Fuzzy Membership Function: Outlier Detecting Methods and Detecting the Outlier by Using Five Number Summary of Data

1Mounika.P, 2Sowbarnikaa.A, 3Shamruthi.M
1Assistant Professor, 2Student, 3Student
Sri Krishna Arts and Science College

***Abstract* - In the fuzzy system describing the membership function is very important activity in data science. An outlier is the thing which is consider extremely dissimilar from the rest of the other objects. Distance based methodology, density based methodology and deviation based methodology and five number summary of data etc., are the methods which plays a vital role in identifying the outliers. Deceptive results in many things is obtained due to the presence of outliers. So it is very important to identify and eliminate the outliers in order to prevent the bad impressions which is created due to the presence of outliers. Then in human error, environmental changes, error in the instrument and in malicious activity etc.; detecting the outliers plays a important role in these applications. In this paper constructing fuzzy membership function by using the five number summary of data method and the other methods which is involved in outlier detection have been discussed and explained respectively. The main advantage is that the accuracy of prediction have been obtained while removing the outliers.**

***keywords* - Outliers, Fuzzy Set, Membership Function, Five Number Summary data, density based method, deviation based method, distance based method.**

I. INTRODUCTION

Fuzziness that is characterized by its membership function is described by using the fuzzy logic. The degree of truth in the fuzzy logic is represented by the membership function. Lotfi A. Zadeh is the one who introduced the membership function first[11]. The most familiar fuzzy membership functions are triangular fuzzy membership function, impulsive fuzzy membership function , gaussian fuzzy membership function and trapezoidal fuzzy membership function[8]. Solving the practical problems by the experience than knowledge is the technique of the membership function. Some of the important causes for the outliers are mentioned as human error. The faulty reporting system or human error, Environmental change may cause the outliers. A simple and new buying pattern among the people or the environmental change itself might cause the outliers presence, instrument defects, instruments error which can be used to calculate and also the report may be the cause for the outliers and the malicious activity; The outliers might exists due to the credit card fraud, network system hacking etc., Then in many fields such as agriculture, medicine, power systems, production, transportation etc., Here uncertainty can be reduced by the fuzzy logic and the membership function plays an important role in these implementations. Also solving different kind of problems in fuzzy mathematics it plays an vital role. There are multiple ways defining membership functions. Five states of fuzzy membership function have been formulated by using five number summary of data method have been discussed in this paper. Triangular shape, semi trapezoidal, and trapezoidal are the linear representation of these membership function which will help us to rectify the outliers in the data set. We can plot the graph of the membership function. Five number summary of data method is the simplest method to understand which we can apply in any type of the problem in outlier detecting. The observation whose value exceeds the values of other observations or the thing which is located away or detached from the most system is said to be as outlier. A special approach have been discussed to detect the outliers by using the information of membership functions. The low outer fence, low inner fence, 1st quartile, 3rd quartile, upper inner fence and upper outer fence are the six points which involves in constructing the fuzzy membership functions. Five states of the fuzzy membership functions have been developed in this process. By using the first state and last state of the degree of membership function the outliers can be identified. In this process the obtained results match also the obtained results of the box plot. Membership functions of values is created by using the concept of five number summary of data which also rectifies the outliers if any. This is procedure is very simple. Methodologies like distance based methodology, density based methodology and deviation based methodology helps in rectifying the outliers. Fuzzy set and the membership function definitions have been given, algorithm of the 5 number summary of data and box plot have been discussed. Real world examples for the five number summary of data method are presented. Methods which helps in detecting the outliers like distance based method, density based method and deviation based outlier technique have been explained. Conclusion of the paper will be discussed in the last.

