ISSN 2278-3091

Volume 8, No.6, November – December 2019

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse130862019.pdf

https://doi.org/10.30534/ijatcse/2019/130862019

Outlier Detection for Improving Data Robust by ODAD Clustering Technique

Deepti Mishra[1], Devpriya Soni[2]

[1]Department of Computer Science & Engineering, NIU, Greater Noida
India: itsdeepti.s@gmail.com
[2]Department of Computer Science & Engineering, JIIT, Noida
India: devpriyasoni@jiit.ac.in

ABSTRACT

The paper presents the concept of outliers and its detection by applying an altogether a new approach. Outliers are the odd man out data points falling under the domain of data mining. Data Mining is the evolving heading, now a days because of its ability to deal large amount of data. This paper identifies the outliers in the dataset through an algorithm named outlier detection based on angle and distance based (ODAD) which is based on clustering techniques (which is combination of the angle based and distance based approaches). It encompasses basic five steps: density calculation, cluster identification, angle calculation, Euclidean distance calculation and finally outlier identification and detection. It first calculates the density between the data points to identify the "clusters". Further, to circle out the outliers, the distance-based method of clustering is applied collectively with angle based method to calculate the distance between the data points. The algorithm assigns the rank value to top most outlier data points. Data points having highest rank values considered as outliers. ODAD is implemented in both R and MSSQL.

Key words : Clustering; Datamining ; Outliers; Outlier Detection

1. INTRODUCTION

Outliers are the data points which are quite different from rest of the datasets. In other words, the outlier is the data which is very or apparently different from its neighbours. Presence of outliers "grossly" changes the patterns generated through knowledge for information processing. Outlier detection is an innovative field of data mining. It utilizes many techniques of data mining which can be illustrated as classification and clustering. Data mining is now a rapidly emerging topic in the field of computer technology. It is the extraction of "pure" and "discrete" information from composite data for information retrieval. It is an approach of extraction of information and knowledge from large amount of data. It has the ability to provide "specific", "precise" and almost accurate data for variant purposes. It can also be concluded as the art of knowledge discovery with information processing.

Actually, data mining could be applied to any type of information repository, which comprises relational database, internet database, data of data warehouses, transactional datasets and many more [1]. The methodologies and techniques of data mining can be different depending upon the storage system variety. Data mining comprises summation of techniques from various topics, for example: database techniques, statistics, learning, high level computing, neural network, data visualization, information retrieval, image processing and signal processing. The knowledge which is discovered by applying data mining techniques can be used for conclusion making, information management, pattern designing [2]. Hence data mining can be considered as of pivotal importance in managing large database systems.

The techniques of data mining are based on different kinds of learning. Techniques of data mining are classification, clustering and association rule. Classification is based on supervised learning where user have pre-processed models and training sets [3]. These training sets are used to model the new approached data in the database. The model thus generated can have multiple representation for example classification rules, decision trees and neural network. Clustering is based on unsupervised learning. There are no prepared models or training sets for calculation. It deciphers data objects without a class level initially. Objects are clustered or grouped on the basis of maximum intra group similarity and minimal inter class similarity.

Identification of outliers is a sub topic of data mining. Outlier analysis is a research field for scientists. Outliers are the data points those cannot be fitted in any type of clusters. The outliers are "confounding factors". These objects are somehow different from other objects in the data set. Outliers may be different from complete data sets or may be difficult from its neighbourhood only. Presence of outliers make the results in a confusable state. The patterns generated after the calculations from the data are not authentic and precise because of the outliers. Outlier analysis can discover credit cards fraud by detecting spurious purchases for a given account number as well as location and type of purchase and frequency of purchase. Outliers existence can be detected in conventional data sets or big data [4]. According to Jiawei Han, Outliers usually get generated either by measurement or execution error or consequence of inherent data variability [5]. Data mining algorithms aims at minimizing or eliminating the influence or presence of outliers. But on the other hand, outlier elimination could also lead to missing of a pivotal hereby non-conspicuous information. Outliers presence indicates fraudulent activity in con inquest cases. Outliers are therefore a necessary evil. One of the important data mining tasks, which is referred as "outlier mining" involves outlier detection and its critical analysis. It is very helpful in online fraud detection.

There are two types of outliers: (i) local outliers and (ii) global outliers. When the objects are entirely different from the remaining of the data sets then it is pointed as global outliers. Local outliers are different from its neighbourhood only.

