Refining the clustering of the k-mean clustering algorithm using of the Leader-follower algorithm.

Efficiency of clustering using Leader-follower

Sowndarya. S
B.Tech, Information & Comm. Technology
SASTRA University, Tirumalaisamudram
Tanjore, TamilNadu, India

Sridhar. A
B.Tech, Information Technology
SASTRA University, Tirumalaisamudram
Tanjore, TamilNadu, India

ABSTRACT-This paper is presented with the implementation of Leader-Follower clustering algorithm to improve the efficiency of clustering in k-means clustering algorithm. This Leader-follower algorithm, though unstable, can be implemented with clustering algorithms, leading to the improved efficiency of the clustering. Clustering is the process of grouping a given data set upon the characteristics they possess. K-means clustering algorithm has its main concentration on the centroids, which leads to clustering of the data. Leader-follower is a clustering algorithm that clusters data upon a given threshold. Here, the output of Leader-follower is taken and is fed as input to k-means, which refines the clustering process with improved centroid selections, that leads to effective clustering.

Keywords: leader-follower, k-means, clustering, outlier.

I. INTRODUCTION

Data mining – Clustering:

Data mining is the process extracting of information, patterns, and etc., from large quantities of data. Data mining is classified into two major types: They are:

- Directed data mining
- Undirected data mining

Directed data mining is like a black box, where, the user need not bother about the functions inside, but cares only about the input and the output. Clustering comes under the category of undirected data mining and it is an unsupervised learning technique, that groups data on the characteristics they possess. The general criterion for a good clustering is that, the data objects within a cluster are closely related to each other but are different from the objects in other clusters.

The objective of the Clustering algorithms (k-mean clustering) is to group the similar data together depending upon the characteristics they possess. Clustering plays a major role in pattern recognition, image analysis, market and business research and etc. Clustering algorithms can be classified as follows:

- Density Based clustering (DBSCAN)
- Partitional clustering
- Spectral clustering
- Model based clustering
- Graph based clustering
- Hierarchical clustering

Convergence Criteria-Squared Error criterion:

When the old mean value and new mean value becomes equal, then it is said that the convergence criteria is met.

The below is the theoretical equation for the convergence criterion, i.e., the squared error criterion.

$$J = \sum_{j-1}^{k} \sum_{i-1}^{n} \| x_i^{(j)} - c_j \|^2$$

Outliers:

Outlier detection and handling is a very important part of any modeling exercise. A failure to detect outliers or their inefficient handling can have serious ramifications on the validity of the inferences drawn from the

exercise. There are a large number of techniques available to perform this task, and often selection of the most appropriate technique poses a big challenge to the practitioner. To be even more practical, there is no standardized method for outlier detection. Some of the outlier detection techniques are:

- Distance based outlier detection
- Deviation based outlier detection
- Clustering based outlier detection
- Density based outlier detection
- Depth based outlier detection.

II. IMPLEMENTATION OF LEADER FOLLOWER WITH K-MEANS CLUSTERING ALGORITHM

1. K-Means:

K- means clustering algorithm groups data upon the distance between the centroids and the input values of the data set. Mean value is found out for every cluster and is set as the updated centroid value. This is iterated until the convergence criterion is met.

Algorithm (k-means):

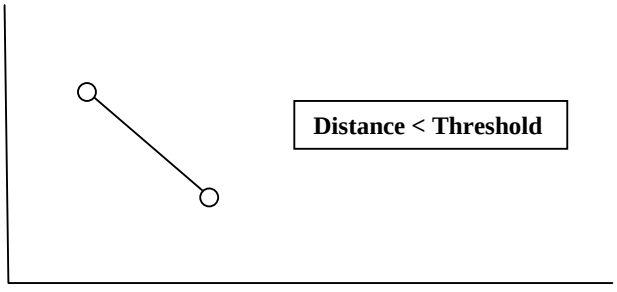
- Assign initial values for means m1, m2….mn.
- Assign each input data to the cluster which has nearest distance.
- Calculate the new mean value for each cluster until the convergence criterion is met.

