Volume 2, Issue 4, April 2012 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A New Hybridized K-Means Clustering Based Outlier Detection Technique For Effective Data Mining

H.S.Behera
Computer Science and Engineering,
VSSUT, Burla Odisha, India
hsbehera_india@yahoo.com

Abhishek Ghosh.
Computer Science and Engineering,
VSSUT, Burla, Odisha, India.
abhishekgh3@gmail.com

Sipak ku. Mishra
Computer Science and Engineering
VSSUT, Burla, Odisha, India.
sipakmishra208@gmail.com

Abstract- **Nowadays million of databases have been used in business management, Govt., scientific engineering & in many other application & keeps growing rapidly in present day scenario. The explosive growth in data & database has generated an urgent need to develop new technique to remove outliers for effective data mining. In this paper we have suggested a clustering based outlier detection algorithm for effective data mining which uses k-means clustering algorithm to cluster the data sets and outlier finding technique (OFT) to find out outlier on the basis of density based and distance based outlier finding technique.**

Keywords:- **K-means clustering algorithm. Density based outlier detection, distance based outlier detection, Outlier Detection Technique (OFT).**

I. INTRODUCTION

Outlier detection is a fundamental issue in data mining; specifically it has been used to detect and remove anomalous objects from data. Outliers, also called contaminant observations, are data points that deviate from other data points that they seem to be generated because of any of the faulty condition in the experimental setup.

When observations that are taken in the experimental setup are subjected to analysis there is a possibility of the two conditions based on the outlier. Either (i) outliers negatively influence the results of analysis, or (ii) the search for outliers is the main task of data analysis. In data mining outlier detection is also regarded as the detection of anomaly. In many applications, a set of training values is required to define 'normality'. Security applications are examples, in which atypical behavior by people or technical systems has to be detected. Outlier tests are used with the statistical model for generating the observations and presume some knowledge of the number of assumed outliers. Many of them can only cope with a single outlier.

Outlier detection is used in various domains in data mining. This has resulted in a huge and highly diverse literature of outlier detection techniques. A lot of these techniques have been developed in order to solve problems based on some of the particular features, while others have been developed in a more generic fashion.

It has been argued by many researchers whether clustering algorithms are an appropriate choice for outlier detection. For example, in (Zhang and Wang, 2006) [2], the authors reported that clustering algorithms should not be considered as outlier detection methods. This might be true for some of the clustering algorithms, such as the k-means clustering algorithm (MacQueen, 1967) [3]. This is because the cluster means produced by the k-means algorithm is (Laan, 2003) [4].But here we propose an algorithm that uses clustering efficiency of the k-means algorithm and uses a hybridized outlier detection technique of finding outlier through clustering.

II. RELATED WORK.

Partitioning Around Medoids (PAM) clustering technique has been used for clustering the data and the clustered data is used for outlier detection as discussed in Moh'd Belal Al-Zoubi [1].Various clustering algorithm's effectiveness was discussed in (Zhang and Wang, 2006) [2]. It was discussed why we should not use clustering based techniques for outlier detection.

In efficiency of k-means clustering algorithm for outlier analysis is highlighted in (MacQueen, 1967) [3] & (Laan, 2003) [4]. Here it has been discussed about the sensitivity of the k-means towards the noise and outliers present in data set.

Distance based approach for outlier detection has been discussed in (Knorr, 1998) [5] & (Knorr, 2000) [6].Density based approaches for outlier detection has been discussed in (Breunig, 2000) [7] & (Papadimitriou, 2003) [8].

Clustering based approaches for outlier analysis has been discussed in (Loureiro 2004) [9] & (Gath and Geva 1989) [10].Distance based outlier detection has been discussed by V. Hautamäki, I. Kärkkäinen and P. Fränti (2004) [11].

Data sets for outlier detection is given by Chiang, J. and Z. Yin [12] & , C. L. & C. J. Merz, (1998) [13].