This paper presents the algorithm for detecting outliers from the data named ODAD [6]. The algorithm combines the methodology of 'Angle based' approach and 'Distance based' approach to detect outliers. First the data set is defined into the clusters which itself a new method of clustering, Further, outliers are detected using the concept of "Euclidean" distance and "angle" detection.
The contribution of paper is to enlighten the researchers by exposing the knowledge of outliers and their detection techniques. The aim of the paper is to clarify the researchers that how to apply data mining techniques to identify outliers in the dataset so that the user can obtain precise outcome. The objective of the paper is to analyse and compare ODAD with other pre-existing algorithms.
The rest of the paper designed as, section 2 includes the study of previous work done in outlier detection. Section 3 presents some basic notations and knowledge. Further the section 4 in the paper, presents a new approach for outlier detection. Further the next section 5 presents the complete functioning of algorithm. Section 6 presents its implementation results. Final section presents comparison and discussions on algorithm following the conclusion and future work.

2. RELATED WORK

There are many definitions given by researchers in context of outliers. "An observation that deviates so much from other observations so as to arouse suspicion that it was generated by different mechanism" is defined by Hawkins. [7]

Others defined outliers as those points that differs entirely from other observations. "Outlier is an observation that is significantly different from the other values in a dataset." [8]. outliers also defined the type of dataset as nominal, ordinal, interval and ratio scaled. The different kinds of data are Univariate, Bivariate and Multivariate. In simplest form, it can be said that the outliers are "Odd man out". Analysis and identification of outliers can be applicable to agriculture region, like Biomedical and DNA analysis, Fraud detection, banking [9] etc etc. Outlier detection can be followed by three ways statistical approach, distance based approach and deviation based approach. Statistical approach is based on calculation of probabilities for a data point in the data set.

Usually clustering that is unsupervised learning is applied to detect outliers in the data set. There are various techniques of clustering such as density based and distance based which are used to identify outliers. The data points which are not similar to the identified data points in the clusters can be considered as outlier points.
As suggested by authors [10] Outliers detection can be defined as the part of data cleaning while applying data mining. Outliers can be defined as the point that can be clearly differentiated from the normal point. The paper refers to the concept of new algorithm LDOF to find out the outliers. A process is discoursed for finding outliers applying a new technique, ODHDP named as (outlier detection in High dimension based on projection) which finds the outliers on the basis of projection from dataset [11]. In one of the research paper [12] it has been suggested to use the distance based method to detect outliers. The approach is based on clustering. Suggested by Charu Agarwal [13] Research has been done on high dimensional dataset for outlier detection. New technique has been presented and discussed which uses projection on datasets. A new approach has been proposed by [14], It is a method using evolutionary search for outlier detection which is based on Euclidean distance and density based techniques.

High dimensional data is now a day's an emerging trend [15]. A paper has been published showing a new approach for high dimensional data using weighted hypergraph. One new technique ABOD has been introduced for outlier detection is that It is based on angle calculation. [16] . This mentioned technique is also based on high dimensional data. [17]In their paper authors have applied angle based approach for outlier detection. Their technique calculates the angle between the data points.
A clustering based outlier detection algorithm OFT has been presented for outlier detection which uses K means clustering algorithm [18]. Another paper has [19] presented a class outlier factor (COF) for outlier detection. A new approach has been discussed for outlier detection [20]. Authors have proposed a method which makes use of sub space based and example based methods. There is another approach for outlier detection for large multidimensional datasets given by [21].

3. PRELIMINARIES

3.1 Algebra

In this paper, for detecting outliers, methods of elementary algebra are applied while designing and implementing the algorithm [22], like for detecting coordinates values, applying square root and many more. It is the most basic form of mathematics.

3.2 Trigonometry

Some trigonometric functions are used to calculate the values and computations. In the algorithm some mathematical calculations are performed which are based on trigonometry such as generating the values of angles like SIN ϴ [23]. Sine is trigonometric function of an angle obtained by the division of perpendicular (side opposite to the angle) by hypotenuse. This ratio is given by length of the side opposite to the angle concerned to the length of the longest side of the triangle. In usual considerations in mathematics sine of acute angle is usually defined in context of a right angle.

$$SIN\ \Theta = \frac{c}{a} \qquad (1)$$

It can be understood, the value of SINϴ which is computed in equation 1. The equation is showing the angle between two lines.

3.3 Geometry

Geometry is a branch of mathematics concerned with questions of shape, size, relative position of figures, and the properties of space [24]. The circle and its properties are used, like tangents while designing and implementing the algorithm for outlier detection. Tangents are always perpendicular to the radius of the circle. The meaning of perpendicular is to make angle of 90 by tangent on radius.

3.4 Statistics

Some techniques of statistics such as distance calculations are applied in the paper. [25] Mostly Euclidean distance and Mahalanobis Distance are used for implementing the algorithm. The formula for Euclidean Distance in equation 2 is

$$\sqrt{(x2 - x1)^2 + (y2 - y1)^2} \qquad (2)$$

The formula for Mahalonobis is mentioned in section 7.

4. MATERIAL AND METHODOLOGY

4.1 Problem Statement

In the technique, it is attempted to solve the problem of outliers by detecting both kind of outliers (i.e. local and global) by using clustering techniques.

