculate distance between upcoming point and centroid of given clusters suppose it is D_{UC}

Step3: If ($D_{UC} <= D_{FC} + (D_{FC} * (1/3))$)
Then put this point into existing cluster
Else
Ignore to merge till |D| / 4 points

Step4: If outside points reaches |D| / 4 count, Then Run K-means algorithm on outside points consider centroids are existing centroid and choose one more centroid point from remaining ones

Step5: Repeat from step 2 for more upcoming points

IV.EXPERIMENTAL RESULTS

An IRIS dataset and Mushroom datasets are taken which are available at the UCI machine learning Repository. IRIS flower dataset consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor) so it consists of total 150 instances. Four features the length and the width of the sepals and petals, in centimeters were measured from each sample. Mushroom dataset consists of 8124 instances with 22 attributes, which are categorical. Here existing algorithm and Modified clustering Algorithm is applied on Iris dataset as well as mushroom dataset and we will compare clustering accuracy and time of both original and modified. After testing following results are generated and The results of the experiments are tabulated in Table 1 and Table 2:

Table 1: Result of Clustering Time

Datasets / Approaches	Time (ms)	
	IRIS	**Mushroom**
Incremental K-Means	12	2410
Proposed Algorithm	6	1863

When We perform analysis of both algorithm based on computation time by using graphical method then we can see in above Figure 1 computation time of proposed algorithm is reduced for small dataset as well as large dataset.

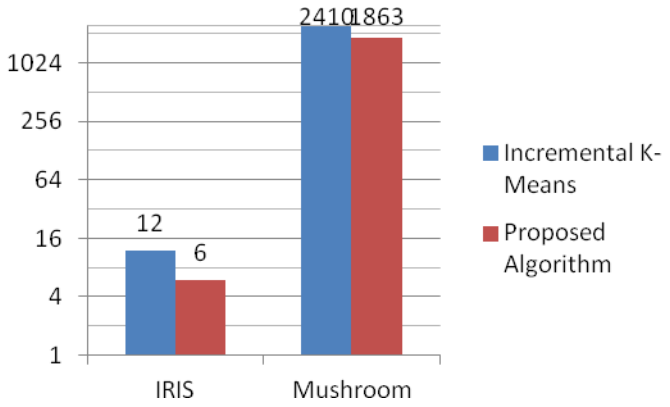

Figure 1: Comparison of Clustering Time

Result of clustering accuracy is given in below Table**:**

Table 2: Result of Clustering Accuracy

Datasets / Approaches	Accuracy (%)	
	IRIS	**Mushroom**
Incremental K-Means	90.52	78.78
Proposed Algorithm	93.23	86.23

When We perform analysis of both algorithm based on accuracy by using graphical method then we can see in above Figure 2 proposed algorithm has higher accuracy than original Incremental K-means algorithm.

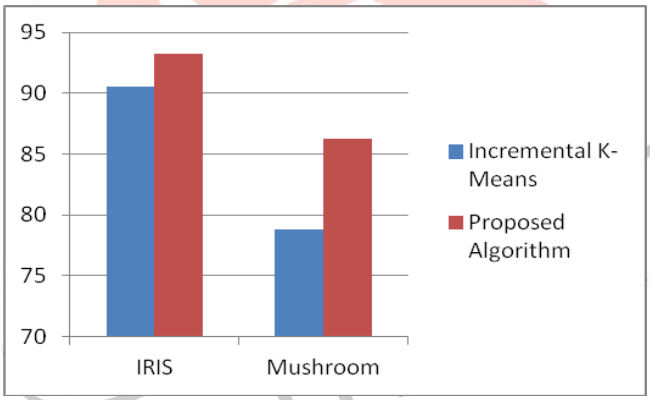

Figure 1: Comparison of Clustering Accuracy

V. Conclusion and Future Work

Clustering is a challenging task for incremental data and bulk of updates and for that Incremental Clustering algorithm is very innovative approach to handle incremental data. It define and evaluate threshold which is used to define upcoming point is in existing cluster or not. If not then it generate new cluster, which is directly affect to the accuracy of clusters as well as computation time. With the help of modified approach for Incremental K-means algorithm, clusters can be generated dynamically without rerun K-means on whole dataset for outside points only and this way algorithm is more efficient in terms of accuracy and computation time.

In future, we will plan to apply Proposed K-means Algorithm in Parallel Incremental K-means Algorithm.

VI. 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.
[8] Kehar Singh , Dimple Malik and Naveen Sharma."Evolving limitations in K-means algorithm in data mining and their removal"in IJCEM International Journal of Computational Engineering & Management ISSN: 2230-7893Vol. 12, April 2011.
[9] Juntao Wang, Xiaolong Su." An improved K-Means clustering algorithm" in 978-1-61284-486-2/111$26.00 ©2011 IEEE.
[10] Sowjanya, . "A Cluster Feature-Based Incremental Clustering Approach to Mixed Data", Journal of Computer Science, 2011.