KDBSCAN: A Hybrid Approach in Big Data

[1]Ekta Joshi, [2]Dr. D.A. Parikh
[1]ME Computer Engineering, [2]HOD Computer Engineering
[1]Computer Engineering
[1]L.D. College of Engineering, Ahmedabad, India

***Abstract :** The goal of the data mining process is to extract given information from a data set and transform it into a useful structure for further use. Clustering is one of the most important tasks in knowledge discovery from data. The goal of clustering is to discover the nature structure of data or detect meaningful groups from data. But for Big data application, clustering models are faced with the problem of analysing large dataset and hence, result in the need for more efficient algorithms to quickly analyse large datasets. Clustering techniques, like K-Means are useful in analysing data in a parallel fashion. K-Means largely depends upon a proper initialization to produce optimal results. However, DBSCAN algorithm has the quadratic time complexity, making it difficulty in real application with large dataset. Proposed approach presents a method which effectively reduce time complexity of clustering modelling based on K-Mean algorithm along with block operation which effectively reduces time costs of clustering modelling. Redesigning of distance function by using Manhattan distance instead of common Euclidean distance to simplify the calculation. The focus of this paper is to select a good initial seeding in less time, facilitating fast and accurate cluster analysis over large datasets.*

IndexTerms - K means clustering, Fast Clustering, DBSCAN, Min-Max, block operation.

I. INTRODUCTION

There has been a tremendous growth in the volume of data in the recent times. Data, whether it be structured or unstructured contribute to this enormous collection. To draw meaningful insights from this mountain of data we need algorithms which can perform analysis on this data. Clustering is the process of grouping data into groups called clusters, so that the objects in the same cluster are more similar to each other and more different from the objects in the other group [1]. Clustering divides the data into groups (cluster) that are meaningful, useful or both. Clustering problem has a long history [11]; K-means method is introduced in 1957, hierarchical clustering and graph-based clustering, while DBSCAN [7] is presented in 1996. Clustering has wide application such as biology, statistics, pattern recognition, information retrieval, and data mining.

Nowadays, the data being generated is not only huge in volume but is also stored across various machines all around the world. We need to process this data in parallel to reduce the cost of processing. K-Means is one of the most famous algorithms in the field of data mining [6]. Its scalability to large datasets and simplicity can be considered as one of the major reasons for its popularity. It is simple in data analysis and provides good performance. But it has a great dependence on the initial cluster centre. The selection of initial cluster centres determines the quality of clustering. Therefore, it is an important step to select a reasonable set of initial cluster centres in K-means algorithm. Density based clustering is one of the attractive methods that has been used in many applications. The key idea of density-based clustering is that the clusters are the areas with high density bounded by areas with lower object density. The algorithm DBSCAN uses the idea above to detect clusters and it can detect arbitrary-clusters with noise. However, DBSCAN has a quadratic complexity, it makes difficult in application with large data sets.

In this paper, we focus on the problem of boosting the performance of DBSCAN. The key idea of our approach is divide and conquer method including basically two steps: one is divide the datasets into blocks and then form k clusters: second apply choices to sample data using min-max method and then by DBSCAN hence recovered to obtain final clusters.

II. THE TRADITIONAL K - MEAN ALGORITHM [6]

An K-means algorithm is a clustering algorithm based on partition, proposed by McQUeen in 1976. The aim of Kmeans algorithm is to divide M points in N dimensions into K clusters so that the precision rate and the recall rate are maximum. It is not practical to require that the solution has maximum against all partitions, except when M, N are small and K=2. The algorithm seeks instead of “local” optima solution, such that no movement of an object from one cluster to another will reduce the within-cluster sum of squares.

