INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

Survey of Partition Based Clustering Algorithm Used for Outlier Detection

Shruti Aggrwal[1], Prabhdip Kaur[2]

[1]AssistantProfessor, [2] Research Scholar
SGGSWU, Fatehgarh Sahib
[1]shruti_cse@sggswu.org,[2]prbhdippandher@ymail.com

Abstract**:** *Data mining is a process of extracting hidden and useful information from the data. The knowledge discovered by data mining is previously unknown, potentially useful, and of high quality. There are large numbers of data mining techniques used to extracting hidden and useful information from the data. Clustering is one of the most important techniques in data mining. Outlier detection is one of most important issue in clustering.*
Outlier represents that data which represent different behavior from others. Therefore, it is important to detect outlier from the extracted data. Outlier detection is a task that finds objects that are dissimilar or inconsistent with the remaining data. Outlier detection is used in many applications like fraud detection, network intrusion detection and clinical diagnosis of diseases. In this paper we describe the large number of Partition based clustering algorithm used for outlier detection and also describe the comparative study of these algorithms so user can choose particular algorithm according their requirement.

Keyword: *Outlier , K-mean, Partition around Mediod Algorithm(Pam), Clustering large application Algorithm(Clara), Clustering Large Application Based on Randomization Search Algorithm (Clarans), Clustering Large Application with Triangular Irregular Network Algorithm (Clatin), ECLARNS (Enhanced CLARANS).*

1. INTRODUCTION

Data mining is a process of extract important and valuable knowledge from large database. Such extraction helps us in decision making. There are large number of technique and algorithm are used to extract hidden pattern in database and finding the between them. Clustering is one of the most important techniques in data mining. Clustering is mainly used to grouping the same data based on their similarity and outlier detection is one of most important issue in clustering [1]. Outlier is a pattern that is dissimilar to another pattern in database [2].outlier detection is a task finding the objects that are dissimilar or inconsistent with remaining object .outlier consider noise or errors which are removed once detected [1]. Different domain has different reason for outlier detection. Outlier detection is used in wide variety of application such as fraud detection, data analysis, network intrusion detection, clinical intrusion detection. [1]

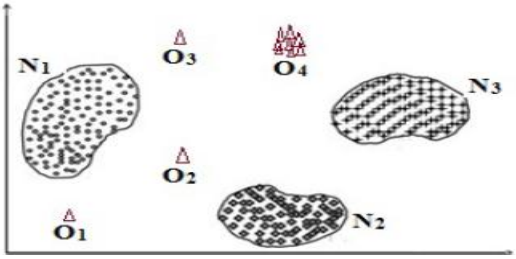


Figure 1: Outlier's detection in two dimensional dataset [15]

Illustrating the outlier in two dimensional dataset. There are N1, N2, and N3 are the three normal regions. Points that are sufficiently far away from the normal region such as points O1, O2, O3 and points in O4 regions are outliers.

2. APPROACHES USED FOR OUTLIER DETECTION

There are several Approaches used for outlier detection. Each of these technique has own advantage and disadvantage. The technique which user used for outlier detection should consider two steps. First identifies an outlier around the dataset using set of inliers. Second step is data request for analyze and identify as outlier when attribute are different from attribute of inliers. [1].

2.1 Distance Based Outlier Detection

Distance-based method was originally proposed by Knorr and Ng. [15] Distance based approach is used to outlier detection according to given threshold value. This is given by user. This technique is used to calculate the maximum distance value for each cluster if the maximum distance of cluster is greater than threshold value then the cluster will we declared as outlier [3]. In distance based outlier detection approach user can use any of metrics like Euclidean distance for measure the distance between points [6] Example of distance based clustering is an object x is marked as an outlier, if there are less than k objects in a distance at most R from x, excluding x itself. According to this definition, to detect distance-based outliers two parameters k and R are Required, to control the d. [18]

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

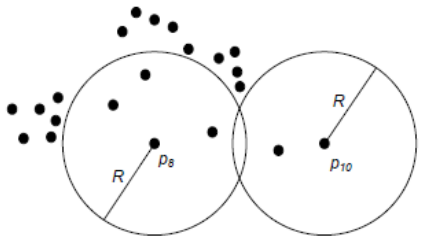


Figure 2.1: An example data set with two distance-based outliers [18]

Figure 2.1 depicts an example. If k = 4 and the parameter R is set to a fixed value, then an object x is marked as an outlier if there are less than four objects in a distance at most R from x (excluding x itself). It is not hard to check that objects p8 and p10 are outliers based on the values of k and R.

