Volume 5, Issue 5, May 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

An Approach for Parallel K-means based on Dunn's Index

Sonal Sharma, Preeti Gupta, Pooja Parnami
Amity University, Rajasthan,
India

Abstract— Clustering is a method of unsupervised classification, where data points are grouped into cluster based on their similarity. K-means is a well-known and widely used algorithm for partitioned clustering. K-means algorithm is effective in producing clusters for several practical applications. In this Work a modified K-Means algorithm is proposed in order to overcome the problem of fixed number of clusters and the sequential execution. K-means algorithm takes the fix number of clusters at the time of input. But in practical scenario it is not possible to fix the number of clusters to get the optimal results

Keywords— K-Means, Dunn's index, Parallel execution, Improved K-Means

I. INTRODUCTION

Clustering is a method of unsupervised classification, where data points are grouped into cluster based on their similarity [1] .Clustering is also a method of partitioning a specified set of objects into the disjoint clusters. Clustering is used to group set of objects into classes such that similar objects are placed in the same cluster while dissimilar objects are in separate cluster [2].

K-means clustering is mostly used clustering algorithm which is used in various areas such as, information retrieval, and pattern recognition &computer vision. In standard k-mean algorithm we have to give the value of k in the number of cluster in advance. Practically it is very difficult to fix the value of k in advance. If the value of k fix in advance there can be chances that empty cluster problem will occur. K-means algorithm spends more execution time in computing the distance between data objects and cluster canters.

The Dunn index (DU) is a metric for evaluating clustering algorithms [15] Dunn's index well used for minimize the intra-cluster distance while maximizing the inter-cluster distance.

This modified algorithm proposed an approach to calculate the number of clusters with the help of Dunn's index. For the parallel implementation of K-means we will apply the Microsoft's Task Parallel Libraries (TPL) to execute our modified K-Means algorithm in a parallel manner. The modified algorithm is also tested with different input data and the results are compared it with original K-means to prove the efficiency of our modified algorithm

II. CLUSTERING

Clustering is a data mining technique of grouping set of data objects into multiple groups or clusters so that objects within the cluster have high similarity, but are very dissimilar to objects in the other clusters. Dissimilarities and similarities are assessed based on the attribute values describing the objects. Clustering algorithms are used to organize data, categorize data, for data compression and model construction, for detection of outliers etc. Common approach for all clustering techniques is to find clusters centre that will represent each cluster. Cluster centre will represent with input vector can tell which cluster this vector belong to by measuring a similarity metric between input vector and all cluster centre and determining which cluster is nearest or most similar one [4].

2.1 Classification of Clustering Algorithm:

Clustering is the main Part of Data Mining. And it is done by the number of algorithms. The most commonly used algorithms in Clustering are Hierarchical, Partitioning, Density and Grid based algorithms.

2.1.1 Hierarchical Clustering:

Hierarchical clustering is a method of cluster analysis which seeks to make a hierarchy of clusters. This algorithm also called the connectivity based clustering algorithm. [14]The hierarchical clustering outputs a hierarchy a structure that is more informative than the unstructured set of cluster.

2.1.2 Partitioning Clustering:

Partitioning algorithms divide data into several subsets. The reason of dividing the data into several subsets is that checking all possible subset systems is computationally not feasible [14]; there are certain greedy heuristics schemes are used in the form of iterative optimization.

2.1.3Density based Clustering:

Clusters are dense regions in the data spaces, separated by regions of lower object density. A cluster is defined as a maximal set of density connected points.

2.1.4 Grid based Clustering:

Grid-based clustering where the data space is quantized into finite number of cells which form the grid structure and perform clustering on the grids.[14] Grid based clustering maps the infinite number of data records in data streams to finite numbers of grids.

III. K-MEANS CLUSTERING

K-means is one of the most well-liked methods used to solve the clustering problems. A segment is to define k centroids, one for each cluster. The next segment is to take each points distance is generally considered to determine the distance between centroids and data points. When all the points are grouped in some clusters, the first step is done and an early grouping is done. At this point we need to recalculate the new centroids, as the addition of new points may lead to a change in the cluster centroids. When find k new centroids, a new binding is to be created between the same data points and the nearest new centroids, generating a loop. As a result of this loop, the k Centroids may change their position in a step by step manner. Eventually, a situation will be reached where the centroids do not move anymore. [3, 10, 11].

