ISSN:0975-9646
Manikandan.G et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 (4) , 2011, 1747-1749

Performance Analysis of Clustering Algorithms in Outlier Detection Based on Statistical Models and Spatial Proximity

[1]Manikandan.G, [1]Rajendiran.P, [2]Kamarasan.M, [1]SowndaryaShekar
[1]*School of Computing, SASTRA University, Thanjavur-613401, Tamil Nadu, India.*
[2] *CSE Department, Annamalai University, Chidambaram-608002, Tamil Nadu, India.*

Abstract **- This paper presents the analysis of leader- follower, k-means and k-medians clustering algorithms in outlier detection based on some statistical models and spatial proximity. Clustering and classification plays a vital role in data mining. Clustering groups the similar data together based on the characteristics they possess. Clustering, which is so much used in pattern recognition, reduces the searching load. Leader-follower algorithm is the simplest one. K-means clustering algorithm clusters the similar data with the help of the mean value and squared error criterion whereas in k-medians algorithm, median value is used. Outliers, the one which is different from norm, should be detected and handled properly. Otherwise, it will affect the original data in clustering in a great manner. Dataset for simulation has been generated using "weka" software.**

Keywords:leader-follower, k-means, k-medians, clustering, outliers

I. INTRODUCTION

Data mining is the process of posting queries to large quantities of data and gathering useful information from them. Data Mining has evolved from multiple technologies, including data management, data warehousing, machine learning, and statistical reasoning.Clustering is a kind of unsupervised classification technique, which is used to group the data into different classes or clusters, without class label predefined. The general criterion for a good clustering is that the data objects within a cluster are similar or closely related to each other but are very dissimilar to or different from the objects in other clusters. Many fields imply on data mining like games, business, surveillance, science and engineering etc.

II. LITERATURE REVIEW

Clustering plays a major role in pattern recognition, image analysis, market and business research and it reduces the searching load and time. Clustering algorithms can be grouped into different categories such as hierarchical clustering, partitional clustering and spectral clustering.

Desirable Features in Clustering Algorithms are:

- Scalability
- Robustness
- Order insensitivity
- Minimum user-specified input
- Point proportion admissibility: Duplicating data set and re-clustering should not change the results.

Cluster analysis is an important activity. The basic requirements of clustering in data mining are: Scalability, ability to deal with noisy data, ability to deal with different types of attributes, usability and interpretability.

In K-means algorithm, we have to determine the number of clusters in advance. The number of clusters must be estimated via the procedure of cluster analysis. The selection of number of clusters in advance,may distort the real clustering structure, so only Leader follower is needed.

Outlier detection is very important in data mining activities and involves identifying a set of observations whose values deviate from the expected range. Outlier arises due to the changes in the system behaviour, human error or instrument error. It is a bad practice to ignore the outliers. Either it should be handled or at least it should be detected from the rest of the inputs.

Effects of outliers are: Less efficient outputs, more error prone and improper results

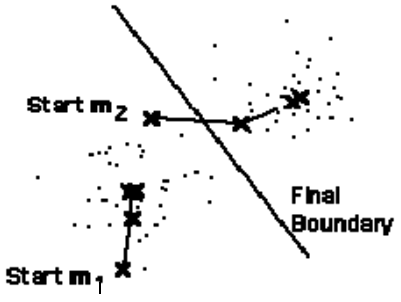
III. ALGORİTHM DESCRİPTİON

a. Leader-Follower Algorithm:

- Assign initial values for clustering.
- Specify threshold distance.
- If the distance is above threshold value, create new cluster.
- If the distance is below threshold value, add instance to cluster.

b. k-means Algorithm:

- Assign initial values for means m1, m2….mn.
- Assign each item to the cluster which has nearest mean.
- Calculate new mean for each cluster until the convergence criteria is met.



c. k-medians Algorithm:

- Assign initial values for means m1, m2….mn.
- Assign each item to the cluster which has nearest mean.

- Calculate new median for each cluster until the convergence criteria is met.
- If the total number of elements in the cluster ends with an even number then take the middle two values and calculate the new median
- If the total number of elements in the cluster ends with an odd number then take the middle value as median.

d. Convergence Criteria:

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

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

where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre c_j, is an indicator of the distance of the *n* data points from their respective cluster centres.

IV. SIMULATION AND RESULT

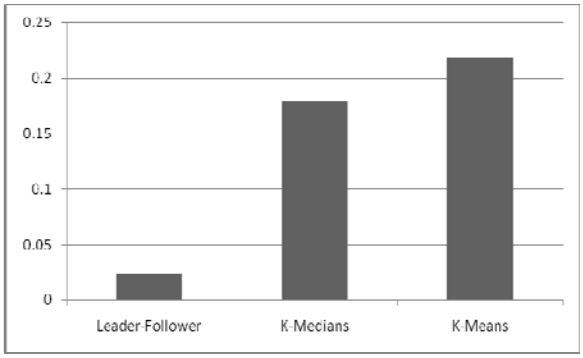
The above algorithms have been implemented in C and the result was tested in Intel Core 2 Duo Processor with 1GB of main memory and windows 7 operating system. CPU Dataset has been fed as input for the algorithms which was taken from Weka software. Weka was mainly developed for Data Mining using Java. It is open source software. The algorithms can either be applied to a dataset or can be called from the code. Weka contains tools for classification, regression, clustering, association rules and visualization.