II. FUZZY SET AND MEMBERSHIP FUNCTION

The word fuzzy represents the vagueness, uncertainty. If the boundary of a piece of information is not well-defined then fuzziness occurs. For example consider the words like high, young, tall are fuzzy. There is no quantitative single value which defines the term "young". Some people, seems young at the age 24 and for some other seems young at the age 34. The term young has no exact boundary. The elements which have degree of membership between the values 0 and 1 is said to be fuzzy set theory and it is the extension of the classical set theory. An element which is either in A or not in the set A, here $x \in A$ or $x \notin A$, in the traditional set theory, in which this kind of set is considered as crisp set [8]. The set which is characterized by the fuzzy membership function $\mu_A(x) \in [0,1]$ is the fuzzy set. If $\mu_A(x)= 0$, which implies that $x \notin A$ and on the other hand $\mu_A(x) = 1$ where $x \in A$ [8].

Definition

The two distinct notations are most common employed in the literature to denote membership functions. One of the membership function of a fuzzy set "A" is represented by $\mu_A(x)$, which is μ_A: $X\rightarrow [0,1]$. In the other one the membership function is represented by A(x) and it has the same form $A: X \rightarrow$:X→1]. The m fuzzy set "A" of the membership function on the universe of discourse X is considered as μ_A: $X \rightarrow [0,1]$, where the each element of X is mapped to the values between 0 and 1 [14].

Definition

The membership function represents fully the fuzzy set. If suppose X is the universal set. Here the fuzzy set "A" is called as the set of ordered pairs such that A = {(x, μ_A(x)): x ∈ X, X, $_A$(x)∈ [0, 1]}[0,1]}. Let X the nonempty set of objects which is referred as the referential set and the unit interval [0, 1] is called as an valuation set and ∀ x ∈ X which denotes the grade of membership of x [14].

Example

Let us consider three fuzzy sets. Y_{a1}: $X\rightarrow [0,1]$, $M_{a1}: X \rightarrow [0,1]$ and $O_{a1}: X \rightarrow [0,1]$ this represents the concepts of the young, middle-aged and old-aged respectively. The reasonable expression of this method by the trapezoidal membership functions are $Y_{a1}(y)$, $M_{a1}(y)$ and $O_{a1}(y)$. These functions are defined on the interval Y = [0,80], where Y is the set of ages of human beings such that

$$\text{Young aged, } Y_{a1}(y)= \begin{cases} 1; & y \leq 20 \\ \frac{35-y}{15}; & 20 < y < 35 \\ 0; & y \geq 35 \end{cases}$$

$$\text{Middle aged, } M_{a1}(y)= \begin{cases} 0; & y \leq 20 \\ \frac{y-20}{15}; & 20 < y < 35 \\ 1; & 35 \leq y \leq 45 \\ \frac{60-y}{15}; & 45 < y < 60 \\ 0; & y \geq 60 \end{cases}$$

$$\text{Old aged, } O_{a1}(y)= \begin{cases} 0; & y \leq 45 \\ \frac{y-45}{15}; & 45 < y < 60 \\ 1; & y \geq 60 \end{cases}$$

The membership function of this linear representation is given,

FIGURE.1. Membership functions indicating the concepts of a young-aged, middle-aged, and old-aged persons.

III. FIVE NUMBER SUMMARY AND BOX PLOT

In this five number summary data method it consists of 5 data sets. Then that 5 data sets are minimum, Q_1 (The first quartile or 25th percentile), median, Q_3(third quartile or 75th percentile), maximum. The other name of box plot method is whisker plot method[9][15]. Here it is the visual representation of a data set which consists of five number summary of data or six summary of data. Here the box plot consists of the following steps [9][16]

Primary data should be shorted the first.

Next the Median, Quartiles Q1 (25th percentile), Q3 (75th percentile), and Inter-quartile range: $IQR = Q_3 - Q_1$ should be calculated.

1.5× IQR below Q1 and 1.5 × IQR above Q3 are the values to be calculated. Then these points are considered as the lower and the upper inner fences, respectively.

The central box plot extends from the 25th percentiles to the 75th percentiles. This box is divided into two parts at the median value of the data sets.

The line segments which is projecting away from the box is extending in both directions to the adjacent value. The points which is 1.5 times the length of the box far away from either quartile are the adjacent values. The other data points which is outside this range are considered as the little circles; then these are represented as the outliers or extreme observations which is detached away from the other data set. The points which is observed that fall outside the two inner fences are represented as the outliers. Outliers are classified into two types. They are mild outliers and the extreme outliers. Here, we describe the two outer fences as the lower outer fence at 3 × IQR below first quartile and the upper outer fence at 3 × IQR above third quartile.

