Outlier Detection: A Clustering-Based Approach

Vijay Kumar, Sunil Kumar, Ajay Kumar Singh

Abstract— Outlier detection is a fundamental issue in data mining; specifically it has been used to detect and remove anomalous objects from data. It is an extremely important task in a wide variety of application domains. In this paper, a proposed method based on clustering approaches for outlier detection is presented. We first perform the Partitioning Around Medoids (PAM) clustering algorithm. Small clusters are then determined and considered as outlier clusters. The rest of outliers (if any) are then detected in the remaining clusters based on calculating the absolute distances between the medoid of the current cluster and each one of the points in the same cluster. Experimental results show that our method works well.

Index Terms— PAM, Clustering, Clustering-based outliers, Outlier detection.

I. INTRODUCTION

Cluster analysis or clustering is the task of assigning a set of objects into groups called clusters so that the objects in the same cluster are more similar in some sense to each other than to those in other clusters.

Clustering is unsupervised classification:[1] no predefined classes. Clustering is the process of grouping a set of physical or abstract object into classes of similar objects is called clustering.

Cluster: A collection of data objects-
- Similar to one another within the same cluster
- Dissimilar to the objects in other clusters

Principle: Maximizing intra-class similarity & minimizing interclass similarity as shown in figure 1.
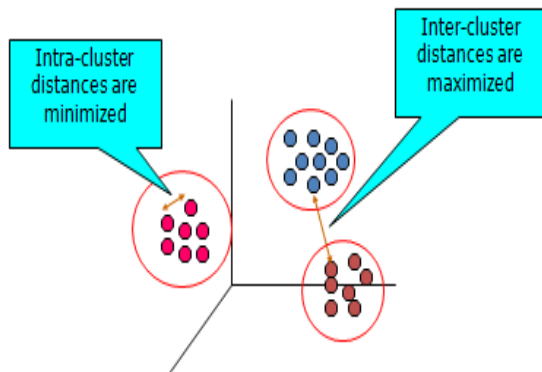

Fig. 1: Cluster Analysis

Conventional clustering measures the similarity of the data objects based on the geometric distance (i.e. Minkowski distance).

Minkowski distance of order p between two points P=(x1, x2,….xn) and Q=(y1,y2….yn) is defined as:

$$d(x, y) = \sum_{i=1}^{n} (| x_i - y_i |^p)^{1/p}$$

Minkowski distance is typically used with p being 1 or 2. Clustering is an important tool for outlier analysis with the key assumption that normal objects belong to large and dense clusters, while outliers form very small clusters.

Outliers are the set of objects that are considerably dissimilar from the remainder of the data. Outliers may be detected using statistical tests that assume a distribution or probability model for the data, or using distance measures where objects that are at a substantial distance from any other cluster are considered as outliers.

II. EXISTING SYSTEM

Clustering is not a new concept but data clustering together with outlier detection is a young scientific discipline under vigorous development. As a branch of statistics, cluster analysis has been studied extensively for many years, focusing mainly on distance based cluster analysis techniques. But there is a revolution happening right now in the way of interpretation of outliers, as now importance is laid on the detection of object whose behavior is completely different from the other objects in dataset.
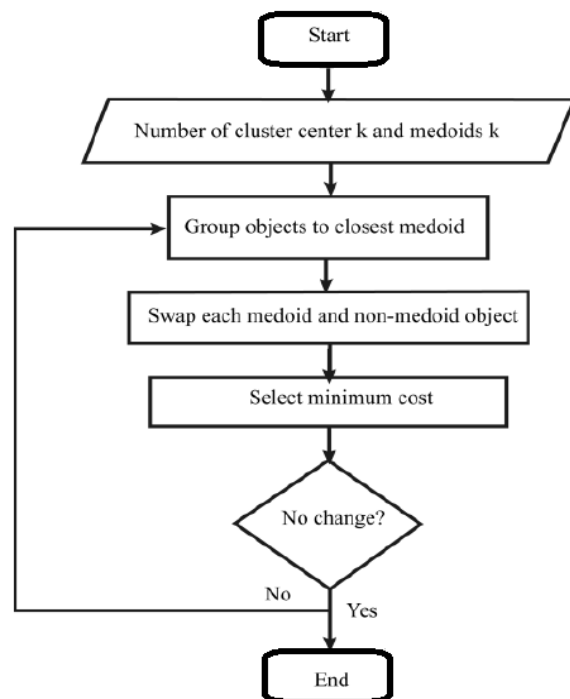

Fig.2: Flow Chart of PAM Algorithm

It has been argued by many researchers whether clustering algorithms are an appropriate choice for outlier detection. For example, in [2], the authors reported that clustering algorithms should not be considered as outlier detection methods. This might be true for

Manuscript Received June 07, 2013.
Asst. Prof. Vijay Kumar, Dept. of CSE MIET, Meerut (U.P.), India.
Associate, Prof Sunil Kumar, Dept. of CSE, JPIET, Meerut (U.P.), India.
Prof. (Dr.) Ajay Kumar Singh, Dept. of CSE MIET, Meerut (U.P.), India.

some of the clustering algorithms, such as the k-means clustering algorithm. In paper [3] the authors proposed a clustering- based approach to detect outliers. The K-means clustering algorithm is used. As mentioned in [4, 5] the k-means is sensitive to outliers, and hence may not give accurate results.