Table 1. Experiment Results

	Total inputs	Leader Follower*	K-Medians*	K-Means*
Spatial Proximity	20	0.024	0.179	0.219
Statistical Model	20	0.028	0.216	0.232

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

Chart 1:



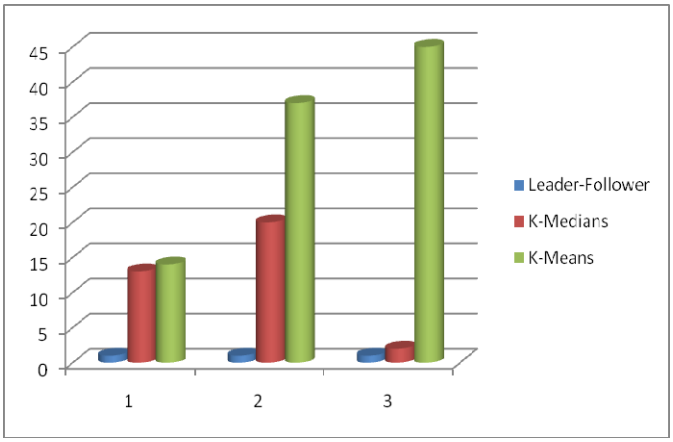
As a result, we can say that, leader follower is fastest algorithm when compared to the other two.

From the above results, leader follower takes (0.024 sec) to compute the clustering while other clustering algorithms take 0.179 and 0.219 seconds to compute the results.

Table 2:

No. of inputs	Leader Follower* (No. Of Outliers)	K-Medians* (No. Of Outliers)	K-Means* (No. Of Outliers)
20	1	13	14
40	1	20	37
60	1	2	45

Chart 2:



The above chart explains the outlier detection for different set of input values. The k-medians clustering algorithm detects the outlier in an efficient manner even though its run time is more when compared to leader follower clustering algorithm. K-medians detects upto 1 norm distance whereas this k-means detects only upto 2 norm distance.

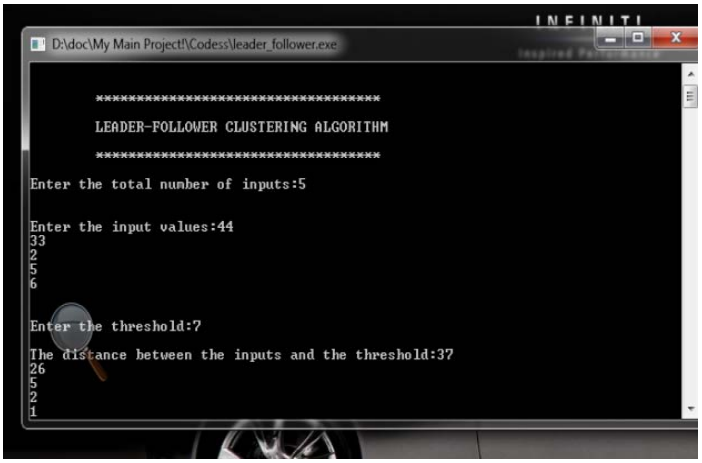


Fig : 1 - Leader Follower:

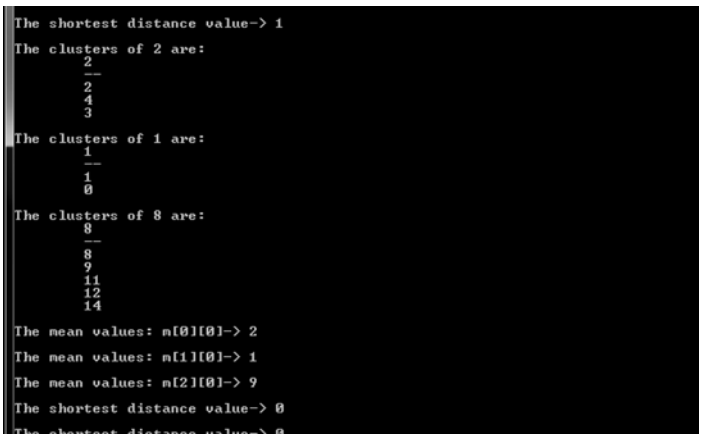


Fig : 2- Initial Clustering:

V. CONCLUSION

Leader Follower algorithm is the best clustering algorithm as it is very fast when compared to the k means and k medians. In minimizing the errors, K-Medians Clustering algorithm is efficient enough than K-Means Clustering algorithm. In future, this work may be extended with the implementation of some other algorithms, to check which is giving the best result.

VI. 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] 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

[4] Rui Xu, Donald Wunsch, "Survey of clustering algorithms," IEEE Transactions on Neural Networks, vol. 16, no. 3, May 2005, pp. 645-678

[5] 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.

[6] 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.

[7] Srinivasa K G, Venugopal K R, L M Patnaik , 'Feature Extraction using Fuzzy C-means Clustering for data mining systems', International Journal of Computer Science and Network Security,Vol.6,No.3A, March 2006, pp230-236.

[8] U. Boryczka, "Finding groups in data: Cluster analysis with ants," Applied Soft Computing Journal, vol. 9, pp.61-70, 2009.

[9] K. J. Cios, W. Pedrycz, and R. M. Swiniarsk, "Data mining methods for knowledge discovery, IEEE Transactions on Neural Networks, vol. 9, pp. 1533- 1534, 1998.