2.2 Clustering Based Outlier Detection

Clustering based approach is used when the number of normal attribute is more than abnormal behavior attribute .this technique provides more positive result. This approach is used in those situation when large and dense cluster have normal data and data which does not belong to any cluster or small cluster (low dense cluster) are consider as outlier. [4] In cluster based approach normal data records belong to large and dense clusters, while outliers do not belong to any of the clusters or form very small clusters

- Cluster the data into groups of different density
- Choose points in small cluster as candidate outliers
- Compute the distance between candidate points and non- candidate clusters.
- If candidate points are far from all other non-candidate points, they are outliers

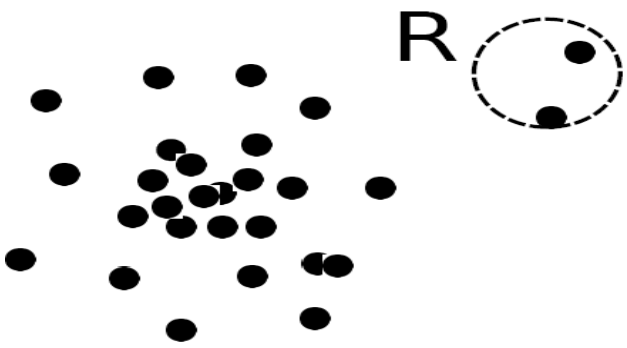


Figure 2.2: Cluster Based Outlier Detection approaches [16]

The objects in region *R* are outliers. All points not in R form a large cluster the two points in R form a tiny cluster, thus are outliers

2.3 Density Based Outlier Detection

Density Based approach method involve the investigation not only local density but also studied local density of its nearest neighbors [5]. This method identify the outlier by checking the main features or characteristics of object in database the object that are deviate from these feature are consider as outlier. [8] In this method the nearest neighbors are relatively close, then the data point is considered to be normal, otherwise it is considered to be an outlier. In this method we compute local outlier factor (LOF) of a sample p as the average ratios of the density of sample p and the density of its nearest neighbors [17].

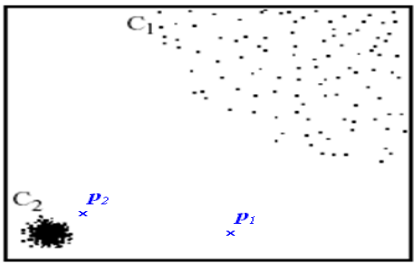


Figure 2.3 Density based outlier detection approach [17]

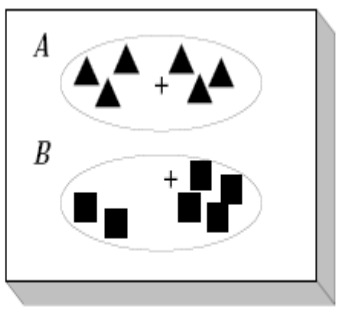
According to nearest neighbors approach, p_2 is not considered as outlier, but according to local outlier factor approach find both p_1 and p_2 consider as outliers

2.4 Distribution Based Outlier Detection

It Develop statistical models from the given data and then apply a statistical test to determine if an object belongs to this model or not. Objects which have low probability to belong to the statistical model are declared as outliers. However, Distribution-based approaches cannot be applied in multidimensional scenarios because they are univariate in nature. Most distribution models typically apply directly to the future space and are univariate i.e. having very few degrees of freedom. Thus, they are unsuitable even for moderately high-dimensional data sets. [14] In this approach, a prior knowledge of the data distribution is required, making the distribution-based approaches difficult to be used in practical applications. [3]

