Outlier Detection: A Research and Modified Method Using Fuzzy Clustering

S. Rajalakshmi, P. Madhubala

Abstract: ***Data mining is becoming increasingly popular in many application fields. Due to the advancement, Researchers show great interest to find unexpected behaviour over large amount of datasets. Outlier detection is studied extensively in data mining and developed for certain application domains, while others are generic in nature. It is one of the important and hottest topic in research which faces a series of new challenges. It occurs due to change in system behaviour, mechanical fault, human error, natural deviations and instrumental error. The purpose of this paper briefly provides a survey on outlier detection and a modified approach to detect outlier using Fuzzy clustering. Also, it provides a better understanding of different dimensions that applied in various substantive areas.***

Keywords : Data Mining, Fuzzy Clustering, Outlier Detection

I. INTRODUCTION

Outlier are patterns that donot conform to a well defined notion of normal behaviour. Outliers are also called as anomalies, abnormality, deviants, discordants, extreme data points and so on. It reveals useful information related to abnormal characteristics and the terms Outlier and Anomaly are used interchangeably. The factors makes the approach very challenging a)to define the boundary between normal and anomalous behaviour b)validation of models c)difficult to distinguish noise. In "Fig. 1", The clusters are N1and N2 and Outliers are O1,O2 and O3.

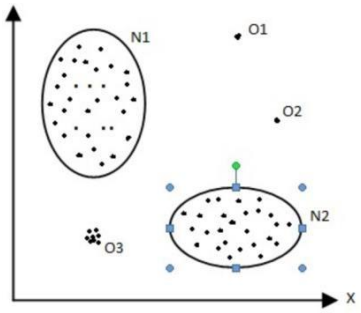

Fig 1:Identifying Outliers

Outlier detection and the notions of prediction are related intimately. The different aspects of outlier detection are used for various domains to detect intrusion, to acknowledge fraudulent in credit-card transactions, detecting events in system network traffics, diagonising the medical reports, enforcing the law amendment, accelerating the damage in industrial sector, figuring the detection of textual anomaly and so on. Session 2 includes the survey of related work, Session 3 includes why the outlier is problematic, Session 4 includes applications, Session 5 includes outlier methodologies, Session 6 includes clustering and fuzzy clustering, Session 7 includes the experimental results and discussion and conclusion concluded by session 8.

II. RELATED WORK

Initial study of outlier framework is studied by Hawkins [2]. Review of outliers and noise accommodation is clearly understood by Rousseeuw and Leroy[5] to immunizing the statistical model of estimation. Outlier analysis referred with supervised model is carried out detecting in real time is presented by Markou and singh[17]. Topic detection and streaming of text tracking is also studied. An excellent review of outlier detection is covered by Chandola et al.,[25], K.Ramamohanarao, R.Zhang, H.Wang, X.Liu, J.Bailey and X.Wu [27]. The key modeling on regression based statistical technique are studied by T.DeVries, M.Houle and S. Chawla [28]. Mining outliers for large data sets is studied from R.Rastogi, K.shim and S.Ramaswamy[13]. k-nearestneighbour using graph has been studied from H.Ville, I.Karkkainen and P.Franti[20].

Erich Schubert, Arthur zimek, Hans peter kriegel presents the subspace(subset of attributes) approaches, being so-called "effect on the concentration of distance", and "dimensionality towards the curse" and specialized algorithms for outlier detection[34].

Temporal Outlier detection has been studied from Charuaggarwal[35], where his initial work is started on time series, and later in multiple data streams, spatio-temporal data, network data, community distribution data has been studied. The statistical technique approaches of temporal outlier detection includes network based, density based, spatio-temporal are studied.

Fast means of outlier detection in high dimensional spaces is studied from C.Pizzuti and F.Angiulli[16]. A pattern based outlier detection is studied from H.Jin and K.Zhang [29].

The challenges of Outlier Ensembles for unsupervised outlier detection has been studied from C.C.Aggarwal[35]. Outlier ensembles on algorithms of theoretical foundations are also studied from C.C.Aggarwal [35]. Feature bagging concept is studied from A.Lazarevic and V.Kumar[23]. Subsampling of unsupervised outlier detection ensembles is also undetstood from the study.

Initially fuzzy clustering studied by Dunn in 1973 and finally generalized by Bezdek in 1981. Fuzzy set theory by Zadeh[10] who reveals uncertainity, which belongs to a membership function, Fuzzy partitioning and Applications

Revised Manuscript Received on January 2, 2020.

