Outlier Analysis Approaches in Data Mining

Krishna Modi[1], Prof Bhavesh Oza[2]

[1,2]*Computer Science and Engineering L D Collage of Engineering Ahmedabad, Gujarat, India.*

***Abstract*—Data Mining is used to the extract interesting patterns of the data from the datasets. Outlier detection is one of the important aspects of data mining to find out those objects that differ from the behavior of other objects. Finding outliers from a collection of patterns is a popular problem in the field of data mining. A key challenge with outlier detection is that it is not a well expressed problem like clustering so Outlier Detection as a branch of data mining requires more attention. Outlier Detection methods can identify errors and remove their contaminating effect on the data set and as such to purify the data for processing. Detecting outliers and analyzing large data sets can lead to discovery of hidden knowledge in area such as fraud detection, Terrorism Activities, telecommunication, web logs, and web document, etc. In this paper, we explained five types of outlier, different approaches to detect outliers, their advantages and disadvantages and applications.**

***Index Terms*—Data Mining, Outliers, Anomalies, Supervised**

I. INTRODUCTION

Outlier analysis is used in various types of dataset, such as graphical dataset, numerical dataset, Text dataset, and can also be used on the pictures etc. The identification of outlier can lead to the discovery of useful and meaningful knowledge. Outlier detection is the process of finding data objects with behaviors that are very different from expectation. Such objects are called outlier or anomalies. Finding outliers from a collection of patterns is a popular problem in the field of data mining. A key challenge with outlier analysis and detection is that it is not a well formulated problem like clustering so Outlier Detection as a branch of data mining requires more attention.

Outlier analysis and detection has various applications in numerous fields such as fraud detection, credit card, discovering computer intrusion and criminal behaviors, medical and public health outlier detection, industrial damage detection etc. General idea of these application is to find out data which deviates from normal behavior of data in dataset. For that they use any of the technique to detecting outlier suitable on their dataset.

In these paper we explained different approaches and most used techniques that is useful to analyze outliers. Rest of this paper is organized as follows. Section II describes various definitions of outlier which give you more deep idea about outlier. Outliers are generally classify into Global outlier, Contextual Outliers, Collective outliers, Real outliers and Erroneous outliers that is well explained in Section III. In section IV we explained different approaches of outlier detection and its useful methods. In V section we put comparison table of different approaches, their advantages and disadvantages and their appropriate applications. At last Section VI concludes with summary of those outlier detection approaches.

II. OUTLIERS ARE DEFINED IN SEVERAL WAYS LIKE

- An outlier is a pattern which is dissimilar with respect to the rest of the patterns in the data set.
- Outlier is defined as an object that deviates from other objects. [42]
- Outlier is an outlying observation.
- Outlier is one that appears to markedly from other members of the sample in which it occurs.
- Observation that deviates so much from other observations. [42]
- A database may contain data objects that do not comply with The general behavior or model of the data. These data objects are outliers.
- Outliers are patterns in data that do not conform to a well-defined notion of normal behavior. [41]
- Data point which is very different from the rest of the data based on some measure. [34]
- An outlier can be explained as every information which occurs to be different or dissimilar with reverence to the remaining data. [1]
- Definition proposed by Barnett and Lewis that "an outlier is an observation which appears to be

inconsistent with the remainder of that set of data.” [1]

- In data mining outlier detection refers to the recognition of data point which does not follow the expected pattern or behavior in a particular dataset or is significantly different from other points in a data. [3]
- The background noise in which the clusters are embedded but recent literature defines the outliers as the points that are neither a part of the cluster nor a part of the background noise; rather they are specifically the points that are very much different from the norm.

III. TYPES OF OUTLIER

Global Outliers

If an individual data instance can be considered as anomalous with respect to rest of the data, then the instance is termed as a point outlier. It is one of the simplest forms of outliers and is the key focus of many outlier detection researches. Global outlier sometimes called point anomalies or point outliers.

Contextual Outliers

If a data instance is a rare occurrence with respect to some specific context and it is a normal occurrence with respect to some another context, then such types of data instances are known as contextual data sets.