3. PARTITIONS BASED CLUSTERING ALGORITHM USED FOR OUTLIER DETECTION

Partition based clustering create k partition of data set with n data object. It is an iterative relocation technique is used to improve the clustering by moving up the object from one group to another. Partition based clustering is represent by centriod or mediod. [7] They use iterative way to produce the clustering. One of the disadvantages of partition based clustering is their high complexity. Even when there are a small number of objects the partition is Huge. [8]



Clustering with k=2

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

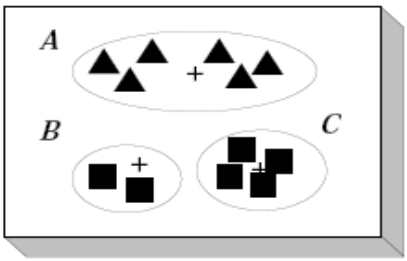


Figure 3: Partitions Based Clustering process. [9]

According to above example of two divisive algorithms performed in the same data set, with different initial parameters. ''+'' sign denotes the centre of clusters, which in this figure is defined as the mean of the values of a particular cluster. [9] The Most Efficient Algorithms Proposed under Partition Based Method used for outlier detection is k-Men.

3.1 K-mean

K-mean describes that given dataset of n object divide into k cluster where k is desired number of cluster. A centriod is defined for each cluster in k-mean all data object are placed in cluster having centriod nearest to all data object. After processing all data object then k-mean centriod is calculated again and again. In each iteration centriod change their location. This process continues step by step until no centriod move. Need to specify k number of cluster in advance. It is unable handle the noise or outlier or handle the cluster very different shapes [7]The complexity of k-mean clustering is O (IKN) where I denote number of iteration and k<<n

Steps of k-mean

- Select the k object as initially center
- Assign each data object to cluster center
- Recalculate the center of each cluster
- Repeat step 2 and 3 until cluster center don't change [4]

This method develops the statical model from given dataset.

3.1.2 PAM (Partition around Mediod)

PAM is developed by Kaufman and Rousseuw in 1987. The algorithm chooses k-mediod initially and then swaps the mediod object with non mediod as a result quality of cluster is improved. It is very robust when compare with k-mean in the presence of noise or outlier. [4] Algorithm work well with small dataset but does not work well with large dataset. [3] The computational complexity of PAM is O (IK (N-K) ^{2}) where I is a number of iteration.

Procedure of PAM

- Input dataset d
- Randomly select K object from dataset G
- Calculate total cost T for each pair of selected Si and non selected sh.
- For each pair if T Si<_0 then it is replaced by SK
- Then find similar mediod for each non selected object
- Repeat the step 2, 3, 4 until find the mediod. [4].

3.1.3 CLARA (Clustering Large Application)

CLARA is developed by Kaufman & Rousseuw in 1990.CLARA algorithm work well with several sample size of N touple in dataset. Then we apply the PAM each sample. [11]It can identify outlier and select the best mediod as output.[4]. This method takes sample of data from dataset instead of taking the full dataset. It randomly selects the data then chooses the mediod using the PAM algorithm. [1]The computational complexity of CLARA is O (K (40+K) ^{2}+k (Ni)). [11] It will not necessarily represent a good clustering whole data set if the sample is biased and in CLARA and if "true" mediod of the initial data are not contained in the sample, then the result is guaranteed not to be the best result.

Procedure of CLARA

- Input the data set D
- Repeat N time
- Draw sample S randomly from D
- Call PAM to get mediod
- Classify entire data set to cost1…cost k.
- Calculate the average dissimilarity from obtained cluster.[3]

3.1.4 CLARANS (Clustering Large Application Based Upon Randomized Search)

Ng and Han a new algorithm in 1994 called CLARANS. It use random search to generate neighbors by starting with arbitrary node and randomly check max-neighbors. If the neighbor represent better partition the process continue with new node otherwise local minimum is found and algorithm restart until numlocal local minima is found (value of numlocal is=2 recommended)the best node return resulting partition.[12]CLARANS take a random dynamic selection of data at each step of process. Thus the same sample set is not used throughout in the clustering process. As a result better randomization source is achieved. [13] CLARANS is accurately detecting outlier than CLARA and it is much less affected by increasing dimensionally and draw the sample of neighbors in each step of search this is benefit of confining the search localize area.