S. Rajalakshmi, Department of Computer Science, Government Arts College for women, Krishnagiri, TamilNadu, India. Email: rajaylakshmiravi7@gmail.com

Dr. P. Madhubala, HOD, Department of Computer Science, DonBosco College, Dharmapuri, Tamilnadu, India. Email: madhubalasivaji@gmail.com

is studied from Bezdek and Harris[1]. Fuzzy clustering comprises three categories that are based on fuzzy relation, objective function and KNN are studied from M.S.Yang[7][8][9]. ISODATA process and its fuzzy relations studied in M.A.Ismail, J.C.Dunn, S.A.Selim and J.C.Bezdek [4], fuzzy relations and fuzzy partitioning are studied in J.C.Bezdek and J.D.Harris[1], pattern classifications studied in M.Roubens[11], fuzzy c-means clustering algorithms studied in, J.C.Bezdek, K.F.Yu, J.Dave, R.Cannon, M.S.Yang [1][7][8][9]. Hoppner et.al made a good framework of FCM survey[12].

III. WHY OUTLIERS ARE PROBLEMATIC?

"All the outliers will not be the intruders". They affect regression line in different ways such as appearing typical predictor value, unusual predictor value following line and unusual predictor value not following line. Incorrect entry, Mis reporting, sampling error and exceptional but true value are some possible causes of outliers[3][4]. Some common ways to deal with outliers are a)Discarding b)Winsorizing c)Variable transformation d)Fit different models e)Dropping but keeping in record f)Use Non-Parametric methods. A point different from all other points called outliers which is useful data to get desired analysis.[20][5]

IV. APPLICATIONS

Fraud Detection, Detecting the network intrusion, Monitoring the system activity and network performance, analyzing the images of the Satellite, detecting the Novelty, Monitoring the series of time, Diagnosing the medical report, Criminal actions in e-commerce, Video Surveillance, Anti-Terrorism, Pharmaceutical research, Data Leakage Prevention.

V. OUTLIER METHODOLOGIES

To alleviate the outlier problem, the detection technique has been classified into a) Detection of neighborhood based b)Detection of Subspace based c)Detection of Ensemble based d)Detection of Mixed type [27][10]. In univariate, it is based on clustering, distance, distribution, and density.

A. Neighbourhood Based Detection:

Detection based on independent data objects. The outlier score is defined by average or weighted distance(KNN). It is measured by LOF(Local Outlier Factor). To identify similar neighbour rank based detection algorithm is used[9][14]. He and Tang formulated three types of neighborhoods, such as shared, reverse and k-nearest. This leads to random sampling procedure, inlier scores as observabilityfactor(OF) where k is assigned as an entropy and estimated for its occurrence times.

B. Subspace Based Detection:

It detection is manipulated by shifting different subsets of high dimensional data. Relevant subspace and Sparse subspace are the types for representing the subspace[27][10]. Aggarwal proposed a evolutionary algorithm to improve exploration efficiency. Detection of correlation dataset deviates the subspace and so called SOD.

C. Ensemble Based Detection: (Outlier ensembles):

Ensemble, which is also called meta algorithm, combines the output of multiple algorithm. It is widely used in many data mining techniques such as clustering,association rules and classification [25][15][6][7]. Ensemble methods thus used to develop clustering quality. The methods are a)boosting b)bagging c)stacking d)subsampling. Ensemble analysis is used to reduce the robustness of specific datasets of Data mining problem. It is regularly used in Clustering and Classification techniques.The outlier scoring algorithm results on dependent and independent execution of data mining algorithm. The final result of Boosting technique in classification depends on the derivation of each model and ensemble score. Theoretical formulatives of outlier ensemble analysis are identified using effective process of subsampling and feature bagging techniques[19][8]. According to C.C.Aggarwal, algorithmic patterns of outlier detection ensembles and complementary we focussed on core ingredients of outlier detection ensembles[17][9]. i)Behaving model of specific dataset. ii) Ensemble clustering and multiview clustering iii)Randomly clustering the forest to the extended level. iv)In classification, stacking, random forest, boosting, bagging, has been considered for effective results[23][3].

D. Mixed Type Detection:

This method shows the strategy of discretizing numerical features in to categorical data. He et al handles the categorical data of frequent pattern based technique of some real examples which includes RELOADED and LOADED so on.

VI. CLUSTERING

Clustering identifies the similar groups of data in a given dataset. Clustering is an unsupervised learning technique where it is subjective in nature.

A. Fuzzy Clustering:

Fuzzy clustering is a soft clustering, which tends to a straightforward implementation with robust behaviour makes the model uncertain. It is developed by Jim Bezdek in 1981. It detects known variants using membership [0,1] and prototype matrix for each member in the dataset. It is evaluated using Euclidean distance and statistical properties of clusters.The range between 0 and 1 is called as membership grade (or) degree of membership.