PAM clustering algorithm in [6] uses the most centrally located object in a cluster (called medoid) instead of the cluster mean. It was shown in [4, 7] that the medoids produced by PAM provide better class separation than the means produced by the K-means clustering algorithm. Flow chart is shown in figure 2.

PAM Clustering Algorithm:
Input:
  K: The no. of clusters.
  D: Dataset containing n objects.
Output: A set of K-clusters.
Method:
1. Arbitrarily choose k objects in D as the initial medoids.
2. Repeat.
3. Assign each remaining object to the cluster with the nearest medoid.
4. Randomly select a non-medoid object, O_{random}.
5. Compute the total cost (S) of swapping initial medoid (O_j) with O_{random}.
6. If S < 0 then swap O_j with O_{random} to form the new set of k-medoids.
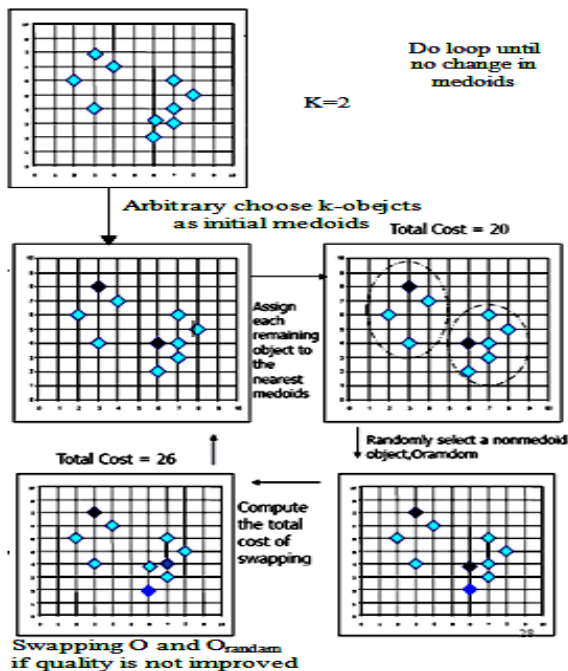7. Until there is no change in the medoids.

Fig. 3: Example of PAM

- PAM is more robust than k-means in the presence of outliers because a medoid is less influenced by outliers [8] or other extreme values than a mean
- This indicates that cost and size of the datasets are directly proportional, that is small datasets have low computation cost, as only a few iterations are required and makes PAM an impractical solution for large datasets.

Thus we need an efficient outlier detection mechanism which finds all the outliers in a given dataset. An example is given in figure 3.

III. PROPOSED APPROACH

Outlier detection is a very important research work in the field of data mining. Here we argue about the PAM clustering algorithm and a new approach for outlier's detection is proposed.
In this approach we tend to find outliers in *two phases.*
Initially we Perform PAM clustering algorithm to produce a set of k clusters.
-In first phase every cluster is checked for number of objects it contains. If a cluster contains less than average number of objects (n/k) then all the objects in the cluster are detected as outliers. That is, the entire cluster is selected as outlier cluster.
-In the second phase for each cluster the absolute distance between each object and the cluster medoid is calculated. This is called the Absolute Distance between the Medoids and Point (ADMP) [9] distance. For each cluster a threshold value is also calculated as follows:

Threshold (T) = (Average ADMP) x 1.5
Threshold is the lowest possible input value of similarity required to join two objects in one cluster.
The value of T is calculated as the average of all ADMP values of the same cluster multiplied by 1.5.
Henceforth each ADMP value is compared with the corresponding value of threshold of its cluster. If the ADMP value is less than the threshold value the object is retained in the cluster, but, if the ADMP value is greater than the threshold value the object is detected as outlier.
The basic structure of the proposed method is as follows:
Step-1: Perform PAM clustering algorithm to produce a set of k clusters.
Step-2: Determine small clusters and consider the points (objects) that belong to these clusters as outliers.
For the rest of the clusters not determined in Step 2:
Begin
Step-3: For each cluster, compute the ADMP and T values.
Step-4: For each object in cluster, if ADMP > T then classify that object as an outlier; otherwise not.
End

IV. RESULT AND DISCUSSION

In this section, we will investigate the effectiveness of our proposed approach when applied on bench marked data set: data set, presented in [10], with two dimensions.
We start our experimentation with data set, which is shown in figure 4. It is clear that the set contains two outliers, located at the far right of the data set inside the circle as depicted in figure 4.