Let x be the set of database;
T be the number of data points in data set
C be the set of clusters= {C1,…………Cm}

Then, let D be the number of top most outliers to be detected i.e. the problem to be solved.

Outlier detection technique presented by the previous authors in the above-mentioned papers was either angle based or distance based techniques. Authors used either partitional clustering or density based clustering. Authors in the paper, detect either global outliers or local outliers.

4.2 Procedure

In this paper, a new approach is introduced and proposed [6] to find the outliers in the dataset called "Outlier Detection using Angle Based and Distance based Approach" (ODAD). It consists of four steps: density calculation, cluster identification, angle calculation, Euclidean distance calculation. Density calculation is done on various squares in a rectangular plane. Further clusters are designed on the basis of densities of each square. Outliers are detected from the respective clusters.

One of the definition of outliers given by [26] can be stated as "Outlier analysis can be outlined as: Let in a set of data points n, the expected number of outliers are k, problem is to find the top most k outliers or in other words dissimilar objects. The outlier mining problem can be considered as "find and state data that could be labelled as erratic within the given dataset and thereby make out an effective way to decoct the outliers thus mentioned".

Our aim is to detect the outliers. The algorithm combines the methodology of "angle based" approach and "distance based" approach. First the data set is defined into the clusters, Further, outliers are detected using the concept of "Euclidean distance" and "angle detection". To find the solution angle calculations are done by measuring the angles arising between the data points of the provided dataset. At the outset, to find out the angles in between the points and then to mention the clusters.

In first step the data points are plotted in the XY coordinate system. Further, the minimum coordinates and the maximum coordinate values are identified to design a rectangle plane. That rectangle is further divided into squares. The square having high density of data points is selected. Again the neighbourhood squares are observed for matching the densities with a high density square to create a cluster. The process is continued till all the clusters are designed.

Furthermore, the centroid of square having highest density is searched. Euclidean distance is calculated between the centroid and with the point lying outside the clusters or on the boundary range. The angle Sine ϴ is to be calculated between the tangents and the outer point. If the value of the angle is smaller than the threshold value then that point is considered as outlier.

- STEP 1 Plotting points into rectangular plane

The plotted points are set into a rectangular plane by finding
A = MINIMUM OF X, MAXIMUM OF Y = min(x), max(y);
B = MAXIMUM OF X, MAXIMUM OF Y = max(x), max(y);
C = MAXIMUM OF X, MINIMUM OF Y = max(x), min(y);
D = MINIMUM OF X, MINIMUM OF Y = min(x), min(y);

Now there are four coordinates namely A,B,C,D as four corners of the rectangle. All the plotted points will fall within the rectangle as shown in figure 1.
Evaluate the boundary values (co-ordinates) of the plane on which the bivariate data is distributed.

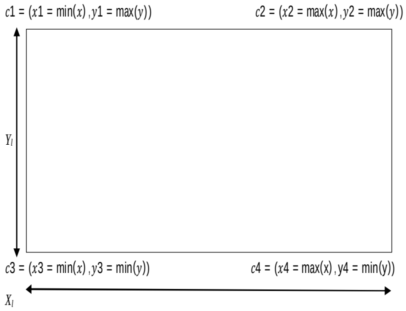
$c1 = (x1 = \min(x), y1 = \max(y))$
$c2 = (x2 = \max(x), y2 = \max(y))$
Y_l
$c3 = (x3 = \min(x), y3 = \min(y))$
$c4 = (x4 = \max(x), y4 = \min(y))$
X_l

Figure 1: showing a outline of dividing plane

- STEP 2 Dividing plane into squares

Now the plane is divided into squares by taking input from the user. As per the assumption, the number of squares in the rows should be equal to the number of squares in the columns.
Now define the value for variable N i.e. number of squares in which it is supposed to divide the plane.
If Xl = Yl
Then, ideally number of squares on each axis should be Ni = $\sqrt{N}$
So, the length of square side is assuming figure 1;

$$Xvar = (C4 - C3)/Ni \quad (3)$$

Or $Yvar = (C1 - C3)/Ni \quad (4)$

$$\text{Square side length } (Sl) = \begin{cases} Xvar & , if\ Xvar > Yvar \\ Yvar & , Otherwise \end{cases} \quad (5)$$

But the designed plane is a rectangular plane, so, the total number of squares in the row side will differ from the total number of squares in the column side.
Which means,
X axis of rectangle ≠ Y axis of rectangle.
So
If Xl ≠ Yl then the number of squares on each axis will be distributed as follows

$$\text{Number of squares on X axis } Nx = \left| ceiling(\sqrt{Ni} \times \left(\frac{Xvar}{Yvar}\right)) \right| \quad (6)$$

$$\text{Number of squares on Y axis } Ny = \left| ceiling(\sqrt{Ni} \times \left(\frac{Xvar}{Yvar}\right)) \right| \quad (7)$$

While calculating the number of squares on each axis in the above given equations ceiling function is referred which translates the decimal value to the next whole number e.g. if the number of squares on X axis is 10.2 then ceiling function translate this value to the next whole number i.e. 11. Due to this the optimum value of total number of squares (i.e. Nf will be >= Ni (Total number of square values which was assumed initially by the user).