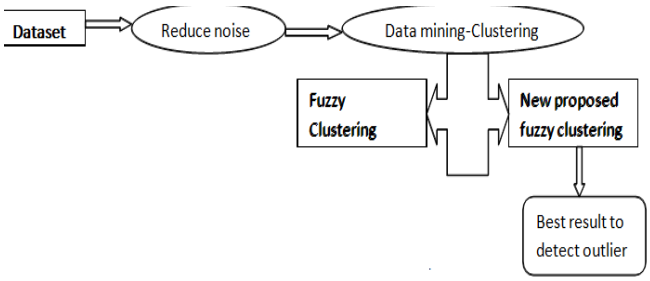

Fig 2 : Basic structure to detect outliers

Consider a set S, subset A(x) and elements of set S is defined by,

A(x) = { 1, if membership degree is ≤ subset value
0, if membership degree is > subset value }

The fuzzy membership lies between 0 and 1. It assigns membership to each data point and the distance is calculated between the data point and center of cluster. The data objects has high degree of membership at the centroid center rather than the borders. Each data point has corresponding membership value one. Dataset should always be unity as shown in (1)[14][24], fuzzy membership iteration in (2), fuzzy center in (3)

$$\sum_{i=1}^{c} u_{ij} = 1, \forall j = 1,\ldots,n \quad (1)$$

$$\mu ij = 1/\sum_{k=1}^{c} (dij/dik)^{(2/m-1)} \quad (2)$$

$$vj = (\sum_{i=1}^{n} \mu ij)^{m} xi)/(\sum_{i=1}^{n} (\mu ij)^{m}), \forall j = 1,2,\ldots..c \quad (3)$$

where, *'n'* - data points

'vj' - *jth* center of cluster

'm' - index of fuzziness showing m € [1, ∞].

'c' - cluster center

'μij' - membership

'dij' - Euclidean distance

By minimizing the fuzzy clustering by,

$$J(U,V) = \sum_{i=1}^{n} \sum_{j=1}^{c} (\mu ij)^{m} || xi - vj ||^{2} \quad (4)$$

'||xi – vj||' - Euclidean distance.The objective function for FCM

$$J(U, c_{1,\ldots..}c_c) = \sum_{i=1}^{c} J_i = \sum_{j}^{c} \sum_{j}^{n} u_{ij}{}^{m} d_{ij}{}^{2} \quad (5)$$

The fuzzy clustering determines the cluster center *ci* and membership matrix of U by the following steps

Step 1: Initializing the membership matrix - U

Step 2: selecingt fuzzy cluster centers[ci], where i=1….c.

Step 3: Calculating fuzzy membership 'μij' and fuzzy centers 'vj' to compute the objective function[OF].

Step 3a: Stop, until the iteration value is below certain threshold value ε.

The cluster center of grouping fuzzy '*i*' is given by the equation(6)[12][25]

$$c_i = \sum_{j=1}^{n} u_{ij}^{m} x_j / \sum_{j=1}^{n} u_{ij}^{m} \quad (6)$$

Membership matrix U for each iteration is given by equation (7)[12][25]

$$u_{ij} = 1/\sum_{k=1}^{c} (d_{ij}/d_{kj})^{2/(m-1)} \quad (7)$$

.

Euclidean distance is given by equation (8)[12][25]. It measures the compactness (goodness) of the clustering

$$d_{ij}. = || c_i - x_j || \quad (8)$$

B. Algorithm:

Execute FCM to generate objective function [OF]
Determine small clusters to identify outliers.
Rest of the points (not in step 2)
Begin
Sum=0.5
For each point Pi in the set Do
Remove a point Pi
calculateOFi
Calculate DOFi=(OF-OFi)/2
Sum=Sum + DOFi
Return Pi
End do
Avg DOF=sum/n
For each point Pi
Do
If (DOFi> T) classify point pi as an outlier and return
Else stop
End do
End

C. Key Points:

1) Fuzzy clustering uses Euclidean distance metric to compute sum of squared distance between cluster centers and data objects

2) Dividing OF by N(no of data points in a set) produces AVG DIST(AD)

3) Calculate OFi,after removing a point from the set. Subtract DOFi=(OF-OFi). Multiply AD by 1.5 to produce threshold T which determines outliers

4) Datapoint belong to cluster depends on the degree of membership between 0 and 1.

5) The dataset degree for the fuzzy related sum is equal to unity.

Before clustering, i) One should know how many clusters to be specified. If not, it is problematic. ii) Before writing the algorithm, one should know how to optimize(k=3) and solve a problem. iii) Know your data on what basis it should group into no of clusters. Partitioning the collection of FCM data points xi, i=1...n for FCM data points divides the cluster centers in to fuzzy groups.