When the specified number of clusters is *three*, these two outliers are detected from Steps 2 in our approach, since they form a small cluster.

However, when the number of clusters is specified as ***two***, our approach detects these two outliers by applying Steps 3-4.

Table 1 shows the data objects in two dimension i.e x and y coordinates. ADMP

values are generated from Step-3.

It is clear from table 1 that the last two shaded objects have the greatest ADMP values. Applying Step-4, with a calculated threshold T=2.47, we found that the last two points were outliers.
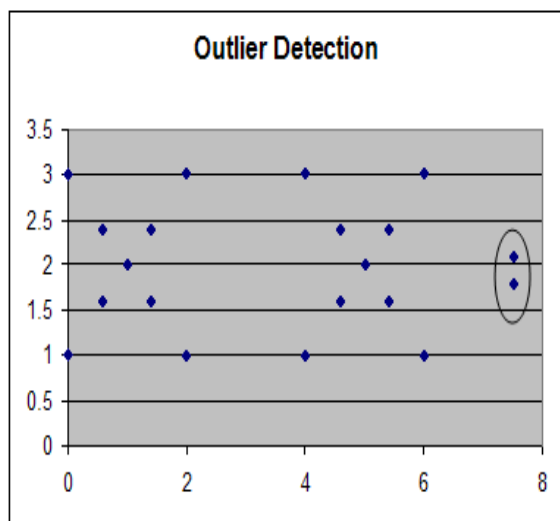

Fig. 4: Data set containing 20 data objects

This means that our proposed approach gives the effective result for outlier detection.

X	Y	ADMP	Cluster
0	1	2	1
0	3	2	1
0.6	1.6	0.8	1
0.6	2.4	0.8	1
1	2	0	1
1.4	1.6	0.8	1
1.4	2.4	0.8	1
2	1	2	1
2	3	2	1
4	1	2	2
4	3	2	2
4.6	1.6	0.8	2
4.6	2.4	0.8	2
5	2	0	2
5.4	1.6	0.8	2
5.4	2.4	0.8	2
6	1	2	2
6	3	2	2
7.5	2.1	2.6	2
7.5	1.8	2.7	2

Table 1: X, Y coordinates of each point, ADMP of each point and each cluster points belongs to

V. APPLICATIONS OF OUTLIER DETECTION

Fraud detection
- Purchasing behavior of a credit card owner usually changes when the card is stolen.
- Abnormal buying patterns can characterize credit card abuse.

Medicine
- Unusual symptoms or test results may indicate potential health problems of a patient.
- Whether a particular test result is abnormal may depend on other characteristics of the patients e.g. gender, age etc.

Public health
- The occurrence of a particular disease, e.g. tetanus, scattered across various hospitals of a city indicate problems with the corresponding vaccination program in that city.
- Whether an occurrence is abnormal depends on different aspects like frequency, spatial correlation, etc.

Sports statistics
- In many sports, various parameters are recorded for players in order to evaluate the players' performances
- Outstanding in a positive as well as a negative sense, players may be identified as having abnormal parameter values.
- Sometimes, players show abnormal values only on a subset or a special combination of the recorded parameters.

Detecting measurement errors
- Data derived from sensors e.g. given in scientific experiment may contain measurement errors.
- Abnormal values could provide an indication of a measurement error
- Removing such errors can be important in other data mining and data analysis tasks
- One person's noise could be another person's signal.

VI. CONCLUSION

We first perform the PAM clustering algorithm. Small clusters are then determined and considered as outlier clusters. The rest of outliers are then found if any in the remaining clusters based on calculating the absolute distances between the medoid of the current cluster and each of the points in the same cluster.
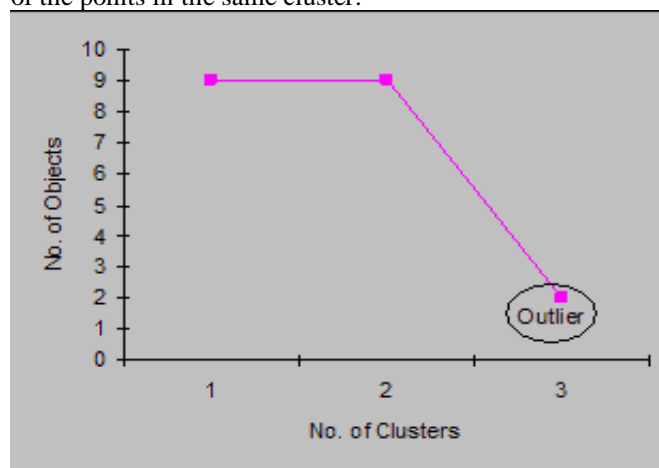

Fig. 5: Graph showing no. of objects in each cluster with outlier

The test results show that the proposed approach gave effective results when applied to a data set because like k-means, PAM is not sensitive to outliers.