The basic principle of the traditional K-means algorithm is: firstly, each data object in the data set is regarded as a single cluster, randomly select K data objects as the initial clustering centers; secondly, successively calculate the distance of the rest data objects to each of the K cluster center, each data object will be categorized into the nearest cluster, and then recalculate the centroid of each cluster; repeat iteratively until the cluster partition is no longer changed. The process of K-means algorithm is as follows:

Input: data set contained n data objects, k(the number of clusters) ;
Output: k clusters;
Step1: Randomly select K data objects as the initial cluster centres;
Step2: Calculate the distances from the remaining data objects to initial cluster centres, assigned the remaining n-k data objects to the nearest cluster;
Step3: Recalculate the cluster centres of each cluster;
Step4: repeat step2 and step3 until convergence;

K-means algorithm is a simple and efficient clustering algorithm [6]. Its time complexity is close to O(n*k). When the differences between categories are small or the scale of data set is large, K-means algorithm will perform more efficient, and get better clustering results. It has two major drawbacks- (1) A priori fixation of the number of clusters (2) Random selection of initial centres. So, there are different methods to improve the algorithm while maintaining its simplicity and efficiency.

III. THE DBSCAN ALGORITHM

The density-based clustering DBSCAN is an algorithm for clustering spatial data with the presence of noise proposed by M Ester et al. in 1996 [7]. We present here after the principle of the method.

Given a data set D with N objects, a similarity function (called dist) between two objects, parameter ε € R+.

Definition 1: (Eps-neighbourhood of the points) the Eps-neighbourhood of a point p, denoted by NEps(p), is defined by NEps(p) = {q € D | dist(p,q) ≤ Eps }

Definition 2: (directly density-reachable) A point p is directly density-reachable from a point q wrt. Eps, MinPts if

1) p € NEps(q) and
2) | NEps(q)| ≥ MinPts (core point condition)

Definition 3: (density-reachable) A point p is density-reachable from a point q wrt. Eps and MinPts if there is a chain of points p1,p2,…,pn, p1=q, pn=p such that pi+1 is directly-reachable from pi.

Definition 4: (density-connected) A point p is density-connected from a point q wrt. Eps and MinPts if there is a point such that both, p and q are directly-reachable from o wrt. Eps and MinPts.

Definition 5: (Cluster) Let D be a data base of points. A cluster C wrt. Eps and MinPts is a non-empty subset of D satisfying the following conditions:

1) □ p, q: if p € C and q is density-reachable from p wrt. Eps and MinPts, then q € C.
2) □ p, q € C: p is density-connected to q wrt. Eps and MinPts.

Definition 6: (noise) Let C1,…,Ck be the clusters of the database D wrt. Parameters Epsi and MinPtsi, i=1,…,k. Then we define the noise as the set of points in the database D not belonging to any cluster Ci.

To construct a cluster, DBSCAN starts with an arbitrary point p and retrieves all points density-reachable from p wrt. Eps and MinPts. If p is a core point, this procedure yields a cluster wrt. Eps and MinPts. If p is a border point, no points are diensity-reachable from p then DBSCAN visits the next point of the database.

IV. KDBSCAN USING K-MEANS AND DBSCAN

We proposed KDBSCAN including two steps: one is partition the datasets into blocks and then form k clusters using K-Means: second apply choices to sample data using min-max method and then apply clustering by DBSCAN hence recovered to obtain final clusters.

A. Block Operation: Let dataset D has M attributes and n instances. For each attribute, range is divided into f equal width. The feature space of D is separated to blocks of a size. N instances are assigned to these blocks and processed as one instance but weighted by number of instances in single block.

B. Apply Min-Max: The idea of the Min-Max Approach is to build a set of points Y from a dataset X such that the points in Y are far from each other and ensures a good coverage of the each cluster.

First a starting point y1 is randomly chosen from the dataset D. then, all the other points in Y are chosen among the points of X that Maximize their minimal distance from the points already in Y. the underlying idea of the Min-Max in the context of active learning is to select the points that is the farthest from the points that have already been used to formulate a query to the user.

C. KDBSCAN: the pseudo code of the algorithm presented in algorithm 1. We use block to partition data. We use K-Means in the first step to guarantee that the blocks chosen for step 2 will cover the whole data set. The second parameter t is the percent of data that will be chosen for DBSCAN. After using K-Mean, KDBSCAN extracts t percent of blocks by the Min-Max method (Algorithm 2).