An observation which is outside either of two inner fences but within either of two outer fences, it is described as the mild outlier. An observation which is outside either of the two outer fences is described as extreme outlier.

The fuzzy membership function can consists of multiple states which depends on the domain of the data sets. The states of the fuzzy variable are fuzzy sets indicating the concepts as very low, low, medium, high, very high etc.; we define the five states of the fuzzy membership function here by using five number summary of a data sets. Here we find the algorithm of five number summary to produce the first investigation of data. From that summary we will select six points as lower outer fence, lower inner fence, first quartile, third quartile, upper inner fence and upper outer fence. By using these five points we can construct five states of membership function as follows. We also draw the graph of membership functions by using EXCEL programming and identify the outliers from this process. If $X = \{ x_1, x_2, x_3, \ldots\ldots x_n\}$ is a set of n observations, and A_1: X → [0,1] , A_2: X→ [0,1] , A_3:X→ [0,1] , A_4:X→[0,1] and A_5: X→[0,1] are the fuzzy sets defined on X indicating the concept of the smallest, small, medium, large and the largest value respectively then the proposed five states fuzzy membership functions are given as follows[1]:

Smallest, $A_1(X) = \begin{cases} 1; & x \le Q_1 - 3I_{qr} \\ \frac{Q_1 - (x + 1.5I_{qr})}{1.5I_{qr}}; & Q_1 - 3I_{qr} \le x \le Q_1 - 1.5I_{qr} \\ 0; & x \ge Q_1 - 1.5I_{qr} \end{cases}$ =1;Q_1 -(x+1.5I_qr)/1.5I_qr0;;

Small, $A_2(x) = \begin{cases} 0; & x \le Q_1 - 3I_{qr} \\ \frac{x - (Q_1 - 3I_{qr})}{1.5I_{qr}}; & Q_1 - 3I_{qr} \le x \le Q_1 - 1.5I_{qr} \\ \frac{Q_1 - x}{1.5I_{qr}}; & Q_1 - 1.5I_{qr} \le x \le Q_1 \\ 0; & x \ge Q_1 \end{cases}$

Medium, $A_3(x) = \begin{cases} 0; & x \le Q_1 - 1.5I_{qr} \\ \frac{x - (Q_1 - 1.5I_{qr})}{1.5I_{qr}}; & Q_1 - 1.5I_{qr} \le x \le Q_1 \\ 1; & Q_1 \le x \le Q_3 \\ \frac{(Q_3 + 1.5I_{qr}) - x}{1.5I_{qr}}; & Q_3 \le x \le Q_3 + 1.5I_{qr} \\ 0; & x \ge Q_3 + 1.5I_{qr} \end{cases}$

Large, $A_4(x) = \begin{cases} 0; & x \le Q_3 \\ \frac{x - Q_3}{1.5I_{qr}}; & Q_3 \le x \le Q_3 + 1.5I_{qr} \\ \frac{(Q_3 + 3I_{qr}) - X}{1.5I_{qr}}; & Q_3 + 1.5I_{qr} \le x \le Q_3 + 3I_{qr} \\ 0; & Q_3 \ge Q_3 + 3I_{qr} \end{cases}$

Largest, $A_5(x) = \begin{cases} 0; & x \le Q_3 + 1.5I_{qr} \\ \frac{x - (Q_3 + 1.5I_{qr})}{1.5I_{qr}}; & Q_3 + 1.5I_{qr} \le x \le Q_3 + 3I_{qr} \\ 1; & x \ge Q_3 + 3I_{qr} \end{cases}$

Q_1 = First quartile

Q_3 = Third quartile and

The defined fuzzy membership function follows two properties:

For all the states of cross over point is 0.5 which indicates these sets are symmetric.