So, Optimum number of total squares i.e.

$$Nf = (Nx \times Ny) \quad (8)$$

Co-ordinates of square (SQrc) = {(Crc,1), (Crc,2),(Crc,3),(Crc,4)} ; where r= row number and c = column number and count starts from bottom left side of the square. Equations to derive the co-ordinate values are given below

$$Crc,1 = \{Xrc + (Sl \times (c-1)), Yrc + (Sl \times (r-1))\}, \quad (9)$$

$$Crc,2 = \{Xrc + (Sl \times (c)), Yrc + (Sl \times (r-1))\}, \quad (10)$$

$$Crc,3 = \{Xrc + (Sl \times (c)), Yrc + (Sl \times (r))\}, \quad (11)$$

$$Crc,4 = \{Xrc + (Sl \times (c-1)), Yrc + (Sl \times (r))\}, \quad (12)$$

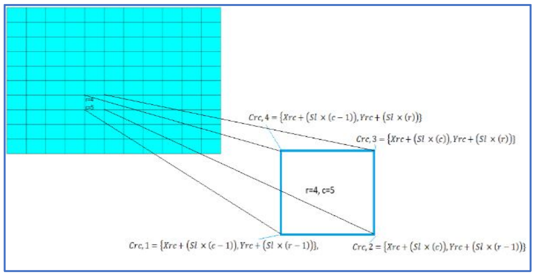
r=4
c=5
$Crc,4 = \{Xrc + (Sl \times (c-1)), Yrc + (Sl \times (r))\}$
$Crc,3 = \{Xrc + (Sl \times (c)), Yrc + (Sl \times (r))\}$
r=4, c=5
$Crc,1 = \{Xrc + (Sl \times (c-1)), Yrc + (Sl \times (r-1))\},$
$Crc,2 = \{Xrc + (Sl \times (c)), Yrc + (Sl \times (r-1))\}$

Figure 2: specifying the visualization of squares divided

In the algorithm as shown in figure 2, the analysis and value of coordinates of squares are calculated and put the entry in the square table. Now, the plane is divided into squares, the formula is given in equation (5) and equation (6).

- STEP 3 Now identify in which square each point falls.

Point belongs to that square in which points X,Y variable values falls between the maximum and minimum value of the square X,Y co-ordinates.

In the case where point X,Y co-ordinate values fall on the square boundary that point can get calculated in two squares. So, for ease of calculation it is kept that point in the square where it get calculated first or the square which contains rc (i.e. r= row no., c= column no.; e.g. for row no. 5 and column no. 6, rc value will be 56) value smaller than the other.

Identifying clusters

Furthermore, the algorithm calculates the density of each square. The square having the highest density put in the cluster, name as cluster 1.
Then, the densities of each adjacent square are checked till the value of density gets 0 or high from the calculated cluster i.e. cluster 1.

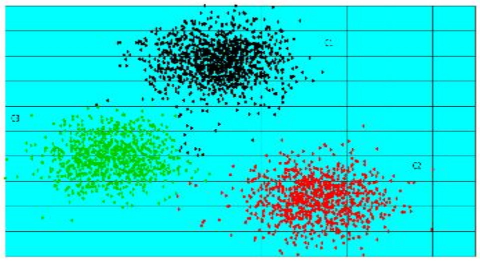

Figure 3: specifying the formation of clusters

This procedure will design the clusters in the plane as shown in figure 3.

Now evaluate the density of each square i.e. count of data-points resides in the square. And then select the Square which has the highest density. This square will be the part of first cluster i.e. C1. Then iteratively calculate the density of the adjacent squares till the density will be either evaluated as 0 or higher from the calculated square. These squares will be the part of the cluster C1. Once the calculations are done for cluster C1 then repeat these steps to identify the other clusters in the plane.

- STEP 4 Angle based approach to detect the outlier points in clusters

After completing this procedure, it is to be found out those squares which are lying outer side the cluster. It is because the outlier would lie on the outer square only.
ODAD will find the high-density square within each cluster. The algorithm identify the centroid of each high -density square. Centroid is the centre point inside the high-density square. It is calculated by crossing diagonals or by taking half of X and Y axis of the square.

A circle is drawn inside the high-density square touching the edges.

Now, those points are selected which lies outside or within the outer square because outlier lies on the outer square only.

Then tangents are drawn to that outer point. These tangents will create an angle Ɵ.