VII. RESULT AND DISCUSSION

In this section, we implemented using R programming version 3.6.1, by installing and loading the basic package 'ppclust' (july-23-2019). Using iris dataset(Anderson, 1935) and stock_data dataset, our method shows better results at the accuracy of 81.66%. It also works well for the dataset of advertising, airport noise monitoring, auto, directory of food stamps, directory of job centers, dycd-contractors, heart, jobs-proximity-index, linked in topskills, nyc-jobs and data-user-modeling.

Table – I : Data set(stock_data)

No of Iterations	Fuzzy Membership Matrix		
	Cluster 1	*Cluster 2*	*Cluster 3*
1	0.8870714	0.03874533	0.07418328
2	0.8870843	0.03879425	0.07412141
3	0.8872636	0.03872362	0.07401278
4	0.8921110	0.03690018	0.07098878
5	0.8949061	0.03582245	0.06927145
6	0.8952708	0.03566608	0.06906314
2996	0.05295371	0.7721128	0.1749334
2997	0.05641268	0.7618989	0.1816884
2998	0.05674812	0.7613694	0.1818825
2999	0.05818080	0.7571338	0.1846854
3000	0.06170653	0.7458116	0.1924819

Table – II : Descriptive statistics for the membership degreesby clusters

Size	Data set : stock_data					
	Min	*Q1*	*Mean*	*Median*	*Q3*	*Max*
C1 1405	0.4610	0.8968	0.9041	0.9562	0.9709	0.9819
C2 789	0.4611	0.7277	0.7799	0.7885	0.8582	0.9469
C3 806	0.4545	0.5754	0.7313	0.7223	0.8940	0.9636

Table –III : Detecting Outlier using tests

Class ID	Outlier detected for Data set : stock_data		
	't'	*'chisquare'*	*'z'*
1	TRUE	FALSE	FALSE
3	TRUE	TRUE	TRUE
5	TRUE	TRUE	TRUE
7	TRUE	TRUE	TRUE
12	TRUE	TRUE	TRUE
13	TRUE	TRUE	TRUE
14	TRUE	TRUE	FALSE
17	TRUE	TRUE	TRUE
18	TRUE	TRUE	TRUE

VIII. CONCLUSION

This paper presents an extensive overview of outlier detection over huge literature survey on various techniques. The test result of our modified approach gives good result for different data sets. Among the three test, 't' test detects the outlier efficiently. Outlier detection methodologies for temporal data, mixed data, high dimensional data and their challenges are discussed. In future, the work is focused on ensemble based outlier detection.

ACKNOWLEDGMENT

I am very thankful to my research Supervisor Dr. P. Madhubala for her continuous support and encouragement to bring this research work successful. She is currently working as Head of the department in computer science at Dharmapuri - Don Bosco college, Tamilnadu , India.