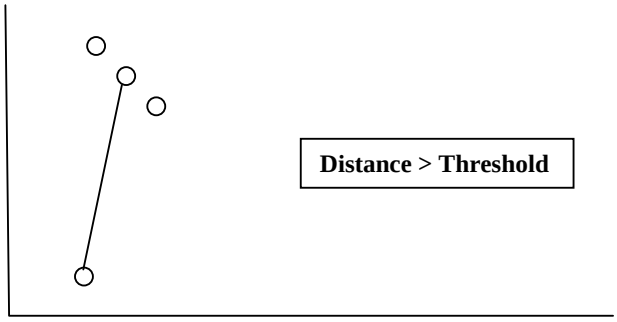
2. Leader-Follower:

The leader follower algorithm calculates the number of clusters and the cluster value with the help of a threshold value. By initializing the threshold value, every input from the input data set is compared with the threshold and if it is greater, a new cluster is created, else it is added as an instance to the cluster containing data lesser than the threshold. There is no advanced cluster selection. Only the threshold value is specified.

Algorithm (Leader-Follower):

- Initialize the input data set.
- Specify the threshold distance.
- Find the closest cluster centre.
- If the distance from the cluster centre is above threshold? Create new cluster.
- Else, add as an instance to the cluster.





3. Leader-follower with k-means:

In K-means clustering algorithm, we have to determine the number of clusters in advance. The selection of number of clusters in advance, has a major effect on the overall output of the algorithm. If the selected clusters are not optimum ie., if its much higher or lower than needed, it may distort the real clustering structure. To avoid such situation and to get a proper idea over the required number of clusters for efficient running of the algorithm Leader-follower is implemented, where the total number of clusters is decided upon a threshold value.

Algorithm:

- Initialize the input data set.
- Specify the threshold value.
- Find the closest cluster center.
- If the distance of the cluster center is above the threshold? Create new cluster.
- Else, add it as an instance to the cluster containing values lesser than the threshold.
- Find the overall output - total number of clusters created.
- Feed the above output as the total number of 'm' values – clusters in the k-mean algorithm for the same input data set
- With m values selected, select the centroids with any of the mean value selection methods.
- Assign each input data to the cluster which has nearest distance.
- Calculate the new mean value for each cluster until the convergence criterion is met

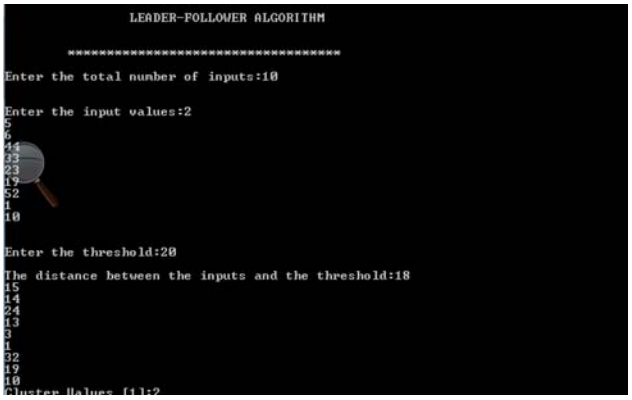
III. SIMULATION AND RESULT

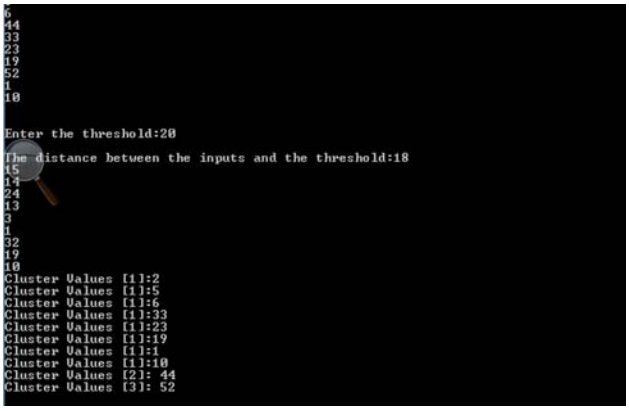
1. Tabular Column Based for Leader-Follower and K-Means Clustering Algorithm