Outliers are those data points in a cluster which are far away from the normal cluster area or the dense area of the cluster and induces the amicable level of error while calculating the descriptive statistical parameter values.
The following approach is developed in ODAD to detect the outliers and it is mainly based on the measurement of the angle (Ɵ) created by the point on the circle which is drawn inside the highest density square of the cluster. This circle radius is equal to half of the length of the square i.e. $r = Sl/2$.

And the co-ordinates of the centre of the circle are as follows

$$C(x,y) = \left\{Xrc + \left(\left(Sl \times (c-1)\right) + \frac{Sl}{2}\right),\ Yrc + \left(\left(Sl \times (r-1)\right) + \frac{Sl}{2}\right)\right\} \quad (13)$$

Then calculate the Euclidian distance (d) between the centre of the highest density square of the cluster and the points which are lies outside this square in the same cluster.

$$d = \sqrt{((P(x)^2 - C(x)^2) + (P(y)^2 - C(y)^2))} \quad (14)$$

If a tangent is drawn on the circle from the point and radius will touch the tangent line, then the tangent and radius will be perpendicular to each other. So, the angle made by the point between tangent line and the line from point to the centre of the circle will be calculated as

$$\theta = \sin^{-1}\left(\frac{r}{d}\right) \quad (15)$$

So, the points which makes very small angle on the circle drawn inside the highest density square in a cluster can be identified as outlier and can be further analysed to check their impact on the descriptive statistical parameters as shown in figure 3.

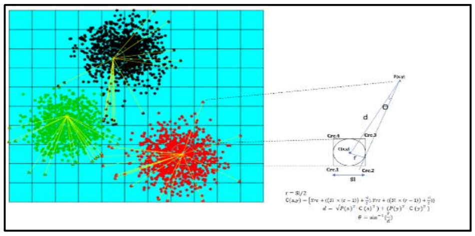

Figure 4: Showing calculation of angle and distance to detect outlier

If the calculated angle is Acute and having value less than the threshold value, then that point is considered as an outlier as in figure 4.
Current defined algorithm ODAD inculcates the notions of both angle as well as distance-based approaches.

5. ALGORITHM

The algorithm ODAD (to detect outliers.)

```
1.  Input: Dataset, number of squares.
2.  Output: clusters and outliers
3.  Initialization:
4.  Read the data from dataset.
5.  Begin:
6.  Plot the data into X-Y coordinate.
7.  Plot the data in a rectangular plane.
8.  For i min x, min y to max x, max y
9.  Divide the plane into squares.
10. End for;
11. While (density < maxdensity)
12. do
13. Find the square with highest density.
14. Design the clusters.
15. End while;
16. Calculate Euclidean distance for the data points.
17. Calculate the angle for the data points far from the cluster.
18. Detect outliers.
19. End
```

Presented above in the algorithm, is a representation showing the methodology for outlier detection procedure followed in ODAD. It depicts an initial large size database from where a designated plot data is carved out. This plot data is then subjected to clustering manoeuvres namely density based, and distance based. Clusters thus formed undergo distance based and angle-based outlier detection techniques. Consequently, generated outliers are then numerically placed in a table to sort out the most effective outliers.

6. EXPERIMENTAL RESULTS

The dataset, which is used, is a bivariate dataset. In statistics, it is defined as the data having two variables. A scatter plot is usually used to represent the variables of bivariate dataset. Now it becomes very easy to detect relationship between the variables if there is one existing.

Dataset XCLARA which is used, is picked from R datasets. It is bivariate dataset which finds 3 clusters [27]. Datafile format is in .xls or .doc file format. This data file is picked to find the clusters. The file contains 3000 record sets (objects) [28]. The dataset is retrieved from below mentioned link of R datasets.

The dataset is retrieved from
https://vincentarelbundock.github.io/Rdatasets/datasets.html
The information and details about the data set can be retrieved from
https://stats.ethz.ch//R-manual/R-devel/library/cluster/html/Xclara.html
Dataset -The rest of information calculated by us about XCLARA is mentioned below
Number of attributes are 2, giving the *x* and *y* coordinates of the points.
Missing values N/A
Total Number of records are 3000.

Software used for the implementation is Microsoft SQL Server 2014. All experiments are conducted on a Windows 8.1 with intel i5 CPU. Experiments are conducted in SQL software. Data is collected from R datasets. The dataset can be used for clustering and classification. It is a special purpose computer language for handling data bases. It manages data very efficiently. It can execute queries and retrieve data very easily. New records can be created and old one can be updated in proper manner. Stored procedures can be created so as to implement the problems for finding the results. SQL provides good platform for interaction of human and databases.