III. METHODOLOGIES

A. K-Means Clustering Algorithm.

K-means is a prototype-based, simple partition clustering technique which attempts to find a user-specified k number of clusters. These clusters are represented by their centroids. A cluster centroid is typically the mean of the points in the cluster. This algorithm is simple to implement and run, relatively fast, easy to adapt, and common in practice.

This algorithm consist of two separate phases: the first phase is to select k centres randomly, where the value of k is fixed in advance. The next phase is to assign each data object to the nearest centre. Euclidean distance is generally considered to determine the distance between each data object and the cluster centres. When all the data objects are included in some clusters, recalculation is done on the average of the clusters.

This iterative process continues repeatedly until the criterion function become minimum.

The k-means algorithm works as follows:

Step 1: Randomly select k data object from dataset D as initial cluster centers.

Step 2: Repeat step 3 to step 5 till no new cluster centers are found.

Step 3: Calculate the distance between each data object d_i(1<=i<=n) and all k cluster centers c_j(1<=j<=n) and assign data object di to the nearest cluster.

Step 4: For each cluster j(1<=j<=k), recalculate the cluster center.

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst to decide what will be considered abnormal. Before abnormal observations can be singled out, it is necessary to characterize normal observations.

The definition can be narrowed to outlier detection can be thought of detection of anomalous data patterns in the data sets by pre learned techniques of the data sets got from the observation of different experimental condition. An outlier will refer to these anomalous patterns in the data. The output of an outlier detection technique could be labeled patterns. Some of the outlier detection techniques also assign a score to a pattern based on the degree to which the pattern is considered an outlier. Such a score is referred to as outlier score.

Outlier detection can be broadly classified into three groups. First is the distance based outlier detection, it detects the outlier from the neighborhood points. Second is the density based outlier detection, here it detects the local outlier from the neighborhood based on the density or the no. of data points in the local neighborhood. Third is the distribution based outlier detection, this approach is based on the finding outlier using some statistical model.

Density based outlier detection

Density-based methods have been developed for finding outliers in a spatial data. These methods can be grouped into two categories called multi-dimensional metric space-based methods and graph-based methods. In the first category, the definition of spatial neighborhood is based on Euclidean distance, while in graph-based spatial outlier detections the definition is based on graph connectivity.

Distance based outlier detection.

In Distance-based methods outlier is defined as an object that is at least d_{min} distance away
from k percentage of objects in the dataset. The problem is then finding appropriate d_{min} and k such that outliers would be correctly detected with a small number of false detections. This process usually needs domain knowledge [11].

C. Outlier Finding Technique (OFT).

Outlier Finding Technique (OFT) is a hybridized form of both distance based and density based outlier finding technique.

Here after cluster formation has taken place with the help of k-means clustering then we are left with the cluster of data points and the cluster center.

Let us denote cluster center as C_k, where k is the no of cluster. We use here a special center called as weight based center let us denote it with W. Where W is defined as follows:-

$$W = \frac{\sum_{n=1}^{k}(t_n \times C_n)}{T} \qquad (1)$$

Where t_n is the total no of data points in the cluster C_n (n= 1,2….k). and T is the total no of data points given.

Based on the weight based center W we can find the cluster that is not closely related to any neighborhood of the data points, thus we can conclude that this cluster is the outlier.

The algorithm can be stated as follows:-

Step 1: Calculate the weight based center W as given in the equation (1).

Step 2: Calculate the distance of each cluster center C_k to the weight based center W, let it be D_k.

Step 3: Arrange the value of D_k in decreasing order.

Step 4: Calculate the no of data points (t_k) in the cluster

Step 5: Calculate the maximum no. of data points (t_{max}) of the k- cluster, where $t_{max} = \max(t_1,t_2,\ldots t_k)$.

Step 6: Calculate the minimum no. of data points (t_{min}) of the k- cluster, where $t_{min} = \min(t_1,t_2,\ldots t_k)$.

Step 7: $t_{critical} = \left|\frac{t_{max} + t_{min}}{2}\right|$

Step 8: Repeat the steps 9-10 till there is no outlier found or there is no cluster left.

Step 9: For each of the value of D_n n= 1 to k compare the value to its t_n to $t_{critical}$.

