Survey on Outlier Detection in Data Stream

Pooja Thakkar
Research Scholar
Dept. of Information Technology
G.H Patel College of Engg. & Tech., Gujarat, India

Jay Vala
Assistant Professor
Dept. of Information Technology
G.H Patel College of Engg. & Tech., Gujarat, India

Vishal Prajapati
Assistant Professor
Dept. of Information Technology
G.H Patel College of Engg. & Tech., Gujarat, India

ABSTRACT
Data mining provides a way for finding hidden and useful knowledge from the large amount of data .usually we find any information by finding normal trends or distribution of data .But sometimes rare event or data object may provide information which is very interesting to us .Outlier detection is one of the task of data mining .It finds abnormal data point or sequence hidden in the dataset .Data stream is unbounded sequence of data with explicit or implicit temporal context .Data stream is uncertain and dynamic in nature. Traditional outlier detection techniques for static data which require whole dataset for modelling is not suitable for data stream because whole data stream cannot be stored. Network intrusion detection ,web click stream analysis ,fraud detection ,fault detection in machines ,sensor data analysis are some of the applications of data stream outlier detection .In this paper, we have described several issues in data stream outlier detection and usual approaches or techniques for finding outlier in data stream .

Keywords
Data mining, Outliers, data stream mining.

1. INTRODUCTION
Data mining is one of the steps of Knowledge Discovery from data (KDD) process.[1].Data mining is the process of analyzing the data and finding patterns that are useful in decision making .Most data mining algorithms assume that the data fits into memory .Association pattern mining ,Clustering ,classification and outlier detection are building blocks in data mining .

1.1.Outlier detection
Definition: "An outlier is a data object that deviates significantly from the rest of the objects, as if it were generated by a different mechanism "[1]

In many applications, data are generated by processes which may reflect either the activity of the system or observation of objects in the system.[2] Outliers may appear in the data due to some reasons like Mechanical fault, Changes in System behavior, Fraudulent Behavior, Human Error, Instrument error, Natural deviations in populations, changes of environment etc.

Outlier can be noise or interesting item. There is no clear distinction between outlier and noise. It depends on interest of the analyst of the system. [2] In both the cases outlier detection is crucial because most of the statistical methods cannot work well in the presence of outliers. [6]

Applications of outlier detection are credit Card fraud detection, Telecom fraud detection, Loan application processing to detect fraudulent applications or potentially problematical customers, Intrusion detection detecting unauthorized access in computer networks. Detecting unexpected entries in databases for data mining to detect errors, frauds or valid but unexpected entries.

Types of Outliers*:*

(1). Global outliers: Global outliers are also called point outliers. A data point is called global outlier if it is different or far from the whole dataset. It is very simple form of outlier. Most of the techniques are designed for finding this type of outliers. For example, Intrusion detection, if a communication behavior of a computer is very different from the normal behavior trends then it is found to be a global outlier. [1] A data stream cannot have a global outlier because it has temporal context available with it. Data stream is infinite and online detection has to be done so only a subset of the data set is available for a specific time instead of the whole data. [9]

(2) Contextual outliers: It is also called conditional Outlier. A data point is called contextual outlier if it is different or far from the other data points in the specific context. Contextual outlier detection techniques provides flexibility for detecting outliers in different contexts. Whether a 30^0 C temperature is outlier or not depends on time and location. If it is winter in Toronto then it is an outlier. If it is summer then it is normal. [1]

In contextual outlier detection data attributes are divided into two groups

Contextual Attributes: Attributes with respect to which a data point is considered an outlier are contextual attributes. It defines context of the object .For example time, spatial attribute (Longitude and latitude), network location etc.[5]

Behavioral attributes: They are non-contextual attributes and evaluated to find outlierness of data point in context in which it belongs. [1] For example Temperature, humidity, pressure, rain fall etc. [5]