Height of the fuzzy set is max $\{\mu_A(x)\} = 1$.
Membership functions grant us to graphically represent a fuzzy set. Then the x axis indicate the universe of discourse, whereas the y axis indicate the degrees of membership in [0,1] interval. Graphs of these functions have trapezoidal, semi trapezoidal and triangular shapes which are most common in many applications. For the outliers we consider the membership functions as $A_1(X)$ and $A_5(X)$. As outliers are the extremely smallest or the extremely largest value of a data set so we do the following two conditions for outliers:

- $\forall x \in X$; If $A_1(x)= 1$ or $A_5(x) = 1$ then x is an outlier.
- $\forall x \in X$; If $A_1(x)< 1$ or $A_5(x) < 1$ then x is a mild outlier.

IV. FIVE NUMBER SUMMARY OF DATA

Here we go with the real life problem that have been discussed below to show the capability of the five number summary of data method. To solve this type of method, the real world data sets have been used. Firstly, the five number summary of the data will be extracted. Next the fuzzy membership functions of the data sets will be defined using this. By plotting the membership function on the graph we can identify the outliers. To show that the results of the this method are effective, the box plot of the graph are drawn for the outliers.

Daily temperature intensity

The intensity of heat is not similar for all the other objects. Now let us consider water, water begins to boil at the temperature100˚C (Celsius) here the melting point of the uranium is 1132˚C. It is critical to find the whether it is warmest day or either the coldest day of the particular month according to its daily temperature. But the complexity in this is reduced due to the membership function. The daily maximum temperature (in ˚C) of Erode district for the month of May2020 36,37,37,38,38,38,36.35,38,38,37,37,37,38,37,34,37,29,34,35,34,35,36,37,37,37,36,37,33,34,35. By using the above given datasets, we need to check whether the days of May are cold or temperate or warn according to its temperature. Firstly, we need to arrange the values in the ascending order, then the data set of observations are:
X= {29,33,34,34,34,34,35,35,35,35,36,36,36,36,37,37,37,37,37,37,37,37,37,37,37,38,38,38,38,38,38}
No. of observation, n = 31 which is odd.

The minimum value, min $x_{min} = 29$

First quartile, $Q_1 = 35$

Median, $Q_2 = 37$

Third quartile, $Q_3 = 37$

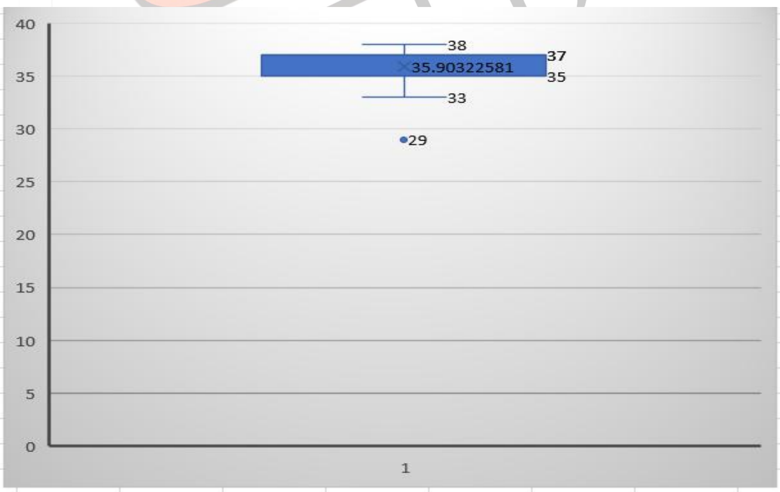
The maximum value, max $x_{max} = 38$

Inter Quartile Range, $I_{qr} = Q_3 - Q_1$=37− 35 35

Inner fence: $Q_1 - 1.5I_{qr} = 32$ and $Q_3 + 1.5I_{qr} = 40$

Outer fence: $Q_1 - 3I_{qr} = 29$ and $Q_3 + 3I_{qr} = 43$

The visual representation of the summary is given below.
FIGURE.2. Box plot of daily temperature.