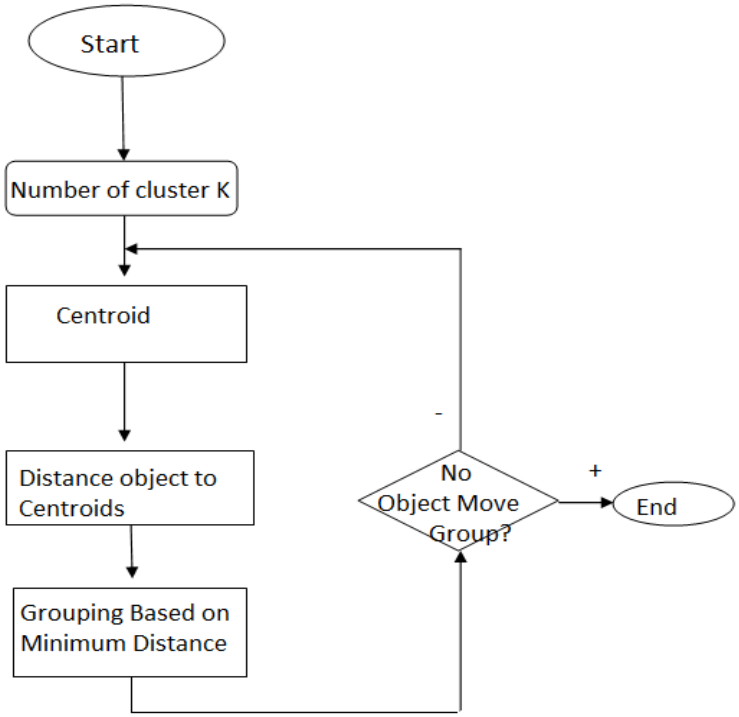

FIGURE 1.STEPS OF K-MEANS ALGORITHM

2 K-Means Algorithm:

Input:

D= [d1,d2,…dn] ///sets of n data items.

K = number of desired clusters

Output:

A set of k clusters.

Steps:

1. Select k data items from D as initial centroids;
2. Repeat the process of selecting the items
3. Assign the each item di to the cluster which has the nearest and the suitable centroids;
4. Calculate the new mean value for each cluster;
5. Repeat the process until the criteria is satisfied. [3], [10], [11].

The working of Algorithm can be explained clearly with the help of an example,

This is shown on Figure 1.Figure 1 shows the graphical representation for working of K-means algorithm. In the first step there are two sets of objects. Then the centroids of both sets are determined. According to the centroid again the clusters are formed which gave the different clusters of dataset. This process repeats until the best clusters are achieved. [9]

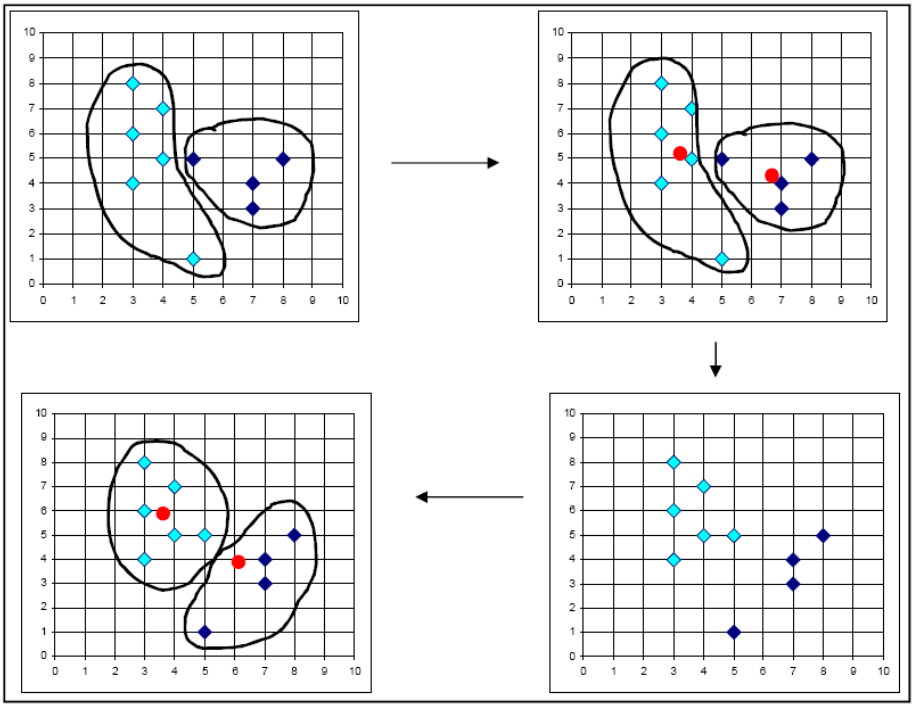

FIGURE 2.WORKING OF K-MEANS ALGORITHM

3.3 K-Means Limitation:

K-means clustering has some of the limitations which need to get overcome. Several people got multiple limitations while working on their research with K-means algorithm. Some of the common limitations are discussed below

3.3.1Number of clusters

Determining the number of clusters in advance is always been a challenging task for K-means clustering approach. It is beneficial to determine the correct number of clusters in the beginning. It has been observed that sometimes the numbers of clusters are assigned according to the number of classes present in the dataset. Still it is an issue that on what basis the number of clusters should be assigned [12].

3.3.2Empty clusters

If no points are allocated to a cluster during the starting step, then the empty clusters problem occurs. It was an former problem with the traditional K-means clustering algorithm [8].