REFERENCES

1. J D Harris, J C Bezdek, "Fuzzy Partition and relation - Aan axiomatic basis for clustering", Fuzzy sets and systems, Elsevier, 1978.
2. D.M. Chapman and Hall, Hawkins, "Identification of Outliers", Springer, 1980.
3. T. Lewis and Barnett, "Outliers in Statistical Data ", Wiley & Sons, 1984.
4. Mohamed A.Ismail, Shokriselim, "On the local optimality of the fuzzy isodata clustering algorithm", IEEE, 1986.
5. A.M.Leroy, P J. Rousseeuw, "Robust regression and outlier detection", 1987.
6. Prof .GourabNath, "Outlier Analysis" – PPT.
7. 7. K.F.Yu and M.S.Yang, "On stochastic convergence theorems for the fuzzy c-means clustering procedure",1990.
8. K.F.Yu and M.S.Yang , "On existence and strong Consistency of fuzzy c-means clustering procedures", 1992.
9. M.S.Yang, "A survey of fuzzy clustering", Pages 1-16, December 1993.
10. LoftiA.Zadeh, "Fuzzy sets, Fuzzy logic and Fuzzy system", Advances in Fuzzy Systems, Applications and theory,Volume 6,1996
11. Roubens, "Fuzzy sets and decision analysis". Fuzzy Sets and Systems, Volume 90, Issue 2, pages 199-206, September 1997
12. F Hoppner, R Kruse, F Klawonn, , T Runkler, "Fuzzy cluster analysis:methods for classification, data analysis and image recognition", John wileyans sons, 2000.
13. S.Ramaswamy, K.shim and R.Rastogi, "Efficient Algorithm for mining outliers from large data sets", ACM SIGMOD, Volume 29, Issue 2, Pages 427-438, June 2000.
14. Mohamed N. Ahmed, A. Farag, Sameh M. Yamany, Nevin Mohamed, Aly and Thomas Moriarty, "A Modified Fuzzy C - Means Algorithm For Bias Field Estimation and Segmentation of MRI Data", IEEE Transactions On Medical Imaging, Vol. 21, No. 3, March 2002.
15. R. J. Hathaway, James C. Bezdek, "Fuzzy C-Means Clustering of Incomplete Data", IEEE, Oct 2001.
16. C.Pizzuti and F.Angiulli, " Fast Outlier Detection in High Dimensional Spaces", European Conference on principles of data mining and knowledge discovery, Sep 2002.
17. Sameer Singh and MarkosMarkou , "Novelty Detection:a Review – part 1:statistical approaches", Elsevier-2003.
18. Petrovskey, "Outlier Detection Algorithms in Data Mining Systems " - M.I., Programming and Computer Software, Vol.29, No.4, pp.228–23, 2003.
19. V.J., Kluver, Hodge, , "A survey of outlier detection methodologies " , Academic Publishers, Netherlands, January 2004.
20. P.Franti, I.Karkkainen, Ville Hautamaki,"Outlier detection using k-nearest neighbour graph", IEEE,2004Roubens, "Fuzzy sets and decision analysis". Fuzzy Sets and Systems, Volume 90, Issue 2, pages 199-206, September 1997
21. Jim Austin and Victoria J. Hodge, "A Survey of Outlier Detection Methodologies", Artificial Intelligence Review, Vol 22, Issue 2,pp 85-126, October 2004.
22. Parsons L, E. H. Liu, Haque, , "Subspace Clustering for High Dimensional Data : A Review ", SIGKDD Explorations, Vol. 6 (1), pp.90 – 105, 2004.
23. V.Kumar, A. Lazarevic, "Feature Bagging for Outlier Detection", ACM SIGKDD International Conference on knowledge Discovery and Data Mining, January 2005.
24. M.Ester, R.Frank, W.Jin, "Efficiently Mining Regional Outliers in Spatial Data",Springer-2007
25. V.Chandola,,V.Kumar and A.Baneerjee "Anomaly detection:A Survey", ACM Computing surveys, pages1 - 72, 2009.
26. Yang D, E.A.Rundensteiner, M.O.Ward, "Neighbor based pattern detection for streaming data in windows", acm - 2009.
27. X.Liu, X.Wu, H.Wang, J.Bailey, R.Zhang, and K.Ramamohanarao, "Mining distribution change in stock order data streams", ICDE Conference, 2010.
28. T.DeVries, M.Houle and S. Chawla, "Finding local anomalies in very high dimensional space", ICDE Conference, 2010.
29. H.Jin and K.Zhang, "An Effective Pattern Based Outlier Detection Approach for Mixed Attribute Data", Australasian Joint conference on Artificial Intelligence, November 2010.

30. M. Georgiopoulos and Koufakou, A., "A Fast Outlier Detection Strategy For Distributed High - Dimensional Data sets with Mixed Attributes", Data Mining and Knowledge Discovery", Vol.-20. pp. 259–289, 2010.
31. M. Tkacz and Chrominski, K., "Comparison of outlier detection methods in biomedical data", Journal of medical Informatics & Technologies, vol.16, ISSN 1642-6037, 2010.
32. Moh'dBelal Al-Zoubi, Al-Dahoud Ali,.Yahya, Abdelfatah A "New Outlier Detection Method Based on Fuzzy Clustering" – WSEAS Transactions On Information, ISSN: 1790-0832 , Issue 5, Volume 7, May 2010.
33. Sanjay Koti.M and Srimani.P.K, "Application of Data Mining Techniques For outlier mining in Medical Databases " , International Journal of Current Research, Vol. 33, No. 6, pp.402-407, 2011.
34. Arthur Zimek, Hans-Peter Kriegel, Erich Schubert , "A survey on unsupervised outlier detection in high-dimensional numerical data", Wiley, 27 August 2012.
35. CharuC.Aggarwal, "Outlier Analysis", second edition, 25 Nov 2016.

AUTHORS PROFILE

Dr.P.Madhubala, Research Supervisor has 20 years of experience is currently working at Don Bosco college, Dharmapuri. She published many international journals and articles in the field of cloud computing and Data mining.

S.Rajalakshmi, Research Scholar is currently working in Govt Arts College for women, Krishnagiri under Periyar University, Salem, Tamilnadu, India has 8 years experience, interested in Data Mining techniques published 5 journals.