Now, if we define the fuzzy sets, Now if we define fuzzy sets $T_1 X : \rightarrow [0,1]$, $T_2 X :\rightarrow [0,1]$, $T_3X :\rightarrow [0,1]$, $T_4X : \rightarrow [0,1]$ and $T_5X : \rightarrow [0,1]$ which represents the very low, low, medium, high and very high temperatures respectively, then $\forall x \in X$; then the five data set membership functions are given as follows:

$$\text{Very Low, } T_1(x) = \begin{cases} 1; & x \leq 29 \\ \frac{35-x}{3}; & 29 \leq x \leq 32 \\ 0; & x \geq 32 \end{cases}$$

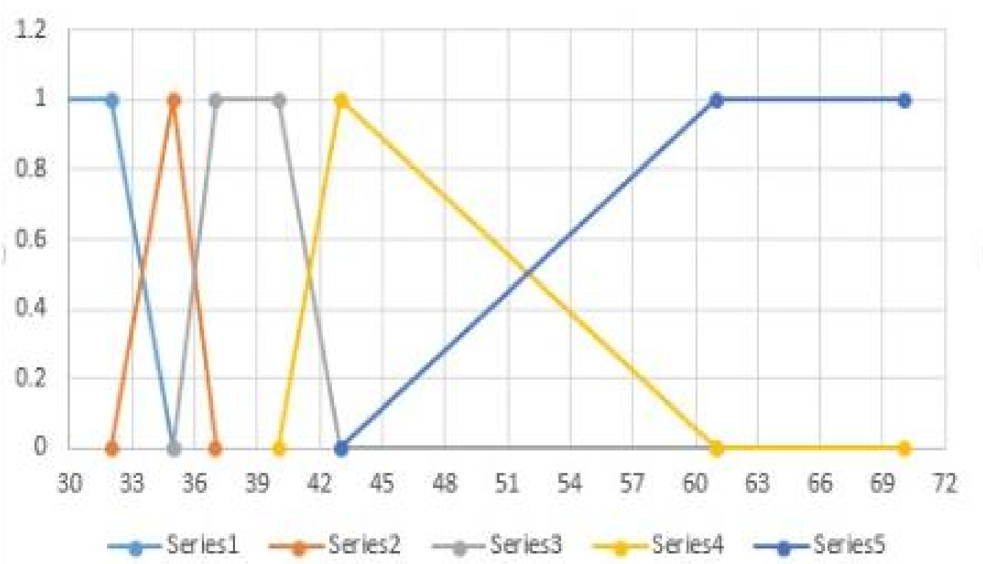
$$\text{Low, } T_2(x) = \begin{cases} 0; & x \le 29 \\ \frac{x-29}{3}; & 29 \le x \le 32 \\ \frac{35-x}{3}; & 32 \le x \le 35 \\ 0; & x \ge 35 \end{cases}$$

$$\text{Medium, } T_3(x) = \begin{cases} 0; & x \le 32 \\ \frac{x-32}{3}; & 32 \le x \le 35 \\ 1; & 35 \le x \le 37 \\ \frac{40-x}{3}; & 37 \le x \le 40 \\ 0; & x \ge 40 \end{cases}$$

$$\text{High, } T_4(x) = \begin{cases} 0; & x \le 37 \\ \frac{x-37}{3}; & 37 \le x \le 40 \\ \frac{43-x}{3}; & 40 \le x \le 43 \\ 0; & x \ge 43 \end{cases}$$

$$\text{Very high, } T_5(x) = \begin{cases} 0; & x \le 40 \\ \frac{x-40}{3}; & 40 \le x \le 43 \\ 1; & x \ge 43 \end{cases}$$

FIGURE.3. Membership function for daily temperature.



V. TYPES OF OUTLIER DETECTION METHODS AND ITS EXPLANATION:

Distance-based Detection Method

Index-based

Nested-loop

Local-outliers

Density-based Detection Method

Deviation based detection method etc.,

DISTANCE BASED METHOD

Data is indicated as a vector of the features.

Based on metric distance function it depends on the computation of the distance value.

The Distance-based anomalies are appropriately referred for k-dimensional datasets for any estimation of k [13].

An article O in the dataset T is a DB(p,D) exception if atleast portion p of the items in T are more prominent than or equivalent to the distance D from O..

