ISSN:0975-9646

Performance Analysis of Clustering Algorithms in Detecting Outliers

Sairam[1], Manikandan[2], Sowndarya[3]
School of Computing, SASTRA University, Thanjavur
Tamil Nadu, India.

Abstract **-** This paper presents the analysis of K-means and K-Medians clustering algorithm in detecting outliers. Clustering is generally used in pattern recognition where if a user wants to search for some particular pattern, clustering reduces the searching load. The k-means clustering and k-medians clustering algorithm's performance in detecting outliers are analysed here. K-means clustering clusters the similar data with the help of the mean value and squared error criterion. K-medians is similar to k-means algorithm but median values are calculated there. Outliers are the one different from norm. If they are not properly detected and handled, they clustering will be affected in a great manner.

Keywords: Clustering, k-Means, k-Medians, Outliers

I.INTRODUCTION

Data mining is the process used to analyse large quantities of data and gather useful information from them. It extracts the hidden information from large heterogeneous databases in many different dimensions and finally summarizes it into categories and relations of data. Clustering and classifications are the two main techniques of data mining followed by association rules, predictions, estimations and regressions. Many fields imply on data mining like games, business, surveillance, science and engineering etc.

II. LITERATURE REVIEW

The objective of the Clustering algorithms 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 it reduces the searching load and time.

Clustering algorithms can be grouped into different categories such as

- Hierarchical clustering
- Partitional clustering
- Spectral clustering

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
- Interpretability

The k-means and k-medians clustering algorithm comes under partitional clustering. Based on the similarity function such as distance, it clusters the input data.

Outliers, not a part of a cluster, are specific points which behave very differently from the norm. 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. They are detected using distance based algorithm here. This algorithm says that outliers are the one which are far away in distance from the rest of the inputs.

Effects of outliers are

- Less efficient outputs
- More error prone
- Improper results

III. Algorithm Description

A. 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.

There are two methods for assigning mean values:

- First k input values
- Randomly assigned

B. 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.

C .Convergence Criteria:

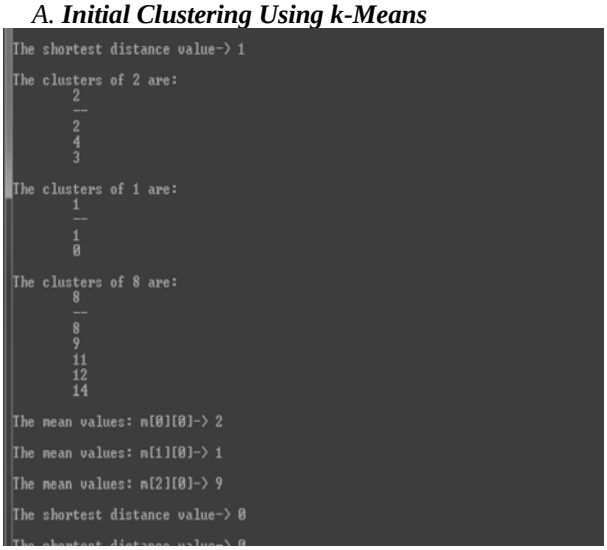
When the old mean value and new mean value becomes equal, then it is said that the convergence criteria is met.

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

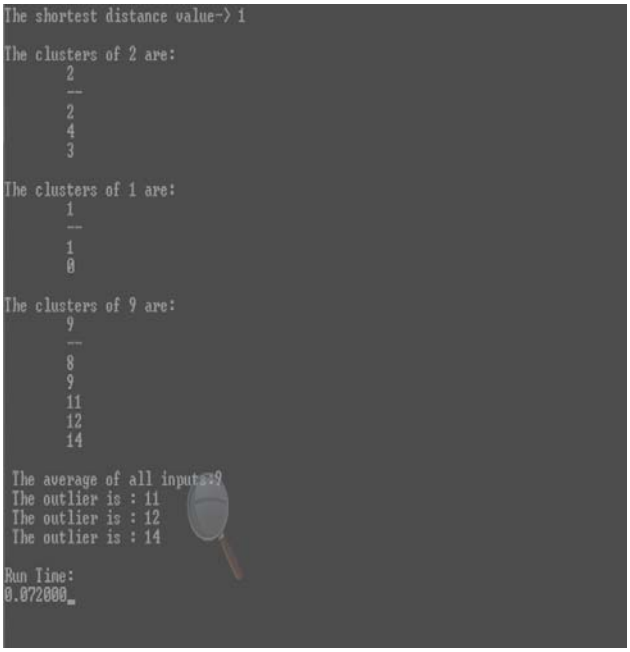
IV. Simulation and Result

Simulations are carried out to compare the performance of the Clustering Algorithms and the results are summarized as follows.

*A. **Initial Clustering Using k-Means***



B. Final Output-Outlier



Total No of Inputs	Total No of Clusters	K-Means* (Run Time)	K-Means * 2-Method (Run Time)	K-Medians* (Run Time)
10	2	0.127000	0.129000	0.112000
10	3	0.115000	0.119000	0.114000
10	4	0.127000	0.128000	0.112000

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

V. Conclusion

K-Means Clustering algorithm is taking more time to compute the outliers also for example, if inputs such as -2,3,1,4,30,12,8,7,10,13 is given, 30 which is an outlier is detected properly by K-Medians Clustering whereas in K-Means Clustering, -2,1,30,12,13 are detected as outliers.

So, in minimizing the errors, K-Medians Clustering algorithm is efficient enough than K-Means Clustering algorithm.

In this paper we have used shortest distance method for detecting outliers. In future, this work may be extended with other algorithms and other methods.

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] Wikipedia search [Online]. Available: http://www.wikipedia.org//

[4] 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

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

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

Sairam Natarajan, received his M.Tech.(Computer Science & Engineering) and Ph.D from SASTRA University,Thanjavur,TamilNadu. He is currently working as Professor in Computer Science and Engineering Department, SASTRAUniversity. Member of Computer Society of India with 15 years of teaching experience. and has 7 publications to his credit..

Manikandan Ganesan ,received his M.Tech.(Computer Science & Engineering) from SASTRA University,Thanjavur,TamilNadu. He is currently working as Assistant Professor in Information and communication Technology Department ,SASTRAUniversity He has much interest towards research in Privacy Preserving Data Mining and Network Security.

Sowndarya Sekhar Pursuing final year B.Tech.in Information and communication Technology From SASTRA University .