(3) Collective outliers: A collection of data point as a whole is different from the entire data set is called collective outlier. An individual data point in a collection may not be outlier. Usually data points are related in collection. Finding subsequence as anomaly in time series data set, finding sub regions as anomaly in spatial imaginary data set or finding sub graph as anomaly in graph data set are examples of collective outliers [5]

1.2. Data Stream
Today many data sources like sensor network , world wide web, Social networking sites , Telecommunication , Internet traffic, online transactions, medical systems ,Real time surveillance systems and other dynamic processes generate tremendous amount of data every time .They are coming continuously . We cannot store those data in limited memory of our computer for processing or analyzing. They are called

stream data. Data streams are uncertain, Dynamic and infinite sequence of data points. [9]

Data streams can be of two types

Time series data stream: In time series, temporal component is stronger than multidimensional data stream. [3]Time series data stream is treated as contextual dataset where time components can be a contextual attribute. Choice of similarity function is very crucial in time series data analysis. Dimensions in a time series data stream can be defined in two ways depending on the application: In multivariate time series, all behavioral attributes are considered as dimensions and in univariate time series, all values are considered as dimensions.

Multidimensional data stream: Multi-dimensional data stream analysis is same as the multi-dimensional static data set analysis but temporal component is added in data stream.[8] Temporal component in multidimensional data stream is weaker than time series data stream .All attributes of multidimensional data stream are treated equally. For outlier detection, time series data require the analysis of each series as a unit, whereas the multidimensional data requires the analysis of each multidimensional point as a unit. [2]

1.3.Outlier Detection in Data Stream

In many cases, the detection of unusual events needs to be performed in a time critical manner so techniques of detecting outliers in data stream must be developed .Most of the algorithms developed are for static dataset which can be stored in memory. Data stream cannot be stored in memory and algorithms have to be applied as it arrives.

Currently, most of the existing outlier detection algorithms only put focus on real-time outlier detection in data stream, but ignore subsequent changes of data stream, which means that these algorithms cannot find the mutual conversion between outliers and normal data points. [16]

Output of outlier detection algorithms can be

Outlier score: outlier score is assigned to data point according to its degree of outlierness. [5] Data point which has higher value of the outlier score has higher probability to be an outlier. [2] Output of such outlier detection techniques is ranked list of outliers. An analyst may choose to either analyze top few outliers or use a cut-off threshold to select the outliers.

Binary Label: These type of techniques assign a label indicating whether or not a data point is an outlier .This type of output contains less information than the first one because a threshold may be applied on the outlier score to convert them into binary labels. [2]

2. TYPES OF OUTLIER DETECTION TECHNIQUES

2.1 Statistical outlier detection

Statistical outlier detection techniques make assumption about normal data and outlier data. [1] It assumes some data distribution in the data set .Outliers are points that have a low probability to be generated by the overall distribution. [6] A good domain specific knowledge is required.

Statistical outlier detection methods are divided into two categories:

Parametric methods assume the distribution model priory. [1] Statistical outlier Detection methods use training data set to build the statistical model .In data stream we cannot assume data distribution because we cannot have entire data set and data distribution may change over time. So same training dataset or model cannot produce true outliers.

In Non–Parametric methods, the model of normal data is learned from the input data. Non parametric statistical methods does not make any assumption about the distribution of the data so it can be used in data stream with single dimension or very low dimensions [7]. They cannot be applied in high dimensional data stream

2.2 Distance Based Outlier Detection

Distance based outlier detection technique decides the outliness of the data point based on its distances to its nearest neighbors.

Definition [Knorr and Ng]: "**Given parameters *k* and *R*, an object is a distance-based outlier if less than *k* objects in the input data set lie within distance *R* from it**."[11]

They are defined for any data type for which distance measure or similarity measure is available and these methods do not require detailed understanding of application domain.[10] In distance based methods k-nearest neighbor distance from the original data points are considered for calculating outlier score instead of pre aggregated data so outlier detection is performed at a finer granularity than other methods like Clustering or density based methods .So they can distinguish between noise and true outlier.[3] Distance based methods do not assume any data distribution. So they can be used in data stream.