Collective Outliers

If an individual data instance is not anomalous but its collection with the entire dataset is anomalous, then it is termed as a collective outlier, the individual data instances which are termed as collective outlier may not be the outlier themselves but their occurrence together as a collection is anomalous, hence it is a collective outlier.

Real Outliers

Observation that help to find the analyst something new and innovative and if they are removed anyhow, we are completely left with the normal region. As they are real outliers, Can’t remove as a noise. Real outliers are subject of interest for system analyst.

Erroneous Outliers

If some observation is noted incorrectly as an outlier, due to some inherent problem, or some catastrophic failure, then these are by mistake outliers or we can say illusive outliers. They really take the outcome of the data in some other way. As they are erroneous outliers, remove them as a noise.

IV. DIFFERENT APPROACHES FOR OUTLIER ANALYSIS

Clustering based Approach

Clustering-based approaches detect outliers by examining the relationship between objects and clusters. An outlier is an object that belongs to a small and remote cluster, or does not belong to any cluster.

Approach we can use to find:

a. Detecting outliers as objects that do not belong to any cluster. [Using a density based clustering method, Ex. DBSCAN]

b. Clustering-based outlier detection using distance to the closest cluster.[Using K mean which assign an outlier score to the object according to the distance between the object and the center that is closest to the object.]

   In this approach, similarity between two objects is measured with the help of distance between the two objects in data space, if this distance exceeds a particular threshold, then the data object will be called as the outlier.

c. Detecting outliers in small clusters.[Using CBLOF / fixed-width clustering]

   The CBLOF score can detect outlier points that are far from any clusters. Small clusters that are far from any large cluster are considered to consist of outliers. The points with the lowest CBLOF scores are suspected outliers.

d. Organize objects into partitions where each partition represents a cluster. [PAM, CLARA, CLARAN are popular partition based methods]

Classification Approach

Outlier detection can be treated as a classification problem if a training data set with class labels is available. The general idea of classification-based outlier detection methods is to train a classification model that can distinguish normal data from outliers.

Due to the number of normal samples likely far exceeds the number of outlier samples, the training set is typically heavily biased which prevent us from building an accurate classifier.it is infeasible to enumerate all potential intrusions, as new and unexpected attempts occur from time to time so to overcome it, classification-based outlier detection methods often use a one-class model (a classifier is built to describe only the normal class other class consider as outlier).

Support Vector Machine (SVM), Bayesian classification are well-known methods for Classification Approach.

Statistical Approach

Statistical approaches were the oldest algorithms used for outlier identification. The statistical approach assumes that data follows some standard or predefined distribution or probability model, and aims to identify outliers with respect to the model using a discordance test (those outliers which do not follow such distributions). A discordance test is used to detect whether a given object is an outlier or not.

The general idea behind statistical methods for outlier detection is to learn a generative model fitting the given data set, and then identify those objects in low-probability regions of the model as outliers. Statistical methods performs poorly on high-dimensional data.

Statistical methods for outlier detection can be divided into two major categories:

a. Parametric methods

Model using parametric technique grow only with model complexity not data size. Assumes that the normal data objects are generated by a parametric distribution.

Regression, scatter-point method are popular parametric method. This figure shows scatter-point method to detecting outliers.

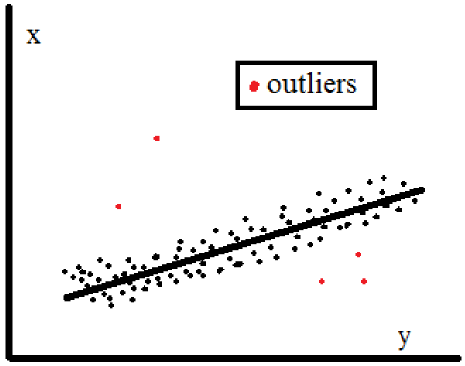

Figure 1 Scatter-point Method to detect outlier

b. Nonparametric methods

The model of normal data is learned from the input data, rather than assuming one a priori. It do not make any assumption about statistical distribution of data.