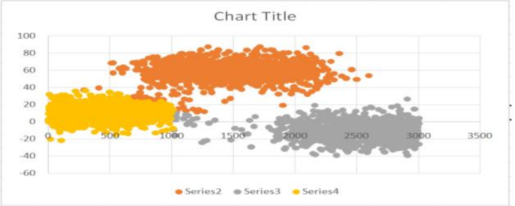

Figure 5: showing the 3 different clusters

Clusters can be shown in the figure. 5. Data with outlier for rank values and different calculated values in ODAD in SQL. Both angle detection values and Euclidean distance values are calculated for identifying the clusters and outliers. The outliers have given the ranks starting from 1 onwards. The outlier having highest rank 1 denotes that it is farthest object from its cluster. It is clearly visible that outlier having top rank also have the most acute angle. As the data points goes far from the clusters the angle becomes acute.

The three things consider the Rank value of outliers –

1. Order By cluster count in ascending order
2. Sine angle in ascending order
3. Euclidean distance in descending order

The above-mentioned points signify that clusters which are lightly dense and the coordinates which are making least or lesser angle from the cluster centroid and far from or in other way the distance is high from the center are considered as the outlier points.

To detect outlier points few things are considered according to the priorities.

1. Density of the clusters
2. Sine angle
3. Distance from centre

The result generates the outliers which are cluster specific.

The field value gets evaluated on the following parameter. In the cluster points which contains the minimum value for the outlierTest field are the most likely outlier coordinates point which falls under the square of least density, and then angle is calculated if it is made by the point is acute, and then the farthest point in the square is observed. The outliers calculations with their Rank values is done. In Fig.6 and 7 clearing the outliers and the clusters. Outliers are clearly visible in different colour with respect to clusters.

Uniqueness of this approach is the utilization of the angle based and the distance-based methods between data points which is quite relevant in contemporary large size data applications. This intuitive and understandable graphical presentation of outlier detection adopted in this thesis will actually enhance the evaluation process by reducing the unwanted data.

Figure 6 showing the results in SQL and R Studio respectively. All the three clusters are undoubtedly noticeable and visible in the figure.

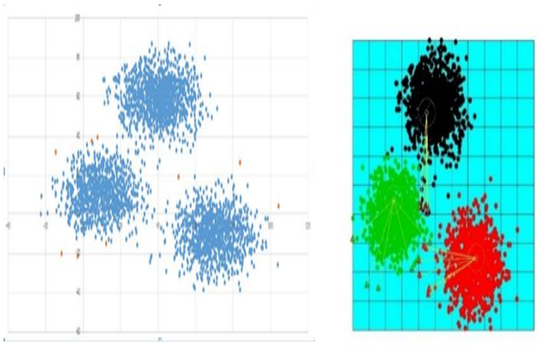

Figure 6: showing clearly the outliers with different colours SQL(a) and R STUDIO(b)

Here the clusters which are displayed in the figure 7 are the implemented by applying the algorithm ODAD using Xclara database in Tableau Software..

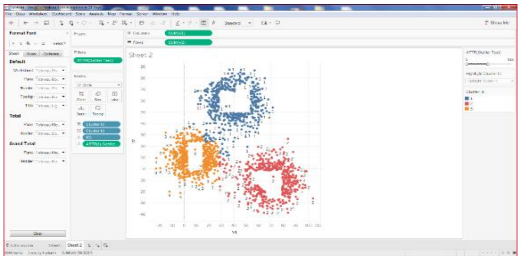

Figure 7: Showing the results of clusters and outliers in Tableau software

7. COMPARISON

The comparison of ODAD is accomplished as it is compared by two methods in different ways in R. In first fragment, ODAD is compared with another method which apply mahalonobis distance for calculation. In other fragment, ODAD is compared with well known algorithm of clustering that is K Means.

In this part, the comparison of algorithm (ODAD) is shown which is implemented both in MSSQL and R further which applies Xclara data set. The comparison is done with implementation of outlier detection in R Studio on the same data set Xclara. This file is applied as CSV file in R Studio. In R Studio, Mahalanobis distance is applied to identify the points which can be stated as outliers.

The Mahalanobis distance for an observation $\vec{x}$= (x1,…..,xn)^T

from a set of observations with mean $\vec{u}$ = (u1,….., un)^T and covariance matrix S

$$D_m(\vec{x}) = \sqrt{(\vec{x}-\vec{u})\, S^{-1}(\vec{x}-\vec{u})} \qquad (16)$$

Comparison of ODAD algorithm is done with the method, Method2, suggested on the following website which is based on Mahalanobis Distance on the same dataset XCLARA.

http://stat.ethz.ch/education/semesters/ss2012/ams/slides/v2.2.pdf

Initially the implementation of the suggested method Method2 , in R studio.

First the command is applied for reading the data file on console of R Studio as CSV file by applying the function

Read.csv()

That CSV file is converted into matrix for further computations by applying

as. matrix() function.

The whole computation needs a covariance formula for matrix data set for calculating Mahalanobis distance as Cov(A), And

Mahalanobis()

In the calculation, it can be clearly visible those data points which are calculated as outliers in ODAD algorithm are not considered as outliers in method followed in R Studio applying Xclara database file.