IV. RESULTS AND DISCUSSION

```
Algorithm 1:

Input: A dataset D, the number of clusters for K-Means k, the proportion of data t;
Output: Clusters and noises;
1. Read Live Dataset and Parse data
2. Remove noise
3. Discrete all instance into blocks using block operation
4. K blocks are selected as initial cluster centres.
5. Partition data by modified k-mean
6. Take a proportion t of points (Min-Max algorithm) from clusters to form a new data set E; build a correspondence list to associate each selected point with its cluster.
7. Perform DBSCAN clustering on the set E
8. Recover the clusters detected by DBSCAN to form final clusters.
```

```
Algorithm 2:

Min-Max method
Input: dataset D, number of samples k
Output: set of points selected by Min-Max Y
• Take any reference point r
• Insert r in Y
• Temp =1
• while |temp|<=k+1
• Find the point x using Manhattan distance that maximize their minimal distance from the point already in Y
• Insert x in Y
• temp =temp + 1
• endwhile
• Remove r from Y
• return Y
```

Experimental dataset: Here is the list of datasets which have used in proposed method. These datasets are from NSE groups companies' data on stock price. The datasets have been taken from https://www.nseindia.com/products/content/equities/equities/eq_security.htm NSE site of stock data, the National Stock Exchange of Indian Ltd.

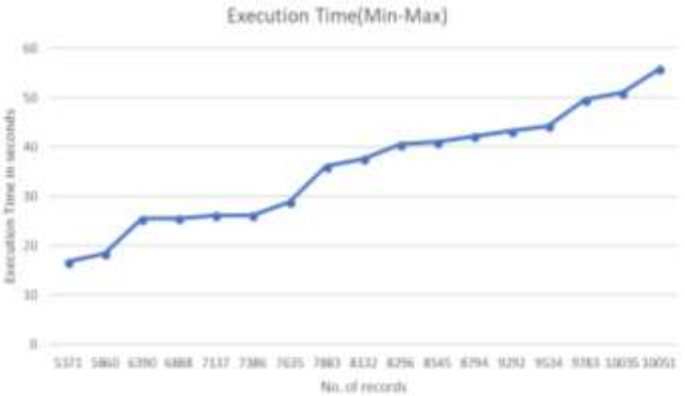

Fig. 1 Execution time of min-max method

The experimental results show that the execution time of Min-Max method as the no. of records increases in the algorithm the time for execution of method also increases gradually. Also in DBSCAN method the clustered elements or blocks increases as we increase the distance of radius or coverage area of the core object and respectively non-clustered blocks value decreases. These results shows the efficiency of DBSCAN algorithm on the blocks for clustering.

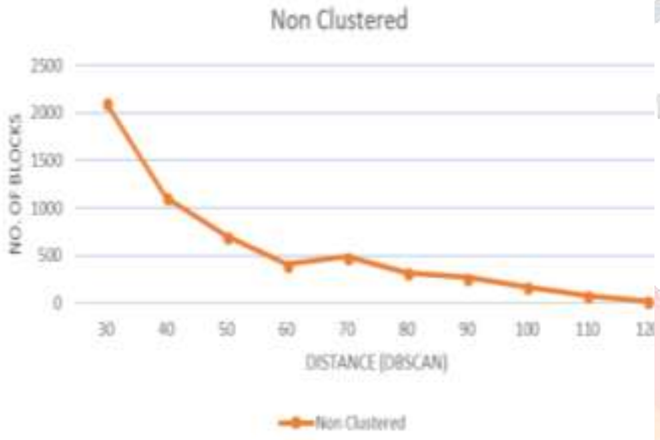

Fig. 2 Non-Clustered blocks after DBSCAN

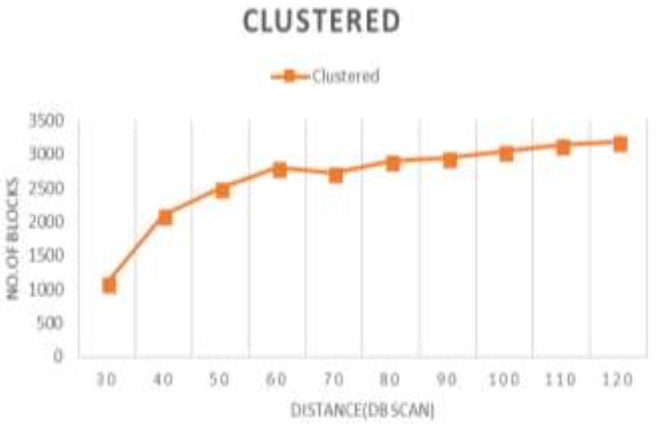

Fig. 3 clustered blocks in DBSCAN

This paper focuses on clustering of the big data modeling, have designed a new block operation method and replaced distance by Manhattan distance, applied a min-max method to select the core objects for DBSCAN algorithm and provide fast clustering compare to original DBSCAN.