Definition for data stream: "**The outlier score of a data point is defined in terms of its k-nearest neighbor distance to data points in a time window of length W**." [3]

It is not effective for high dimensional data stream. High dimensional dataset in real application contains very much noise. Distance between all data points are equal in high dimensional data point .So degree of outlierness are same for all data point. [7]

2.3 Density Based Outlier Detection

In density based outlier detection, density around a data point is compared with the density around its local neighbors. The relative density of a point compared to its neighbors is computed as an outlier score.

Basic assumption in density based outlier detection method: The density around a normal data point is similar to the density around its neighbors. The density around an outlier is considerably different to the density around its neighbors.

In this type of method, outliers are detected by computing Local Outlier Factor (LOF), which is the ratio of local density of the point and the local density of its nearest neighbor. Data point whose LOF value is high is declared as outlier.

In [13], an incremental LOF algorithm which is suitable density based algorithm for data stream, is proposed. It can detect changes in the data behavior. It provides performance equivalent to static LOF algorithm. It cannot distinguish between outliers and new data behaviors.

In [14], an improvement of Incremental LOF algorithm is proposed .It can significantly distinct outliers from new data behavior.

The density based outlier detection methods are more effective than distance based methods [7]. But they are more complicated and computationally expensive because they involve density of both the point and its neighbors also. Density based methods are not effective for high dimensional

data set because the accuracy of the density estimation process degrades with increasing dimensionality.

2.4 Sliding window based outlier detection

Streaming data uses sliding window concept which is used for maintaining the statistical information in data stream. The window is identified by two sliding end points. [11] Both ends are active. During the moving process, both ends are moving in the same direction and shifting the same units. Let W be the window size so only the last W records to arrive in data stream are relevant at any point of time. It has some overlaps between the next window and the last window. Here W is fixed based on the number of records or the interval of time. [15]

If some data points are outliers in one window, they can be inliers in other window because nature of data stream is dynamic and data behavior may change during the time. Hence detecting any changes in a data as outlier is not desirable. So determining outlierness of a data point as it arrives although meaningful can lead us to wrong decisions.
Choosing accurate window size in sliding window based outlier detection is required. Choice of sliding window is independent of data point used for implementation which gives poor result over outlier detection. [6]

2.5 Clustering based outlier detection

Clustering based outlier detection is an unsupervised outlier detection technique in which class label as "normal" or "outliers" are not available. For this reason, clustering means learning by observation rather than learning by examples. [1]Clustering method is used to group similar data points in a cluster. The main requirements for clustering evolving data streams are Summarization, Processing, and Outlier detection. [17] Here we assume that Normal data instances belong to a cluster in the data, while anomalies either do not belong to any cluster, Normal data instances are close to the closest cluster centroid, while anomalies are far from their closest cluster centroid, Normal data instances belong to large and dense clusters, while anomalies either belong to small or sparse cluster. [19]
In [18] Clustering based outlier detection technique for evolving data stream is proposed that assigns weights to attributes according to its relevance in mining task.
The testing phase for clustering based techniques is fast because the number of clusters against which every test instance needs to be compared is small. [19]
The main aim of most of the clustering algorithms is to find clusters rather than outliers and they are not optimized to find anomalies. Most of the existing clustering algorithm requires number of clusters in advance and shape of the clusters are also defined, but in data stream we cannot assume no of clusters in advance. Arbitrary shape clusters also cause some difficulties in realizing exact clusters of the data.

