Analytical Comparison of Some Traditional Partitioning based and Incremental Partitioning based Clustering Methods

Rimi Gupta
Department of Computer Engineering
SVIT, Vasad
Gujarat, India.

Jayna Shah
Assistant Professror
Department of Computer Engineering
SVIT, Vasad
Gujarat, India.

Neha Soni
Assistant Professror
Department of Computer Engineering
SVIT, Vasad
Gujarat, India.

ABSTRACT

Data clustering is a highly valuable field of computational statistics and data mining. Data clustering can be considered as the most important unsupervised learning technique as it deals with finding a structure in a collection of unlabeled data. A Clustering is division of data into similar objects. A major difficulty in the design of data clustering algorithms is that, in majority of applications, new data are dynamically appended into an existing database and it is not feasible to perform data clustering from scratch every time new data instances get added up in the database. The development of clustering algorithms which handle the incremental updating of data points is known as an Incremental clustering. In this paper authors have reviewed Partition based clustering methods mainly, K-means & DBSCAN and provided a detailed comparison of Traditional clustering and Incremental clustering method for both.

1.INTRODUCTION

Data mining, a one of the promising technology, is up to some extent a nontraditional data driven method to discover novel, useful, hidden knowledge from massive datasets.

Data clustering is the most famous and necessary concepts in data mining. Clustering plays an important role in data mining and is applied widely in fields of pattern recognition, computer vision, and fuzzy control. It is also known as unsupervised learning process, as there is no a-prior knowledge about the data. Clustering is to group data points into several clusters and makes the intra-cluster similarity maximum and the inter-cluster similarity minimum [1].

Incremental clustering is the extended version which is suitable for the databases in which data are frequently added. With the development of information technology, especially with the Web, data and environment are varying from minute to minute and more and more space is needed for storing data in memory. For clustering a new appended data required to rescan a datasets again and again. The incremental clustering was proposed with the advantage of limited space requirement since the entire dataset is not necessary to store in the main memory[7]and clustering a newly data without rescanning a dataset so it can save a lots of time also.

The rest of this paper is organized as follows. Section 2 discuss on Traditional and Incremental clustering. Section 3 talks about Traditional K-means clustering and Incremental K-means clustering. Section 4 talks about Traditional DBSCAN clustering and Incremental DBSCAN clustering. Section 5 concludes with a summary of discussed clustering techniques and future scope. Section 6 describes the references.

2.BACKGROUND CONCEPTS

A. Traditional Clustering

Several data mining tasks have been identified and one of them is clustering. Clustering techniques have been applied to a wide variety of research problems such as in biology, marketing, economics and others. Clustering is similar to classification in that data are grouped. However, unlike classification, the groups are not predefined. Instead, the grouping is accomplished by finding similarities between data according to characteristics found in the actual data.

A clustering algorithm partitions a data set into several groups such that the similarity within a group is larger than among groups. Moreover, most of the data collected in many problems seem to have some inherent properties that lend themselves to natural groupings [2].

Clustering algorithms are used extensively not only to organize and categorize data, but are also useful for data compression and model construction.

B. Incremental Clustering

Incremental clustering means applying Clustering Algorithms on incremental Database. Data warehouses get updated periodically due to large set of data sources. It is desirable to perform these updates incrementally than in the batch mode. The objective of incremental clustering algorithms is to

minimize the scanning and calculation effort required to reform the clusters with newly added records.

The term incremental means "% of δ change in the original database" i.e. insertion of some new data items into the already existing clusters. Such as, [5].

% δ change in DB
= ((New data – Old data) × 100) / Old data)

Prime factors that cause to apply Incremental Clustering.

1) Database always gets modified.
2) Traditional Clustering requires lots of time for rescanning the whole database while some new data are added.
3) Traditional Clustering techniques require storing the whole large data set in main memory. So memory space requirement is too high.

Clustering algorithms are mainly divided into two parts: Partitioning and Hierarchical Based Clustering[9].

In this paper review of only Partitioning Based Methods: K-Means and DBSCAN is analyzed.

3. PARTITION BASED K-MEANS CLUSTERING METHOD

The Partitioning Methods means to creates K partitions of the Datasets with N data objects, each partition represent a

Cluster, where k<= N. The partitioning algorithm is required to represent a clusters by the gravity of the centre is known as k-means algorithms.

A. Traditional K-Means Clustering Method

K-means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. The algorithm called K-means where K is the no. of required clusters. Algorithm start with initial set of means and classify clusters based on their distance to the centers. Next compute the clusters mean again and reclassify the clusters. Repeat this process until cluster means don't change much between successive steps. [2].

It is applicable to large data sets but in that it required to accommodate the entire data set into the main memory. So that K-means clustering algorithm may take an amount of time.