Step 10: If the value is found less than $t_{critical}$ then the given cluster C_n is the outlier.

IV. PROPOSED ALGORITHM

Input: Data set $D=\{d_1,d_2.....d_n\}$,where d_i=data points, n= no of data points

Cluster centre $C=\{c_1,c_2,......c_k)$, where c_i=cluster centre ,k = no of cluster centres.

Output: Cluster C_n (n=1,2….k) outlier cluster.

Step 1: Calculate the distance of each data points d_n and the k cluster centers c_k mostly preferred is the eucledian distance.

Step 2: For each data object d_i, find the closest centroid c_j and assign d_i to the cluster with nearest centroid cj..

Step 3: Repeat the following Steps 4-5 till a convergence criteria is met or we can say no new centroids are found.

Step 4: For each data points d_i compute its distance from the centroid c_j of the present nearest cluster.

Step 5: If the calculated distance is less than or equal to the previous calculated distance then the data points stay in the previous cluster.

Step 6: Else, calculate the distance of the data point to each of the new cluster centers and assign the data point to the nearest cluster based on the distances from the cluster centers.

Step 7: Calculate the weight based centre W as given in the equation (1).

Step 8: Calculate the distance of each cluster centre c_k to the weight based centre W, let it be D_k.

Step 9: Arrange the value of D_k in decreasing order.

Step 10: Calculate the no of data points (t_k) in the cluster.

Step 11: Calculate the maximum no. of data points (t_{max}) of the k- cluster, where $t_{max} = \max(t_1,t_2,\ldots t_k)$.

Step 12: Calculate the minimum no. of data points (t_{min}) of the k- cluster, where $t_{min} = \min(t_1,t_2,\ldots t_k)$.

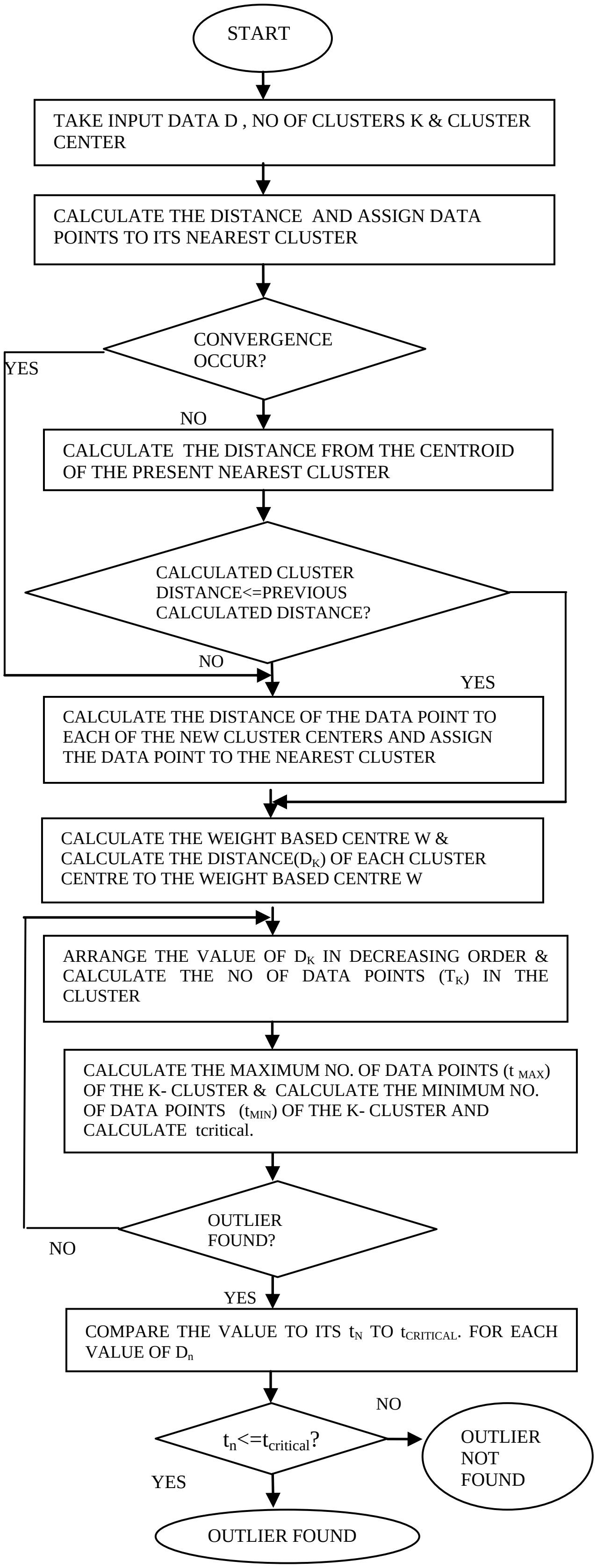
Step 13: $t_{critical} = \frac{t_{ma\ x} + t_{min}}{2}$