Histogram, Kernel density function, Kernel feature space are methods of non-parametric techniques.

Frequency based Approach [18]

We can use Statistical Approach, Density-based Approach and Deviation-based approach for numerical dada. But when data is categorical, we have to map those data into numeric values, but it is not a much easy for all categorical data. Frequency-based approaches have been defined to detect outliers in categorical data.

Proximity based Approach

Proximity-based approaches assume that the proximity of an outlier object to its nearest neighbors significantly deviates from the proximity of the object to most of the other objects in the data set.

There are three types of proximity-based outlier detection methods:

a. Density based Approach

The density-based approach of classic outlier finds the density allocation of the information and identify outliers as those present in low-density region. Partitioning methods are designed to find spherical-shaped clusters but it is difficult to find clusters of arbitrary shape such as ‘S’ and oval. Density-based methods overcome this limitation.

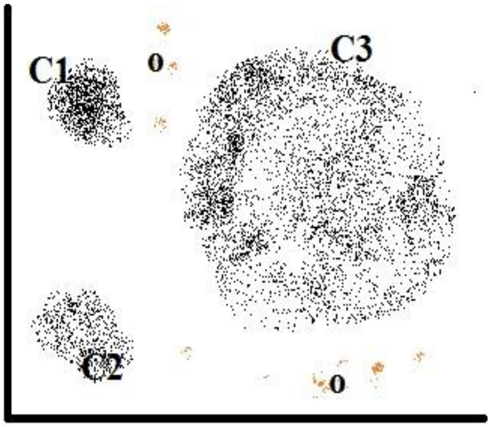

Figure 2 Density based approach

This method compares the density around a point with its local neighbor’s densities. The relative density of a point compared to its neighbors is computed as an outlier score. Density based outlier detection method uses density distribution of data points within data set. Breunig et al. [6] allocate a local outlier factor (LOF) to every point based on the neighboring density of its environs. LOF value of an object is based on the average of the ratios of the local reachability density of the area around the object and the local reachability densities of its neighbors. The size of the neighborhood of the object is determined by the area containing a user-supplied minimum number of points (MinPts). DBSCAN, OPTICS, DENCLUE methods are used for density-based Approach. In this figure C1, C2, C3 stands for clusters and O for outliers.

b. Distance based Approach

Distance based outlier analysis is one of the most used and widely accepted technique that used in data mining and machine learning. It completely depends on the concept of local neighborhood of data points. This concept is also termed as Nearest Neighbor analysis and it can also applied for different purposes such as classification, clustering and most importantly outlier analysis. [2]

In distance based approach similarity between two objects is measured with the help of distance between the two objects in data space, if this distance exceeds a particular threshold, then the data object will be called as the outlier. To measure distance any appropriate distance measure can be used such as Manhattan distance, Euclidean distance, Mahalanobis distance, or some other measure of dissimilarity. In this approach most-used algorithm is K-means algorithm which use Euclidean distance to find outliers. K-Medoids algorithm and kNN algorithms are also types of distance based approaches.

c. Grid based Approach

Grid-based method quantizes the space into a finite number of cells which form a grid structure on which all the operations for clustering are performed. Main advantage of using this method is fastest processing time that depends only on size of grids instead of data size. For problems with highly irregular data distribution, the resolution of grid must be too fine to obtain good clustering quality.

CELL is a grid-based method for distance-based outlier detection. CELL method organizes objects into groups using a grid—all objects in a cell form a group. When the data set is very large so that most of the data are stored on disk, the CELL method may incur many random accesses to disk, which is costly. An alternative method was proposed, which uses a very small amount of main memory (around 1% of the data set) to mine all outliers by scanning the data set three times.

STING, Wave cluster and CLIQUE are other methods for this approach.