In figure 8 There are two variable One is named as Rank and another one is named as ML Rank. Rank denotes the calculated outlier rank in ODAD algorithm and ML Rank denotes the rank value of data points which can be considered as outliers. The data point which is detected as outlier having Outlier Rank value as 1 in ODAD algorithm at the mean time the same data point has ML Rank value 3. As it can be observed other cases also for the data points having different rank values in ODAD and R studio.

In this part, a comparison in outcomes is discussed among ODAD and K Means algorithms. As shown in figure 9 it is clearly visible the results generated from ODAD are more precise than the results generated from K Means algorithm. The comparison is computed in RStudio on the above mentioned data set. Clusters designed by applying ODAD are clearly distinguishable than the clusters generated from K Means algorithm in which data points are mixed with each other within the clusters.

Outliers generated by applying ODAD are also noticeable as outliers are pointed by red triangle in the figure. Whereas Outliers are not detected by applying K Means algorithm in the figure.

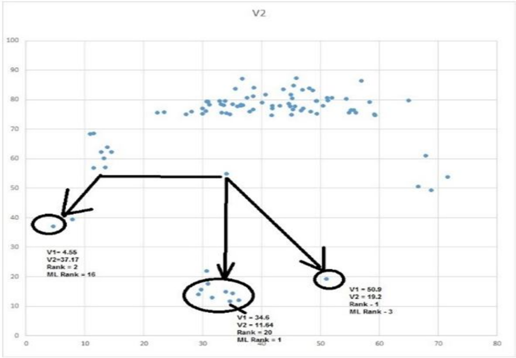

Figure 8: Comparison results

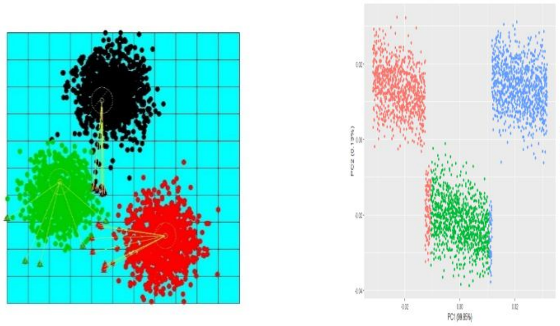

Figure 9: comparison of ODAD (a) with Kmeans (b)

8. DISCUSSIONS

Our algorithm ODAD is more precise than the algorithm which is presented in section 7 comparison, that is R Studio algorithm, as ODAD considers outliers on the basis of distance values as well as density regions instead of only distance values as done in compared one. The algorithm implemented in R Studio is based on only distance calculations. ODAD detects both kinds of outliers global and local.

For comparison, first there is implementation of the algorithm which is mentioned in the link in section 7, then the results are compared with ODAD, in which it is found that ODAD performs better as in results, accuracy, methodology and visualization.

In results of ODAD it is clearly mentioned and visible that it can detect both kinds of outliers. ODAD provides more precise and accurate results in the form of rank values as mentioned in section 7.

The methodology of ODAD is based on firstly, finding the clusters on basis of density and mathematical calculations, then detecting the outliers on the basis of various parameters like angle calculation, distance calculation and density calculation as compared to other algorithm which calculates outliers only on the basis of distance calculation.

Through visualization, outliers can be understood and realized very effortlessly.

According to the comparisons, it can be concluded that the study shows that for high dimensional data the distance based approach does not perform well. Angle based approach is better for detecting outliers in high dimension data set.

9. CONCLUSION AND FUTURE WORK

The proposed study and analysis of algorithm which incorporated both angle which is measured and distance based approaches for outlier detection is implemented in this paper. Bivariate data set is used to show the implementation of the algorithm. Generally, pre-existing algorithms identify either local or global outliers but the algorithm ODAD presented in this paper can identify both kind of outliers i.e. local outliers and global outliers. In addition, ODAD is also compared with K-Means and the results shows that ODAD delivers better cluster outcomes than K-Means. The concept of algorithm is totally based on mathematics and statistics additionally requires only single parameter as input. In current scenario there are numerous techniques are presented to detect outliers for dissimilar data sets. Currently, study exhibits that a herculean effort is required for high dimensional data cases. Further research reveals that there are very few methods of outlier detection for bivariate data set. So forth it has been tested the proposed algorithm with the bivariate dataset but in future there will be a comparison of the results of the algorithm with the benchmark dataset. In future the algorithm can be implemented and works for high dimensional data set. In the last, a definition of outliers can be generated as those points which can induce large errors during cluster related calculations.

REFERENCES

1 D. Mishra and D. Soni. **A comprehensive Overview on Data Mining Approaches and Applications,** *Interntional Journal of Computer Science and Information Technologies IJCSIT,* pp. 7814-7816, 2014.