Procedure of CLARANS

- Randomly choose k mediod
- Randomly consider the one of mediod swapped with non mediod
- If the cost of new configuration is lower repeat step 2 with new solution
- If the cost higher repeat step 2 with different non mediod object unless limit has been reached
- Compare the solution keep the best
- Return step 1 unless limit has been reached (set to the value of 2).[8]

3.1.5 CLATIN (Clustering Large Application with Triangular Irregular Network)

When user is worked with the PAM algorithm then the replacement of current mediod O_i and non mediod object O_h that will be effect only small portion of dataset. Therefore in

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

step 2 of PAM clustering algorithm to calculate the total cost of replacement of mediod object o_i and non mediod object o_h we only need to account the small portion then there is a question in mind how to determine the efficiency of this subset of original dataset. Then we use the triangular network of the mediod to help us find subset. This lead helps us to generate new efficient k-mediod algorithm we call this algorithm clustering large application with triangular irregular network. [1]

Procedure of CLATIN

- Initialization.

1) Select k representative objects arbitrarily as initial medoids.

2) Construct the TIN of these k medoids.

- Compute total cost TCih for all pairs of objects Oi, Oh where Oi is currently selected mediod, and Oh is one of the non-mediod objects.

1) Determine the affected object subset S through a link analysis in the medoids-TIN.

2) Calculate the total cost TCih over the neighboring object subset S.

- Select the pair Oi, Oh which related to min Oi, Oh (TCih). If the minimum TCih is negative,

1) Replace Oi with Oh,

2) update the TIN and clustering results locally,

3) Go back to Step 2.

- Otherwise, for each non-selected object, find the most similar representative object. [1]
-

3.1.6 ECLARNS (Enhanced CLARANS)

This method is different from PAM, CLARA, and CLARANS .This method is improvement of CLARANS instead of selecting random searching operation. This method make a cluster by selecting proper arbitrary node .this method is very similar to CLARANS but this method selected arbitrary node reduced the number of iteration in CLARANS. [1]

Procedure of ECLARANS

- Input parameters numlocal and maxneighbour. Initialize i to 1, and mincost to a large number.
- Calculating distance between each data points
- Choose n maximum distance data points
- Set current to an arbitrary node in n: k
- Set j to 1.
- Consider a random neighbor S of current, and based on 6, calculate the cost differential of the two nodes.
- If S has a lower cost, set current to S
- Otherwise, increment j by 1. If j max neighbour, go to Step 6.
- Otherwise, when j > maxneighbour, compare the cost of current with mincost. If the former is less than mincost, set mincost to the cost of current and set best node to current.
- Increment I by 1. If I > numlocal, output best node. Otherwise, go to Step 4. [3]

4. COMPARATIVE STUDY

Outlier detection is challenging task of data mining than .Large number of Partition based clustering algorithm is proposed tell now for outlier detection. They can be used to solve all problems. But all algorithms are designed under certain assumption and different algorithm is used under different condition. Such as k-mean is used to handle spherical shaped cluster we cannot used to find arbitrary shaped cluster

A comparative study of different partition based clustering algorithm used for outlier detection proposed under this method. So user can choose particular algorithm according their requirement.

Table 4: Describe the Comparative Study of Various Partitions Based Clustering Algorithm Used for Outlier Detection

Sr. No	Name	Cluster shape	Complexity	Remarks
1	k-mean	Spherical	O(IKn)	+ease of implementation, efficiency, -not suitable for clusters of nonconvexshapes or different size, sensitive to noise
2	PAM	Arbitrary	$O(IK(n-k)^2$	+more robust than k-means in presence of noise -processing is more costly than k-means
3	CLARA	Arbitrary	$O(ks^2+k(n-k))$ s is a size of sample	- effectiveness depends on sample selection

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

4	CLARANS	Arbitrary	$O(n^{2})$	+ more effective than PAM & CLARA, Insensitivity to noise is partially, - does not handle high dimensional data
5	CLATIN	Arbitrary	O(n log n)	+ clustering large application with triangular irregular network
6	ECLARANS	Arbitrary		+ improvement of CLARANS,ECLARANS selected arbitrary nodes reduce the number of iterations of CLARANS