The diagram shows all approaches with their methods

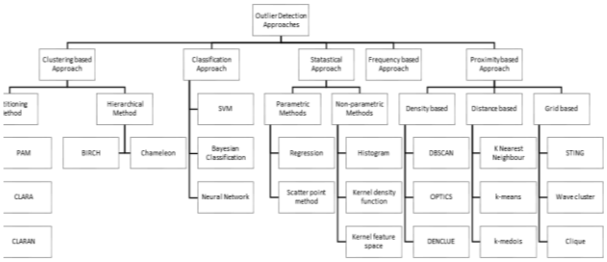

Diagram 1 Outlier analysis Approaches and methods

V. COMPARISON AND APPLICATIONS

Outlier detection is used in applications like fraud detection, Medical health outlier detection, Stock market Analysis, Terrorism Attacks, Credit card fraud detection, Network intrusion system, discovering computer intrusion and criminal behaviors, Medical condition monitoring - such as heart-rate monitors, detecting unauthorized access in computer networks, detecting mobile phone fraud by monitoring phone usage, monitoring the performance of computer networks to detect mislabeled data in a training data set, These applications adopt one or combination of more than one outlier detection approaches. Method used for outlier detection is application specific.

This adopted approach/approaches is based on behavior of attributes used in application. This table shows comparison of these outlier approaches, their methods and their advantage and disadvantages.

Here is list of applications of outlier detection.

- Credit Card Fraud detection [38]
- Medical and Public health outlier detection
- Network Intrusion System
- Wireless Network Sensor [9]
- Terrorism Attack Detection
- Stock market Analysis [37]
- Financial Fraud Detection [39]
- Industrial Fraud detection [40]

Outlier Detection Approaches	**Methods**	**Advantages**	**Disadvantages**	**Comments / Applications**
Clustering based Approach – Partitioning Method – Hierarchical Method – Proximity-based Approach (explained in detail below)	Partitioning Methods – PAM – CLARA – CLARAN Hierarchical Methods – BIRCH – Chameleon	Partitioning method – Less complex – Very effective on high dimensional data Hierarchical Method – No need to assume or define number of cluster initially.	Partitioning Methods – Difficulties in finding clusters of arbitrary shape such as 'S' and oval.	Generally hierarchical methods are not used in outlier detection. Partitioned method – Designed to find spherical-shaped data. – Used in intrusion detection.
Classification Approach	– SVM – Neural Network	SVM – Can classify both linear and nonlinear data Neural Network – Can classify patterns even if not trained data – Efficient to handle noisy data	SVM – Lack of transparency of results. Neural Network – Poor training time – Poor interoperability	SVM – Medical diagnosis Neural Network – Insurance claim fraud detection
Statistical Approach – Parametric Methods – Non-parametric Methods	Parametric Methods – Regression – Scatter point method Non-parametric Methods – Histogram – Kernel density function – Kernel feature space	▪ Most outlier research has been done in this area, many data distributions are known.	▪ Performs poorly on High-dimensional data. ▪ Less efficient ▪ Very complex	▪ This method can use only on numerical data.
Proximity based Approach – Density based – Distance based – Grid based	Density based – DBSCAN – OPTICS – DENCLUE Distance Based – k-Nearest Neighbor – k-means – k-medoid Grid Based – STING – Wave cluster – Clique	Density based – High level of interoperability. – Can find clusters of arbitrary shape Distance Based – Easy to understand and implementation – Do not rely on any assumed distribution to fit the data. Grid Based: – Fastest processing time that typically depends on size of grid instead of data.	Density based – More complex mechanism Distance Based – not effective in high-dimensional space due to the curse of dimensionality Grid Based: – For the problems with highly irregular data distribution, the resolution of grid mesh must be too fine to obtain a good clustering quality.	▪ Wireless Sensor Network ▪ Breast Cancer Analysis

Table 1 Comparison of Outlier detection Approaches

VI. CONCLUSION

The main objective of this paper is to review various outlier detection methods and to study how the techniques are categorized. Selection of method to detect outlier depends on the type of data involved. Method used for outlier detection is application specific. There is no universally accepted gamut of any method or approach to detect and analyze the outliers.

REFERENCES

[1] R. Bansal, “Outlier Detection: Applications and Techniques in Data Mining,” 2016.

[2] Sadawarti, Kamal Malik H, Member IEEE, and G S Kalra. “Comparative Analysis of Outlier Detection Techniques.” *International Journal of Computer Applications* 97.8 (2014): 12–21.