The important approaches are

Nearest Neighbors based

Density based

Index Based

Neighbors of the every object O within radius D around that object can be searched by using the index. Once K (K = N(1-p)) neighbors of object O are found, O is not an outlier. worst case computation complexity is O(K*n2) ,dimensionality is K and number of objects is n in the dataset[12].

Nested-loop Method

The distances of instances to the all other objects to identify the instances k nearest neighbors can be calculated by the nested loop. Nested loop have the complexity O(kN2), (The no. of dimensions is K and the no. of data objects is N), and then no. of passes over dataset is linear to N. In the respect to no. of objects, making it deficient for mining in the large databases, due to quadratic complexity. It is major cost when it comes to the calculation of distance between the objects. For the high dimensionality datasets nested loop is the good choice, it unsuitable when there is large no. of calculation [12].

In this method Buffer space is separated into two halves (1st and 2nd arrays). Break the data into blocks and feed 2 blocks into arrays. Computes the distance between the each pair of objects directly, inside the array or between arrays decides the outlier.

Local Distance-Based Outlier Detection Factor (LDOF)

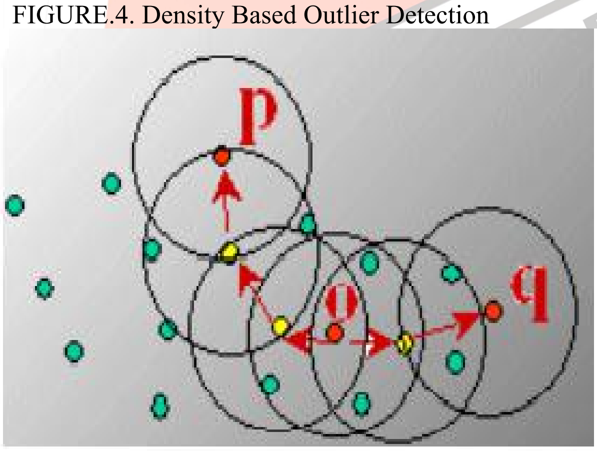
The outlier detection in the above schemes are an average when it involves in detecting the outliers in real world datasets. LDOF (Local Distance-Based Outlier Detection Factor) takes the relative distance from the object to its neighbors to calculate how much that the objects have been deviated from their scattered neighborhood. The point is like an outlier in the upper outlier factor. It has been noted that the outlier detection methods are more reliable in a top-n manner when it is utilized. This suggest that the top n factors have been taken as the outliers, then the n is taken by the user as convinent to the requirements[10].

DENSITY BASED METHOD

It depends on the nearby exception factor (LOF) of each point, which relies upon the nearby thickness of its neighborhood [7]. The area is characterized by the separation to the MinPts-th closest neighborhood. In commonplace use, objects with a high LOF are hailed as anomalies. Tang presented an connectivity based outlier factor (COF) conspire that improves the adequacy of LOF plot. The thickness around a outlier is significantly extraordinary to the thickness around its neighbors. By utilizing this property the outliers are distinguished.

```
DBSCAN(D, eps, MinPts) {
  C = 0
  for each point P in dataset D {
    if P is visited
          continue next point
 mark P as visited
 NeighborPts = regionQuery(P, eps)
 if sizeof(NeighborPts) < MinPts
mark P as NOISE
else {
 C = next cluster
 expandCluster(P, NeighborPts, C, eps, MinPts)
          }
        }
     }
```

FIGURE.4. Density Based Outlier Detection



DEVIATION BASED METHOD

In deviation based technique, given a bunch of information focuses Outliers are focuses that don't fit to the overall attributes of that set, Therefore, the variation of the dataset is limited while eliminating the anomalies. In deviation based technique recognize the anomalies by investigating the attributes of items and consider an articles that digresses these highlights as an anomaly deviation based outlier detection doesn't utilize any factual tests or separation based measures to distinguish outstanding items, Instead of, it distinguishes the outliers by analyzing the fundamental qualities of the objects in a group. Articles which 'deviate' altogether from this depiction are spoken to as outliers. Consequently in this methodology the term deviation is ordinarily used to refer to the outliers.