5. CONCLUSION

There are large number of Partition based outlier detection technique are available. They can be used to solve all problems. But all algorithms are designed under certain assumption and different algorithm is used under different condition. Such as k-mean is used to handle spherical shaped cluster we cannot used to find arbitrary shaped cluster. The main aim of clustering algorithm which is used for outlier detection improves the time complexity and outlier efficiency. Additionally, the efficiency and effectiveness of a novel outlier detection algorithm can be defined as to handle large volume of data as well as high-dimensional features with acceptable time and storage, to detect outliers in different density regions, to show good data visualization and provide users with results that can simplify further analysis.

REFERENCES

[1] **Sivaram, Saveetha,"AN Effective Algorithm for Outlier Detection", Global Journal of Advanced Engineering Technologies, Volume 2, pp 35-40, January 2013.**

[2] **Ms. S. D. Pachgade, Ms. S. S. Dhande," Outlier Detection over Data Set Using Cluster-Based and Distance-Based Approach", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, pp 12-16 June 2012.**

[3] **Deepak Soni, Naveen Jha, Deepak Sinwar," Discovery of Outlier from Database using different Clustering Algorithms", Indian J. Edu. Inf. Manage., Volume 1, pp 388-391, September 2012.**

[4] **P. Murugavel, Dr. M. Punithavalli," Improved Hybrid Clustering and Distance-based Technique for Outlier Removal", International Journal on Computer Science and Engineering, Volume 3, pp 333-339, 1 January 2011.**

[5] **S.Vijayarani, S.Nithya," Sensitive Outlier Protection in Privacy Preserving Data Mining", International Journal of Computer Applications, Volume 33, pp 19-27, November 2011.**

[6] **Ji Zhang," Advancements of Outlier Detection: A Survey", ICST Transactions on Scalable Information Systems, Volume 13, pp 1-26 January-March 2013.**

[7] **Shalini S Singh, N C Chauhan," K-means v/s K-medoids: A Comparative Study", National Conference on Recent Trends in Engineering & Technology, May 2011.**

[8] **Periklis Andritsos," Data Clustering Techniques", pp 1-34, March 11, 2002.**

[9] **Maria Halkidi, Yannis Batistakis, Michalis Vazirgiannis," On Clustering Validation Techniques", Journal of Intelligent Information Systems, pp 107–145, January 2001.**

[10] **Mr Ilango, Dr V Mohan," A Survey of Grid Based Clustering Algorithms", International Journal of Engineering Science and Technology, Volume 2, pp 3441-3446, 2010.**

[11] **S.Vijayarani, S.Nithya," An Efficient Clustering Algorithm for Outlier Detection", International Journal of Computer Applications, Volume 32, pp 22-27, October 2011**

[12] **David Breitkreutz, Kate Casey," Clusterers: a Comparison of Partitioning and Density-Based Algorithms and a Discussion of Optimisations", 2008.**

[13] **Periklis Andritsos," Data Clustering Techniques", pp 1-34, 11 March 2002.**

[14] **Silvia Cateni, Valentina Colla ,Marco Vannucci Scuola Superiore Sant Anna, Pisa," Outlier Detection Methods for Industrial Applications", ISBN 78-953-7619-16-9, pp. 472, October 2008**

[15] **A. Mira, D.K. Bhattacharyya, S. Saharia," RODHA: Robust Outlier Detection using Hybrid Approach", American Journal of Intelligent Systems, volume 2, pp 129-140, 2012**

[16] **Han & Kamber & Pei," Data Mining: Concepts and Techniques (3rd ed.) Chapter 12 ", ISBN-9780123814791**

[17] **Tan, Steinbach, Kumar," Introduction to Data Mining (1st ed.) chapter 10", ISBN-0321321367**

[18] **Maria Kontaki, Anastasios Gounaris, Apostolos N. Papadopoulos, Kostas Tsichlas, Yannis Manolopoulos," Continuous Monitoring of Distance-Based Outliers over Data Streams", Proceedings of the 27th IEEE International**

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

Conference on Data Engineering , Hannover, Germany, 2011.