[3] S. S. Rakhe and A. S. Vaidya, “A Survey on Different Unsupervised Techniques to Detect Outliers,” Int. J. Eng. Technol., pp. 514–519, 2015.

[4] A. M. Said, D. D. Dominic, and B. B. Samir, “Frequent pattern-based outlier detection measurements: A survey,” 2011 Int. Conf. Res. Innov. Inf. Syst. ICRIIS’11, 2011.

[5] D. R. Chandarana, “A Survey for Different Approaches of Outlier Detection in Data Mining,” Int. Conf. Electr. Electron. Signals, Commun. Optim. - 2015.

[6] Breunig, M.M., Kriegel, H.P., and Ng, R.T., “LOF:Identifying density- based local outliers.” ACM conference Proceedings, 2000, pp. 93-104.

[7] K. M. H. Sadawarti, M. Ieee, and G. S. Kalra, “Comparative Analysis of Outlier Detection Techniques,” Int. J. Comput. Appl., vol. 97, no. 8, pp. 12–21, 2014.

[8] B. Kale, “Detection of outliers,” Sankhyā Indian J. Stat. Ser. B, vol. 38, no. 4, pp. 57–70, 1976.

[9] Yang Zhang, Y. Z. Y. Zhang, N. Meratnia, P. Havinga, “Outlier Detection Techniques for Wireless Sensor Networks: A Survey,” IEEE Commun. Surv. , vol. 12, no. 2, pp. 159–170, 2010.

[10] R. T. Ng, E. M. Knorr, R. T. Ng, and V. Tucakov, “Distance-Based Outliers: Algorithms and Distance-based outliers: algorithms and applications,” no. February, pp. 1–17, 2000.

[11] K. Manoj and K. S. Kannan, “Comparison of methods for detecting outliers,” Int. J. Sci. Eng. …, vol. 4, no. 9, pp. 709–714, 2013.

[12] R. L. Devi, “Hubness in Unsupervised Outlier Detection Techniques for High Dimensional Data – A Survey,” vol. 4, no. 11, pp. 797–801, 2015.

[13] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” ACM Comput. Surv., vol. 41, no. September, pp. 1–58, 2009.

[14] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, “LOF: Identifying Density-Based Local Outliers,” Proc. 2000 Acm Sigmod Int. Conf. Manag. Data, pp. 1–12, 2000.

[15] S. Vijayarani, “Sensitive Outlier Protection in Privacy Preserving Data Mining,” Int. J., vol. 33, no. 3, pp. 19–27, 2011.

[16] Z. A. Bakar, R. Mohemad, A. Ahmad, and M. M. Deris, “A Comparative Study for Outlier Detection Techniques in Data Mining,” IEEE Conf. Cybern. Intell. Syst., pp. 1–6, 2006.

[17] V. J. Hodge and J. Austin, “A Survey of Outlier Detection Methodologies,” Artif. Intell. Rev., vol. 22, no. 1969, pp. 85–126, 2004.

[18] Pasha, M.Z. and Umesh, N., “A Comparative Study on Outlier Detection Techniques,” Int. J. Comput. Appl., vol. Vol.66, No, no. 24, pp. 23–27, 2013.

[19] P. Gogoi, D. K. Bhattacharyya, B. Borah, and J. K. Kalita, “A survey of outlier detection methods in network anomaly identification3,” Comput. J., vol. 54, no. 4, pp. 570–588, 2011.

[20] J. Singh, “Survey on Outlier Detection in Data Mining,” Comput. Appl., vol. 67, no. 19, pp. 29–32, 2013.

[21] X. Jian-Qiong, “Local Outlier Detection Method towards Data Stream,” Commun. Softw. Networks (ICCSN), 2011 IEEE 3rd Int. Conf., pp. 479–482, 2011.

[22] M. M. K. Deshmukh and P. A. S. Kapse, “A Survey On Outlier Detection Technique In Streaming Data Using Data Clustering Approach,” Int. J. Eng. Comput. Sci., vol. 5, no. 1, pp. 1–4, 2016.