Step 14: Repeat the steps 9-10 till there is no outlier found or there is no cluster left.

Step 15: For each of the value of D_n n= 1 to k compare the value to its t_n to $t_{critical}$.

Step 16: If the value is found less than $t_{critical}$ then the given cluster C_n is the outlier.

V. FLOW CHART OF THE PROPOSED ALGORITHM


START
TAKE INPUT DATA D , NO OF CLUSTERS K & CLUSTER CENTER
CALCULATE THE DISTANCE AND ASSIGN DATA POINTS TO ITS NEAREST CLUSTER
CONVERGENCE OCCUR?
YES
NO
CALCULATE THE DISTANCE FROM THE CENTROID OF THE PRESENT NEAREST CLUSTER
CALCULATED CLUSTER DISTANCE<=PREVIOUS CALCULATED DISTANCE?
NO
YES
CALCULATE THE DISTANCE OF THE DATA POINT TO EACH OF THE NEW CLUSTER CENTERS AND ASSIGN THE DATA POINT TO THE NEAREST CLUSTER
CALCULATE THE WEIGHT BASED CENTRE W & CALCULATE THE DISTANCE(D_K) OF EACH CLUSTER CENTRE TO THE WEIGHT BASED CENTRE W
ARRANGE THE VALUE OF D_K IN DECREASING ORDER & CALCULATE THE NO OF DATA POINTS (T_K) IN THE CLUSTER
CALCULATE THE MAXIMUM NO. OF DATA POINTS (t_{MAX}) OF THE K- CLUSTER & CALCULATE THE MINIMUM NO. OF DATA POINTS (t_{MIN}) OF THE K- CLUSTER AND CALCULATE tcritical.
OUTLIER FOUND?
NO
YES
COMPARE THE VALUE TO ITS t_N TO $t_{CRITICAL}$. FOR EACH VALUE OF D_n
$t_n<=t_{critical}$?
NO
YES
OUTLIER NOT FOUND
OUTLIER FOUND

VI. EXPERIMENTAL ANALYSIS.

The effectiveness of our proposed algorithm is analyzed on data set1 (Chiang and Yin, 2007) [12], it contain datasets having two dimensions. Then the algorithm is analyzed on data sets of higher dimension, here we have taken Bupa data set (Blake and Merz, 1998) [13]. The data sets are taken for analysis in Moh'd Belal Al- Zoubi [1]. Eventually we can see that using k-means with OFT gives more outlier detection in case of higher dimension data than in outlier detection using PAM by Moh'd Belal Al- Zoubi [1].

First outlier detection is analysed in data set1. The data sets are represented graphically in figure 1. Here three cluster centres are taken.
The first cluster consists of nine data points, second cluster consist of ten data points, and third cluster consists of two data points.

Table 2 gives the cluster data points. Here after the application of k-means to the data points of data set1 cluster formation is done and data .

Weight based centre (W) = 4.75 calculated from (1). Distances of each cluster centre from the W arranged in the decreasing order we get cluster3, cluster2, cluster1. $T_{critical}$ = 6, so based on OFT we find that cluster3 is the outlier cluster.