2 W. Y. Lai, K. K. Kuok, S. Gato-Trinidad and K. X. L. Derrick. **A Study on Sequential K-Nearest Neighbor (SKNN) Imputation for Treating Missing Rainfall Data,** *International Journal of Advanced Trends in Computer Science and Engineering,* vol. 8, no. 3, pp. 363-368, 2019. https://doi.org/10.30534/ijatcse/2019/05832019

3 M. Syamala and N.J.Nalini. **A Deep Analysis on Aspect based Sentiment Text Classification Approaches,** *International Journal of Advanced Trends in Computer Science and Engineering,* vol. 8, no. 5, pp. 1795-1801, 2019. https://doi.org/10.30534/ijatcse/2019/01852019

4 L. Chen, S. Gao and X. Cao. **Research on real-time outlier detection over big data streams,** *International Journal of Computers and Applications,* 2017.

5 Z. Li, Ziyuan Li, Ning Yu and S. Wen. **Locality-Based Visual Outlier Detection Algorithm for Time Series,** *Security and Communication Networks,* 2017. https://doi.org/10.1155/2017/1869787

6 D. Mishra and D. Soni. **An Integrated Method for Outlier Detection with Analytical Study of Distance based and Angle based Approaches,** in *ACM ICTCS*, Udaipur, 2016.

7 X. Jingke. **Outlier Detection Algorithm in Data Mining,** in *second International Symposium IITA,* Shanghai, 2008.

8 V. Ilango, R. Subramanian and V. Vasudevan. **A five step procedure for outlier Analysis in Data Mining,** *European Journal of Scientific Research,* pp. Vol 75 No 3, pp 327-339., 2012.

9 S. Koteeswaram, P.Visu and J.Janet. **A Review on Clustering and Outlier Analysis Techniques in Data Mining,** *American journal of Applied sciences*, pp. 254- 258, 2012.

10 P. Rajendra and N. S. D. Jatindra Kumar. **An outlier detection based on clustering,** in *IEEE second international conference on emerging application of information technology*, 2011.

11 D. Ping, Ji-Yong and W.Yan-Xia. **Outlier Detection in High Dimension Based on Projection,** in *IEEE proceedings of fifth International conference on machine learning and cybernetics*, 2006.

12 S. Pachgade and S. Dhande. **Outlier detection over data set using cluster-based and distance-based approach,** *IJARCSSE, vol 2 issue 6*, 2012.

13 C. C. Agarwal and P. S. Yu. **Outlier Detection for High Dimensional Data,** in *ACM SIGMOD*, California, USA, 2001.

14 A. Banerjee . **Density based evolutionary outlier detection,** in *ACM GECCO*, 2012.

15 L. YinZhao, W. D., R. JiaDong and H. ChangZhen. **An Improved Outlier Detection Method in High-Dimension based on Weighted Hypergraph,** in *IEEE Second international symposium on electronic commerce and security*, 2009.

16 K. Hans-Peter, S. Matthias and Z. Arthur . **Angle based Outlier Detection in High Dimensional Data,** in *ACM SIGKDD*, 2008.

17 P. Ninh and P. Rasmus. **A Near-Linear Time Approximation Algorithm for Angle based Outlier Detection in High Dimensional Data,** in *ACM KDD*, 2012.

18 H. Behera, G. Abhishek and M. S. Ku. **A new hybrid k-means Clustering based Outlier Detection Technique for Effective Data Mining,** *IJARCSSE, vol 2 issue 4*, pp. 287-292, 2012.

19 H. N. M. and S. M. K.. **Class Outlier Mining-Distance based Approach,** *International journal of Electrical and computer engineering*, pp. 448-461, 2007.

20 L. Yuan and K. Hiroyuki. **DB-Outlier Detection by Example in High Dimensional Datasets,** *IEEE*, 2007.

21 A. Leman, T. Hanghang, V. Jilles and F. Christos. **Fast and Reliable Anomaly Detection in Categorical Data,** in *ACM CIKM*, 2012.

22 A. L. Gorodentsev. **Algebra** I, Springer International Publishing, Springer, 2016.

23 I. Gelfand and M. Saul. **Trigonometry,** Birkhauser Basel, Springer , 2001.

24 R. Fenn. **Geometry,** Springer-Verlag London, 2001.

25 C. Heumann, M. Schomaker and Shalabh. **Introduction to Statistics and Data Analysis,** Springer International Publishing, 2016.

26 J. Han and M. Kamber. **Data Mining: Concepts and techniques,** Springer, 2001.

27 "XCLARA," [Online]. Available: https://stat.ethz.ch/R-manual/R-devel/library/cluster/html/xclara.html.

28 A. Struyf, M. Hubert and P. J. Rousseeuw. **Clustering in an Object-Oriented Environment,** *Journal of Statistical Software*, 1996.