[23] X. Su and C. Tsai, “Outlier detection,” vol. 1, no. June, pp. 261–268, 2011.

[24] K. Subramanian and E. Ramraj, “Outlier detection: a review,” Int. J. Adv. Embed. Syst. Res., pp. 55–71, 2011.

[25] D. S. Shukla, A. C. Pandey, and A. Kulhari, “Outlier detection: A survey on techniques of WSNs involving event and error based outliers,” Proc. Int. Conf. Innov. Appl. Comput. Intell. Power, Energy Control. With Their Impact Humanit. CIPECH 2014, no. November, pp. 113–116, 2014.

[26] M. Gupta, J. Gao, C. Aggarwal, J. Han, and J. Gupta, Manish; Gao, Jing; Aggarwal, Charu C; Han, “Outlier Detection for Temporal Data : A Survey,” Tkde, vol. 25, no. 1, pp. 1–20, 2013.

[27] D. Goyal and H. Singh, “Survey Paper on Data Mining Techniques: Outlier Detection and Text summarization,” vol. 5, no. 3, pp. 223–227, 2014.

[28] S. Seo and P. D. Gary M. Marsh, “A review and comparison of methods for detecting outliersin univariate data sets,” Dep. Biostat. Grad. Sch. Public Heal., pp. 1–53, 2006.

[29] G. Williams, R. Baxter, H. He, S. Hawkins, and L. Gu, “A comparative study of RNN for outlier detection in data mining,” Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE Int. Conf., pp. 709–712, 2002.

[30] P. Rana, D. Pahuja, and R. Gautam, “A Critical Review on Outlier Detection Techniques,” Int. J. Sci. Res., vol. 3, no. 12, pp. 2394–2403, 2014.

[31] P. Chauhan and M. Shukla, “A Review on Outlier Detection Techniques on Data Stream by Using Different Approaches of K-Means Algorithm,” pp. 580–585, 2015.

[32] R. Kaur and S. Singh, “A survey of data mining and social network analysis based anomaly detection techniques,” Egypt. Informatics J., vol. 17, no. 2, pp. 199–216, 2015.

[33] Alle, “Procedures for detecting outlying observations in samples,” Technometrics, vol. 11, no. 1, pp. 1–21, 1969.

[34] S. Agrawal and J. Agrawal, “Survey on anomaly detection using data mining techniques,” Procedia Comput. Sci., vol. 60, no. 1, pp. 708–713, 2015.

[35] R. T. Ng, E. M. Knorr, R. T. Ng, and V. Tucakov, “Distance-Based Outliers : Algorithms and Distance-based outliers : algorithms and applications,” no. February, pp. 1–17, 2000.

[36] J. Zhang, “Advancements of Outlier Detection: A Survey,” ICST Trans. Scalable Inf. Syst., vol. 13, no. 1, pp. 1–26, 2013.

[37] L. Zhao and L. Wang, “Price Trend Prediction of Stock Market Using Outlier Data Mining Algorithm,” Proc. - 2015 IEEE 5th Int. Conf. Big Data Cloud Comput. BDCloud 2015, pp. 93–98, 2015.

[38] A. D. Pawar, P. N. Kalavadekar, and S. N. Tambe, “A Survey on Outlier Detection Techniques for Credit Card Fraud Detection,” IOSR J. Comput. Eng., vol. 16, no. 2, pp. 44–48, 2014.

[39] E. W. T. Ngai, Y. Hu, Y. H. Wong, Y. Chen, and X. Sun, “The application of data mining techniques in financial fraud detection : A classification framework and an academic review of literature,” Decision Support System, vol. 50, no. 3, pp. 559–569, 2011.

[40] S. Cateni, V. Colla, and M. Vannucci, “Outlier Detection Methods for Industrial Applications,” no. October, 2008.

[41] Han and Kamber(2007), Data Mining: Concepts and Techniques Morgan Kaufmann publications.

[42] D. M. Hawkins, London, 1980. “Identification of Outliers”. Chapman and Hall.