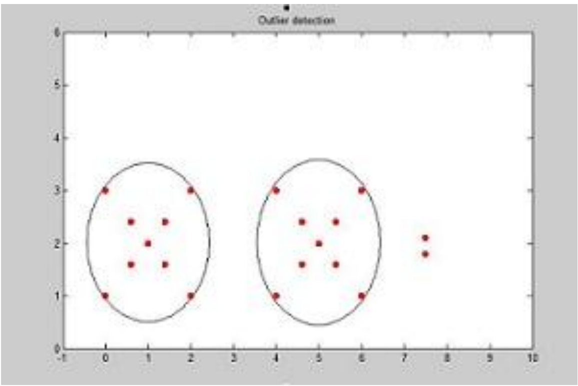


Fig 1- coordinate representation of the data points of data set1.

Table 1 gives the coordinate axes of the data points. Points belonging to the respective cluster are recorded in table

.

TABLE I- coordinate axes of the data points of the data set1

Data points	X	Y
1	0	1
2	0	3
3	0.6	1.6
4	0.6	2.4
5	1	2
6	1.4	1.6
7	1.4	2.4
8	2	1
9	2	3
10	4	1
11	4	3
12	4.6	1.6
13	4.6	2.4
14	5	2
15	5.4	1.6
16	5.4	2.4
17	6	1
18	6	3
19	7.5	2.1
20	7.5	1.8

The third data set is the Bupa data set. This data set has 2 classes and 6 dimensions. Based on PAM clustering in [1] 19 outliers were found in class1 and 17 outliers were found in class2 but our proposed algorithm finds 19 outliers were found in class1 and 20 outliers in class2. It is shown in Table 3.

TABLE II- data points assigned to respective cluster using k-means.

Data points	Cluster	Cluster centre
1	1	
2	1	
3	1	
4	1	
5	1	1
6	1	
7	1	
8	1	
9	1	
10	2	
11	2	
12	2	
13	2	
14	2	2
15	2	
16	2	
17	2	
18	2	
19	3	3
20	3	

TABLE III-Conins the outlier detection of the Bupa data set.

Class 1	Detected by [1]	Detected	Class 2	Detected by [1]	Detected
20	No	No	2	No	No
22	No	No	36	Yes	Yes
25	Yes	Yes	77	Yes	Yes
148	Yes	Yes	85	Yes	Yes
167	Yes	No	111	No	Yes
168	Yes	Yes	115	Yes	Yes
175	Yes	Yes	123	Yes	Yes
182	Yes	Yes	133	Yes	Yes
183	Yes	Yes	134	Yes	Yes
189	Yes	Yes	139	Yes	Yes
190	Yes	Yes	157	No	No
205	Yes	Yes	179	Yes	Yes
261	Yes	Yes	186	Yes	Yes
311	Yes	Yes	187	Yes	Yes
312	Yes	Yes	224	No	Yes
313	No	Yes	233	No	No
316	Yes	Yes	252	No	No
317	Yes	Yes	278	Yes	No
326	Yes	Yes	286	Yes	Yes
335	Yes	Yes	294	Yes	Yes
343	Yes	Yes	300	No	Yes
345	Yes	Yes	307	No	Yes
			323	Yes	Yes
			331	Yes	Yes
			337	No	No
			342	No	Yes

Here the yes symbolizes the outlier detection and no symbolizes that it cannot detect outlier.

The figure 2 depicts the performance of the proposed algorithm for outlier detection.

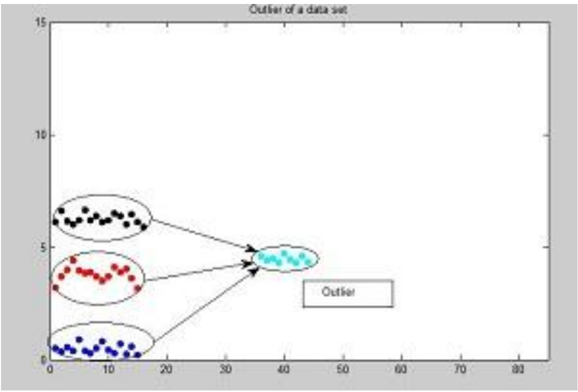


Fig 2- outlier detection using the proposed algorithm

VII. CONCLUSION

From the experimental analysis of data sets both of lower dimension and higher dimension as in the case of Bupa dataset we can see that k-means can be used for outlier analysis.