IV. PROBLEM STATEMENT

The K-means algorithm takes number of clusters (K) as input from the client [6]. Practically it is very difficult to fix the value of k in advance will lead to a poor quality cluster [13] .The k-means algorithm spends mostly time in execution to computing the distances between data objects and cluster centers. To make the system more efficient compute the major parts of the task in parallel [7]

V. PROPOSED SOLUTION

The standard k-means algorithm need to calculate the distance between every data object and the centers of k clusters when it executes the iteration every time; it takes up more execution time mainly for large datasets. The proposed solution overcome the shortcomings of k-means algorithm.The main idea of algorithm is to calculate the optimal no of clusters and speed up the execution to implement k-means algorithm parallel

Modified Algorithm:

Input: k=number of clusters, calculate through Dunn's index

D= A data set containing n objects.

Output:

A set of k clusters.

Method:

1. Randomly choose k objects from D as initial cluster centers.
2. Do again.
3. (Re) assign each object to the cluster to which object is more similar, calculation based on the Dunn'sindex.

Dunn's index calculation

i. Find the Intra cluster Maximum distance from all clusters.
ii. Then calculate the Maximum value from among these distances.

iii. Find the Inter cluster Minimum distance from all clusters.
iv. Then calculate the Minimum value from among these distances.
v. Apply the Dunn's Equation to get the feasible number of clusters

$$Dunn'sIndex = \frac{Mind(i,j)}{Maxd'(k)}$$

4. Update the cluster means,
5. Output the clustering result

VI. CONCLUSIONS

K-means is a clustering algorithm and it is broadly used for clustering large sets of data. This paper elaborates k-means algorithm and analyses the shortcomings of the standard k-means clustering algorithm. If no points are allocated to a cluster during the assignment step, then the empty clusters occurs so the proposed method is Use Dunn's index for finding the optimal no of clusters that will not lead the empty cluster problem. To speed up the execution implement the k-means algorithm parallel. The proposed k-means method is feasible.

REFERENCES

[1] "Data mining concepts and techniques," Michelin andJ.Han,morgan Kauffman, 2006.
[2] "Genetic Algorithm-Based Clustering Technique" U.Maulik, and S.Bandyopadhyay,Pattern Recognition 33, 1999.
[3] M.Murty & K.Krishna,"Genetic k-means Algorithm,"IEEE Transactions on System, Vol.29, No.3, 1999.
[4] " A Comparative Study of Various Clustering Algorithms in Data Mining," International Journal of Engineering Reserch and Applications (IJERA),Manish Verma, Mauly Srivastava, NehaChack, Atul Kumar Diswar, Nidhi Gupta,Vol. 2, Issue 3, pp.1379-1384, 2012
[5] "Comparative Study of Parallel Programming Models to Compute Complex Algorithm",Pradeep Soni Mukul Sharma International Journal of Computer Applications (0975 – 8887) Volume 96– No.19, June 2014
[6] Volume 10, Number 3, June 2005,TSINGHUA SCIENCE AND TECHNOLOGY ISSN 1007-0214 01/21 pp277-281"Improvement and Parallelism of K-Means Clustering Algorithm"TIAN Jinlan ZHU Lin ZHANG Suqin
[7] "Dynamic Clustering of Data with Modified K-Means Algorithm"AhamedShafeeq B M andHareesha K S 2012 (ICICN 2012) International Conference on Information and Computer Networks, IPCSIT vol. 27 (2012) © (2012) IACSIT Press, Singapore
[8] "Top 10 algorithms in data mining", Knowledge and InformationSystems, January 2008, Volume 14, Issue 1, pp 1-37, X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.H. Zhou, M. Steinbach, D.J. Hand, D. Steinberg,.
[9] "Dynamic load balancing on GPU clusters forlarge-scale K-Means clustering, " E. Kijsipongse, S. U-ruekolan2012 IEEE International Joint Conferenceon Computer Science and Software Engineering (JCSSE), vol., no., pp.346,350, May 30 2012-June 1 2012
[10] "Refining Initial Points for K-means Clustering,"P.Bradley, and U.Fayyad, In Proceeding of 15th International Conference on Machine Learning, 1998.
[11] "Improving the Accurancy and Effiency of the K-means Clustering Algorithm ",K.A Abdul Nazeer , M.P Sebastian WCE 2009, London.
[12] O. A. Abbas,"comparisons between data clustering algorithms", the international Arab journal of information technology, vol. 5, no. 3, July 2008.
[13] " A Review of K-mean Algorithm" Monika Sharma,Jyoti Yadav,International Journal of Engineering Trends and Technology (IJETT) – Volume 4 Issue 7- July 2013.
[14] Amandeep Kaur Mann, NavneetKaur**,**"Survey Paper on Clustering Techniques ",Volume 2, Issue 4,ISSN: 2278 – 7798 International Journal of Science, Engineering and Technology Research (IJSETR), April 2013
[15] **"**Review and Comparative Study of Cluster ValidityTechniques Using K-Means Algorithm", RomanaRiyaz, MohdArifwani.International Journal of Advance Foundation and Research in Science & Engineering (IJAFRSE) Volume 1, Issue 3, August 2014.