V. Acknowledgment

The author wants to thank L.D. college of Engineering Ahmedabad, India for their support and guidance. I would also like to thank the members of Lab for their help.

References

[1] Anu Saini, G. B. Pant ,Jaypriya Ubriani “New Approach for Clustering of Big Data: DisK-Means”, 2016 IEEE ,International Conference on Computing, Communication and Automation ,pp 122-126;

[2] Kun niu, zhipeng gao,haizhen jaog ,haijie deng “K-mean+:a developed clustering algorithm for big data”, 2016 IEEE , Proceedings of CCIS2016,pp 141-144;

[3] Vadlana Baby,Dr. N. Subhash Chandra “Distributed threshold k-means clustering for privacy preserving data mining”,2016 IEEE,Conference on Advances in Computing, Communications and Informatics (ICACCI);

[4] Rasim Alguliyev , Ramiz Aliguliyev , Adil Bagirov , Rafael Karimov “Batch Clustering Algorithm for Big Data Sets”;

[5] Caiquan Xiong, Zhen Hua, Ke Lv, Wuhan Hubei ,“An Improved K-means text clustering algorithm By Optimizing initial cluster centers”, 2016 IEEE, International Conference on Cloud Computing and Big Data,pp 265-268;

[6] Jiawei Han, Jian Pei, Micheline Kamber “Data Mining: Concepts and Techniques” 3rd edition;

[7] Vu Viet Thang, D.V. Pantiukhin, A.I. Galushkin “A hybrid clustering algorithm : the FastDBSCAN” 2015 International Conference on Engineering and Telecommunication,pp 69-74;

[8] Tahereh Kamali, Daniel Stashuk “A Density-Based Clustering Approach to Motor Unit Potential Characterizations to Support Diagnosis of Neuromuscular Disorders” 2016 IEEE Transactions on Neural Systems and Rehabilitation Engineering ;

[9] Bin Jiang, Jian Pei, Yufei Tao and Xuemin Lin, Member, IEEE “Clustering Uncertain Data Based on Probability Distribution Similarity” IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 4, APRIL 2013;

[10] Chang Lu , Yueting Shi, Yueyang Chen, Shiqi Bao, Lixing Tang “Data Mining Applied to Oil Well Using K-means and DBSCAN” 2016 7th International Conference on Cloud Computing and Big Data;

[11] Jianbing Shen, Xiaopeng Hao, Zhiyuan Liang, Yu Liu, Wenguan Wang,and Ling Shao, Member, IEEE “Real-time Superpixel Segmentation by DBSCAN Clustering Algorithm” 2016 IEEE TRANSACTIONS ON IMAGE PROCESSING;

[12] Dongming Tang.Affinity propagation clustering for bid data based on Hadoop. Computer Engineering and Applications, 2015, 51(4):29-34;

[13] Joshua M.Dudik a, AtsukoKurosu b, JamesL.Coyle b, ErvinSejdić a,n “A comparative analysis of DBSCAN, K-means, and quadratic variation algorithms for automatic identification of swallows from swallowing accelerometry signals”, Computers in Biology and Medicine 59 (2015);

[14] Jesal Shethna “Data Mining Techniques available from https://www.educba.com/7-data-mining-techniques-for-best-results/” November 7, 2016;

[15] Martin Brown “Key techniques from https://www.ibm.com/developerworks/library/ba-data-mining-techniques/” Published on December 11, 2012;

[16] Data Mining tutorials “Data Mining Techniques from http://www.zentut.com/data-mining/data-mining-techniques/”

[17] Saurabh Arora, Inderveer Chana “A Survey of Clustering Techniques for Big Data Analysis” 2014 5th International Conference-Confluence The Next Generation Information Technology Summit (Confluence),pp 59-65.

[18] Martin Ester, Hans Peter Kriegel, Jorg Sander, Xiaowei Xu, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”, Published in Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96).