Total no of inputs	**Total no. of Clusters**	**Leader Follower***	**K-Means***
10	1	Runtime: 0.015000	Runtime:0.062000 No.of outliers: 5
10	3	Runtime: 0.062000	Runtime:0.012400 No.of outliers:7
10 (**without leader-follower**)	5		Runtime:0.131000 No.of outliers:4

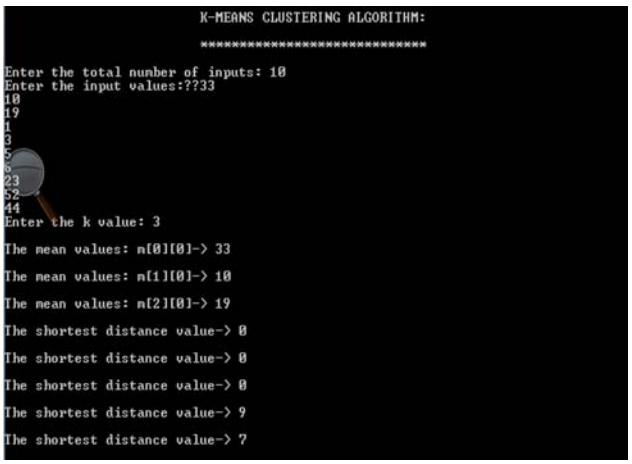
***-runtime depends on the processor for every individual run.**

2. Leader-Follower:

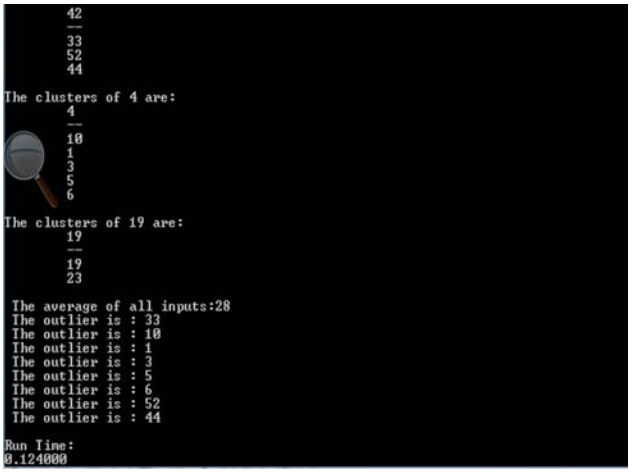




3. k-means-Initial Clustering:



4. Final Output-Outlier Detection:



IV. CONCLUSION

By implementing the leader follower algorithm for the k-means clustering algorithm, the clustering of the k-means algorithm gets refined. By implementing the leader-follower for the input data set (2,5,6,44,33,23,19,52,1,10) the total number of clusters becomes 3 with shorter run time. But when this input data set is implemented in k-means without the leader- follower, the run time increases and also the total number of clusters is unknown which reduces the efficiency.

V. REFERENCES

[1] Yashwanth K Kanethker, *Let Us C*, 5th ed., BPB publications, New Delhi.
[2] J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2000.
[3] (2002) The IEEE website. [Online]. Available: http://www.ieee.org/
[4] Wikipedia search [Online]. Available: http://www.wikipedia.org//
[5] Yinghua Zhou Hong Yu Xuemei Cai, Coll. of Comput.Sci. & Technol., Chongqing Univ. of Posts & Telecommun, Chongqing, China. A Novel K-Means algorithm for Clustering and Outlier Detection, 13-14 Dec 2009
[6] Rui Xu, Donald Wunsch, "Survey of clustering algorithms," IEEE Transactions on Neural Networks, vol. 16, no. 3, May 2005, pp. 645-678
[7] Mu-Chun Su and Chien-Hsing Chou, "A modified version of the K-means algorithm with a distance based on cluster symmetry," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23,no. 6, June 2001, pp. 674-680.
[8] David Arthur and Sergei Vassilvitskii, "k-means++: the advantages of careful seeding," Proceedings of the 18th annual ACM-SIAM symposium on Discrete algorithms, 2007, pp.1027-1035.
[9] A. Sridhar and Sowndarya. S, "Efficiency of K-Means Clustering Algorithm in Mining Outliers from Large Data Sets", International Journal on Computer Science and Engineering, IJCSE10-02-09-088.