VI. CONCLUSION

An outlier is a perception which deviates most from the rest of the other perceptions as to stimulate doubts that it was created by an alternate system .There are numerous variety of application which is used for the outliers detection. Five number summary of data method is proficient technique for generating the membership functions and rectifying the outliers that is discussed in this paper. By applying it, the solution of some real life related problems has been highlighted. The problem is presented graphically and their results are also described. The results utilized are excellent and provides adequate information of the data set. Then there are different methods utilized in outlier detection like distance based, deviation based, density based have been examined.

REFERENCES

[1] Hasan, M.F. and Sobhan, M.A. (2020) Describing Fuzzy Membership Function and Detecting the Outlier by Using Five Number Summary of Data. American Journal of Computational Mathematics, 10, 410-424.https://doi.org/10.4236/ajcm.2020.103022

[2] Study of Distance-Based Outlier Detection Methods by Jyoti Ranjan Sethi A, thesis submitted in partial fulfillment for the degree of Bachelor of Technology, Under the guidance of Prof. Bidyut Kumar Patra Computer Science and Engineering National Institute of Technology Rourkela Rourkela-769 008, India. (2013)

[3] Proceedings of 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE). Dr SUVANAM sasidhar Babu, Sree Narayana gurukalam college of engineering, Methods to detect different type of outlier. https://www.researchgate.net/publication/311610830

[4] New Methods for Deviation-Based Outlier Detection in Large Database. Zhiyuan Zhang and Xia Feng Civil Aviation University of China (2009). https://www.researchgate.net/publication/221088779

[5] Gustavo Henrique Orair, Federal University of Minas Gerais, Wagner Meira Jr. Federal University of Minas Gerais, Presented by Kajol UH ID : 1358284 Distance-Based Outlier Detection: Consolidation and Renewed Bearing.

[6] George, J.K. and Bo, Y. (1995) Fuzzy Sets and Fuzzy Logic: Theory and Applications. PHI New, Delhi, 443-455.

[7] M. Breunig, Hans-Peter Kriegel, Raymond T. Ng‡, Jörg Sander(2000) Local Outlier Factor:IdentifyingDensity-Based Local Outliers,Markus.

[8] Ling, W.K. (2007) Nonlinear Digital Filters: Analysis and Applications. Academic Press, Cambridge.

[9] Thirumalai, C., Vignesh, M. and Balaji, R. (2017) Data Analysis Using Box and Whisker Plot for Lung Cancer. 2017 Innovations in Power and Advanced Computing Technologies, Vellore, 21-22 April 2017, 1-6. https://doi.org/10.1109/IPACT.2017.8245071

[10] Ke Zhang, Marcus Hutter, and Huidong Jin. A new local distance-based outlier detection approach for scattered real-world data. In Advances in Knowledge Discovery and Data Mining, pages 813–822. Springer, 2009.

[11] Zadeh, L.A. (1965) Fuzzy Sets. Information and Control, 8, 338-353. https://doi.org/10.1016/S0019-9958(65)90241-X

[12] Edward Hung and David W Cheung. Parallel mining of outliers in large database. Distributed and Parallel Databases, 12(1):5–26, 2002.

[13] Edwin M. Knorr and Raymond T. Ng. Algorithms for mining distance-based outliers in large datasets. In Proceedings of the 24rd International Conference on Very Large Data Bases, volume VLDB '98, pages 392–403. Citeseer, 1998.

[14] Klir, G. and Yuan, B. (1995) Fuzzy Sets and Fuzzy Logic (Vol. 4). Prentice Hall, Upper Saddle River. https://doi.org/10.1109/45.468220

[15] Vignesh, V., Pavithra, D., Dinakaran, K. and Thirumalai, C. (2017) Data Analysis Using Box and Whisker Plot for Stationary Shop Analysis. 2017 International Conference on Trends in Electronics and Informatics, Tirunelveli, 11-12 May 2017, 1072-1076. https://doi.org/10.1109/ICOEI.2017.8300874

[16] Le, C.T. and Eberly, L.E. (2016) Introductory Biostatistics. John Wiley & Sons, Hoboken.