The outlier detection of the proposed algorithm in the Bupa data set has improved over the algorithm used by Moh'd Belal Al- Zoubi [1].

K-means has sensitivity over outlier data but can be still used with OFT for the detection of outlier data.

Many of the clustering techniques are being developed that are not affected by outliers and can be easily implemented to find outlier. K-means can be improved over noisy data that can be used to find outlier detection.

VIII. REFERENCES

[1] "An Effective Clustering-Based Approach for Outlier Detection" by Moh'd Belal Al- Zoubi European Journal of Scientific Research Vol.28 No.2 (2009).

[2] "Detecting outlying subspaces for high-dimensional data: the new Task, Algorithms, and Performance, Knowledge and Information Systems", Zhang, J. and H. Wang, 2006.10(3): 333-355.

[3]"Some methods for classification and analysis of multivariate observations" by MacQueen, J.,1967. Proc. 5th Berkeley Symp. Math. Stat. and Prob, pp. 281-97.

[4]"A New Partitioning Around Medoids Algorithms" by Laan, M., K. Pollard and J. Bryan, 2003. Journal of Statistical Computation and Simulation, 73(8): 575-584.

[5] "Algorithms for Mining Distance-based Outliers in Large Data Sets" by Knorr, E. and R. Ng, 1998. Proc. the 24th International Conference on Very Large Databases (VLDB), pp. 392-403.

[6]" Distance-based Outliers: Algorithms and Applications" by Knorr, E., R. Ng, and V. Tucakov, 2000 VLDB Journal, 8(3-4): 237-253.

[7] "Lof: identifying density-based local Outliers" by Breunig, M., H. Kriegel, R. Ng and J. Sander, Proceedings of 2000 ACM SIGMOD International Conference on Management of Data. ACM Press, 93–104.

[8]" LOCI: Fast outlier detection using the local correlation integral" by Papadimitriou, S., H. Kitawaga, P. Gibbons, and C. Faloutsos, 2003 Proc. of the International Conference on Data Engineering, pp. 315-326.

[9] "Outlier Detection using Clustering Methods: a Data Cleaning Application" by Loureiro,A., L. Torgo and C. Soares, 2004 in Proceedings of KDNet Symposium on Knowledge-based Systems for the Public Sector. Bonn, Germany.

[10]" Fuzzy Clustering for the Estimation of the Parameters of the Components of Mixtures of Normal Distribution, Pattern Recognition Letters" by Gath, I and A. Geva, 1989

[11] "Outlier Detection Using k-Nearest Neighbor Graph" by V. Hautamäki, I. Kärkkäinen and P. Fränti,In Proceedings of the International Conference on Pattern Recognition, Volume 3 pages 430 – 433, Cambridge, UK, August 2004.

[12] "Unsupervised minor prototype detection using an adaptive population partitioning algorithm, Pattern Recognition" by Chiang, J. and Z. Yin

[13] "UCI Repository of Machine Learning Database" by Blake, C. L. & C. J. Merz, 1998. University of California, Irvine, Department of Information and Computer Sciences.

SHORT BIO DATA OF ALL THE AUTHORS

1. Dr. H.S Behera is currently working as a Faculty in Dept. of Computer Science andEngineering is Veer Surendra Sai University of Technology(VSSUT), Burla, Odisha, India. Hisresearch areas of interest include Operating Systems, Data Mining , Soft Computing and Distributed Systems.

2. Mr. Abhishek Ghosh is a Final year B. Tech. student in Dept. of Computer Science and Engineering, Veer Surendra Sai University of Technology (VSSUT), Burla, Odisha, India.

3. Mr. Sipak ku. Mishra is a Final year B. Tech. student in Dept. of Computer Science and Engineering, Veer Surendra Sai University of Technology (VSSUT), Burla, Odisha, India.