B. Incremental K-Means Clustering Method

To overcome the problems of Traditional K-means clustering the Incremental clustering is designed using the cluster's metadata captured from the K-Means results [5]. In incremental clustering approach, the K-means clustering algorithm is applied to a dynamic database where the data may be frequently updated. This approach computes the new cluster centers from the means of the existing clusters and newly added data instead of applying the K-means algorithm on whole dataset again.

The large dataset is stored in secondary memory and during incremental clustering data items are transferred to the main memory one at a time.

The concept is more clearly discussed with the examples given below [6].

Example: Suppose there is a set of data objects, such as, 15, 11, 8, 14, 3, 1, 7, 5. Assume that the points 15, 5, 1 are three cluster centers. Add two new data item 17 and 9.

Two approach Traditional and Incremental Clustering to solve this problem.

TABLE 1: COMPARISION BETWEEN TRADITIONAL K-MEANS AND INCREMENTAL K-MEANS

Traditional Clustering	**Incremental Clustering**
Case 1: Initial Clustering Input : 15, 11, 8, 14, 3, 1, 7, 5 Output : Apply Manhattan Distance Metric Cluster 1={15, 11, 14} Mean = 13.3 Cluster 2={7, 8, 5} Mean = 6.7 Cluster 3={3, 1} Mean = 2	Case 1: Initial Clustering Input : 15, 11, 8, 14, 3, 1, 7, 5 Output : Apply Manhattan Distance Metric Cluster 1={15, 11, 14} Mean = 13.3 Cluster 1={7, 8, 5} Mean = 6.7 Cluster 1={3, 1} Mean = 2

Case 2: Add new data 17 and 9 Rescan whole dataset again and apply again K-means algorithm. Input : 15, 11, 8, 14, 3, 1, 7, 5,17,9 Output : Apply Manhattan Distance Metric Cluster 1={15, 11, 14, 17} Mean = 14.25 Cluster 2={7, 8, 5, 9} Mean = 7.25 Cluster 3={3, 1} Mean = 2	Case 2: Add new data 17 and 9 Insert new data directly after comparing with the means of existing clusters using Manhattan Distance Metric. Input : Existing Cluster means 13.3, 6.7, 2 and new data 17, 9. Output : Apply Manhattan Distance Metric. Cluster 1={15, 11, 14, 17} Mean = 14.25 Cluster 2={7, 8, 5, 9} Mean = 7.25 Cluster 3={3, 1} Mean = 2
Traditional Clustering and Incremental Clustering both Results are same in any Case and Initial Build in time is same in any approach.	
Traditional Clustering takes more time for adding a data into existing clusters.	Incremental Clustering takes less time than Traditional Clustering.

4. PARTITION BASED DBSCAN CLUSTERING METHOD

A. Traditional DBSCAN Clustering Method

Most popular clustering algorithm is Density Based Spatial Clustering of Applications with Noise (DBSCAN) which has the ability to produce arbitrary shape, size of clusters. Clusters are identified by looking at the density of points. Regions with a high density of points depict the existence of clusters whereas regions with a low density of points indicate clusters of noise or clusters of outliers.

DBSCAN [3] grows clusters according to a density based connectivity analysis. Clusters are a maximal set of density-connected points. The key idea of density-based clustering is that for each object of a cluster the neighborhood of a given radius (eps) has to contain at least a minimum number of objects (Minpts) [8]. DBSCAN starts clustering with an arbitrary staring point that has not been visited. This point's e-neighborhood is retrieved, and if it contains sufficiently many points, a cluster is started. Otherwise, the point is labeled as noise.

B. Incremental DBSCAN Clustering Method

The DBSCAN approach is not suitable for a large database which is frequently updated. In that case, the incremental clustering approach is much better. In a data warehouse, the databases may have frequent updates and thus may be dynamic. Incremental DBSCAN has speed up factor over DBSCAN even for large numbers of daily updates in a data warehouse.

The incremental approach of DBSCAN first forms clusters using Traditional DBSCAN algorithm from the initial objects in dataset and for given radius (eps) and minimum number of points (Minpts). Now, when new data is inserted into the existing database, there is a need to update existing clusters using Incremental DBSCAN algorithm.

In Incremental DBSCAN, first compute the means between every core object of clusters and the new coming data and insert the new data into a particular cluster based on the minimum mean distance. The new data which are not inserted into any clusters, they are treated as noise or outliers.

The concept is more clearly discuss with the example given below [3].

Example: Suppose there is data points, such that 15, 22, 12, 82, 73, 10, 17, 48, 96,152, 8, 85. Now select any point as a core point and check condition of core point. Say 15 and 82 as core point. Add two new data 77 and 124. DBSCAN parameters: Eps= 10 Units and Minpts= 5

Two approach Traditional and Incremental Clustering to solve this problem.