REFERENCES

1. J. Queen Mac, "Some methods for classification and analysis of multivariate observations", Proc. 5th Berkeley Symp. Math. Stat. and Prob, pp. 281-97, 1967.
2. J. Zhang and H. Wang, "Detecting outlying subspaces for high-dimensional data: the new Task, Algorithms, and Performance", Knowledge and Information Systems, vol. 10, no. 3, pp. 333-355, 2006.
3. K. Yoon, O. Kwon and D. Bae, "An approach to Outlier Detection of Software Measurement Data using the K-means Clustering Method", First International Symposium on Empirical Software Engineering and Measurement (ESEM 2007), Madrid, pp. 443-445, 2007.
4. T. Velmurugan and T. Santhanam, "A Survey of Partition based Clustering Algorithms in Data Mining: An Experimental Approach", Information Technology, Journal, vol. 10, no. 3, pp. 478-484, 2011.
5. M. Laan, K. Pollard and J. Bryan, "A New Partitioning Around Medoids Algorithms", Journal of Statistical Computation and Simulation, vol. 73, no. 8, pp. 575-584, 2003.
6. J. Han and M. Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann, 2nd Ed., 2006.
7. P. Bradley, U. Fayyad and O. Mangasarian, "Mathematical Programming for Data Mining: Formulations and Challenges", INFORMS Journal on Computing vol. 11, no. 3, pp. 217–238, 1999.
8. F. Angiulli and C. Pizzuti, "Outlier Mining in Large High-Dimensional Data Sets", IEEE Trans. on Knowledge and Data Engineering, vol. 17, no. 2, pp. 203-215, 2005.
9. K. Niu, C. Huang, S. Zhang, and J. Chen, T. Washio et al. "ODDC: Outlier Detection Using Distance Distribution Clustering", (PAKDD 2007) Workshops, Lecture Notes in Artificial Intelligence (LNAI), pp. 332–343, Springer-Verlag, 2007.
10. J. Chiang and Z. Yin, "Unsupervised minor prototype detection using an adaptive population partitioning algorithm", Pattern Recognition, pp. 3132 – 3145, 2007.

AUTHOR PROFILE

Asst. Prof. Vijay Kumar: Born in 1980 at Jaunpur (U.P.) He had done M.Tech (Computer Science & Engg.) from GGSIP University, Delhi. Work Experience: He had taught in UIT, Almora (U.A.), presently he is teaching in MIET, Meerut as an Asst. Professor. He has published several research papers in National and International Journals. He is pursuing Ph.D in Computer Science & Engineering and his area of research is Data Mining.

Associate Prof. Sunil Kumar: is a faculty in a self-financing engineering college in Meerut (U.P.). He had done M. Tech (Computer Science & Engineering) from Maharshi Dayanand University, Rohtak (Haryana). He is currently doing his PhD (Computer Science & Engineering) from Punjab Technical University, Jalandhar (Punjab). He has published several research papers in national and international journals.

Prof. (Dr.) Ajay Kumar Singh: Born in 1974 at Dhanbad (Jharkhand). He had done B.E (Computer Science & Engg.) from Kumaon Engineering College, M. Tech (I.T) Allahabad, Ph. D (Computer Science & Engg.) Jaypee University of Information Technology. Work Experience: He had been in different institution / university like Radha Govind Engineering College, Meerut, (U.P), Sir Padampat Singhania University, Bhatewar, Udaipur, Rajasthan, Jaypee University of Information Technology, Waknaghat, Solan (H.P), Mody College of Engineering and Technology, Lakshmangarh, Sikar, Rajasthan, Regional Engineering College (Now N.I.T.) Kurukshetra (Haryana), Software Solution Integrated Ltd. (Delhi), Computer Centre CMC. Now he is working with MIET, Meerut, U.P. He has published 12 papers in international Journals like PIER, Asia Magazine EFY Elsevier, JSIP, 8 papers in international Conference out of which 4 of them in IEEE, Published 4 papers in National Conferences, 1 in EFY and presented his papers at Bangalore, IT B.H.U, USA (Washington DC), etc.