TABLE 2: COMPARISION BETWEEN TRADITIONAL DBSCAN AND INCREMENTAL DBSCAN

Traditional Clustering	**Incremental Clustering**
Case 1: Initial Clustering Input: 15, 22, 12, 82, 73, 10, 17, 48, 92, 152, 8, 85 Output: Cluster 1={15, 22, 12, 10, 17, 8} Cluster 2={82, 73, 92, 85} Outlier=152.	Case 1: Initial Clustering Input: 15, 22, 12, 82, 73, 10, 17, 48, 92, 152, 8, 85 Output: Cluster 1={15, 22, 12, 10, 17, 8} Cluster 2={82, 73, 92, 85} Outlier=152.
Case 2: Add new data 77 and 125 Rescan whole dataset again and apply again DBSCAN algorithm. Input: 15, 22, 12, 82, 73, 10, 17, 48, 92, 152, 8, 85 Output: Cluster 1={15, 22, 12, 10, 17, 8} Cluster 2={82, 73, 92, 85, 77} Outlier=152, 125	Case 2: Add new data 17 and 9 Insert new data directly Clustered by calculating the minimum Mean between the newly added data and existing clusters core points. Apply Manhattan Distance Metric. Input: Existing Clusters core points and new data 77, 125. Output: Apply Manhattan Distance Metric. Cluster 1={15, 22, 12, 10, 17, 8} Cluster 2={82, 73, 92, 85, 77} Outlier=152, 125
Traditional Clustering and Incremental Clustering both Results are same in any Case and Initial Build in time is same in any approach.	
Traditional Clustering takes more time for adding a data into existing clusters.	Incremental Clustering takes less time than Traditional Clustering.

Table 3: Comparison Between Traditional Clustering and Incremental Clustering

Data Clustering	Database	Time Complexity for Large dataset	Space Complexity for Large dataset
Traditional	Static	High	High
Incremental	Dynamic	Less	Less

5. CONCLUSION

This paper Conclude that for mining an updated Data Warehousing environment, Incremental K-means and Incremental DBSCAN both are more efficient and applicable than Traditional K-means and Traditional DBSCAN algorithm. The Incremental Clustering is applied on Incremental data after collecting necessary information from the existing clustering and existing dataset. The new data can directly clustered in existing clustered without scanning the dataset and rerunning the algorithm again and again. So, It can conclude that Incremental clustering is More Efficient than Traditional Clustering. In this paper analyze the efficiency of the Incremental Clustering. In Future work, could be analyzing the other popular clustering techniques in Incremental Fashion.

6. REFRENCES

[1] F. Knoll "Survey of Clustering Data Mining Techniques" Pavel Berkhin Accrue Software, Inc. Pavel Berkhin , Accrue Software, 1045., San Jose, CA, 95129

[2] Manish Verma, Mauly Srivastava, Neha Chack, Atul kumar Diswar, Nidhi Gupta, " A Comparative study of various clustering algorithms in data mining", International Journal of Engineering Research and Applications, 2012.

[3] Prof.Sanjay Chakraborty, Prof. N.K. Nagwani, "Analysis and Study of Incremental DBSCAN clustering algorithm",

[4] International Journal of Enterprise Computing and Business Systems, July 2011.

[5] Martin Ester, Hans-peter Kriegel, Jorg Sander, Michael Wimmer, Xiaowei Xu, " Incremental Clustering for Mining in a Data Warehousing Environment", Proceedings of the 24th VLDB Conference New York, USA, 1998.

[6] Prof.Sanjay Chakraborty, Prof. N.K. Nagwani, "Analysis and study of Incremental K-Means clustering algorithm", A. Mantri et a. HPAGC 2011, CCIS 169,pp.338-441,2011.

[7] Prof.Sanjay Chakraborty, Prof. N.K. Nagwani, "Performance Evaluation of Incremental K-Means clustering algorithm",IFRSA International Journal of Data warehousing and Mining , 2011.

[8] C.C. Hsu, Y.P. Hung. Incremental clustering of mixed data based on distance hierarchy. Expert systems with Applications, 2008.

[9] Sauravjyoti Sarmah, Dhruba K. Bhattacharyya, " An Effective Technique for clustering Incremental Gene Expression data", IJCSI International Journal of Computer Science Issues, Vol 7, Issue 3, No 3, May 2010.

[10] Prof. Neha Soni , Prof. Amit Ganatra "Categorization of Several Clustering Algorithms from Different Perspective: A Review" International Journal of Advanced Research in Computer Science and Software Engineering.