3. ISSUES OF OUTLIER DETRECTION IN DATA STREAM

Issues related outlier detection in data stream with respect to data stream characteristics:

3.1 Transient

Data point in data stream are transient in nature .So after some time it loses its importance because it is discarded or archived. Earlier outlier detection techniques construct the outlier detection model using entire dataset and then data point is compared to the model or other data points to detect whether it is outlier or not .For data stream, outlier detection technique should detect outlierness of object immediately as it arrives. **[9]**

3.2. Notion of time stamp

Each data point in data stream is associated with some notion of time implicit or explicit. In explicit association, time is an attribute and in implicit association, exact time is not important but order of the data points is important .If we consider temporal context, a data point is considered outlier if it deviates significantly from the other data points with the same temporal context .an appropriate temporal context has to be decided first and then every data point has to be processed according to its temporal context.

3.3 Infinite

Data stream is an infinite sequence of data points as they keep coming from a data source indefinitely. So at any specific time the entire dataset cannot be available so many static outlier detection techniques which require whole dataset for detecting outlier cannot be used. Outlier detection method for data stream should store summary of the data set and summary should be computed incrementally. A data point is compared with the summary of the data points instead of other data points. Thus an outlier detection model has to be incremental and cannot assume the availability of the entire dataset. [9]

3.4. Arrival rate

Arrival rate may be fixed or variable. Outlier detection technique for data stream has to process data point before next data point arrives .The set of data points or the summary of the data points, to which the current data point is compared to detect outlier-ness, should be adjusted based on the available processing time.[9]

3.5. Concept Drift

The distribution of data may change over time in data stream and it is called concept drift. Data point which has detected as outlier for one data distribution may change its outlierness with changing data distributions. So outlier detection technique for data stream cannot assume any fixed distribution of data. [12]

3.6. Uncertainty

Today new hardware technology such as sensors are generating large amount of data. But data contain missing, inconsistent or erroneous values .Uncertainty indicates it is impossible to determine whether the information available is true or not. Uncertainty in data stream is a key challenge for outlier detection. **[10]**

4. CONCLUSION

In this paper. Various types of outliers, types of data stream, various approaches for outlier detection and issues related to outlier detection in data stream have been discussed.. Outlier can be a noise or an interesting information in many applications. Data stream is infinite sequence of data which cannot be stored and dynamic in nature so any changes in a data behavior cannot be considered as an outlier. Thus detecting outlier as it arrives can lead us to wrong decisions. Here it can be said that for data stream unsupervised method is more suitable because unavailability of labelled streaming data. A number of open challenges still remain in data stream algorithms particularly in DataStream clustering.

5. REFERENCES

[1] Jiawei Han, Micheline Kamber and Jian Pei, "Data mining Concepts and Techniques", Third Edition, Morgan Kaufmann Series in Data management Systems.

[2] Charu C. Aggarwal, "Outlier Analysis", Springer, 2013.

[3] Charu C. Aggarwal, "Data Mining: The Textbook", Springer, 2015.

[4] QIANG YANG, "10 CHALLENGING PROBLEMS IN DATA MINING RESEARCH",International Journal of Information Technology & Decision Making Vol. 5, No. 4 ,2006.

[5] Karanjit Singh and Dr. Shuchita Upadhyaya**, "**Outlier Detection: Applications And Techniques**",** IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 3, January 2012.

[6] Sreevidya S S, "A Survey on Outlier Detection Methods", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (6), 2014.

[7] Ji Zhang, "Advancements of Outlier Detection :A Survey", ICST Transactions on Scalable Information Systems ,January-March 2013 , Volume 13 issue 01-03 ,e2.

[8] Manish Gupta, Jing Gao, Member, IEEE, Charu C. Aggarwal, Fellow, IEEE, and Jiawei Han, Fellow, IEEE , "Outlier Detection for Temporal Data: A Survey", IEEE

[9] Shiblee Sadik and Le Gruenwald, "Research Issues in Outlier Detection for Data Streams", SIGKDD Explorations, Volume 15, Issue 1.

[10] Neeraj Chugh, Mitali Chugh, Alok Agarwal, "Outlier Detection in Streaming Data A research Perspective", International Conference on Parallel, Distributed and Grid Computing, IEEE, 2014.

[11] Fabrizio Angiulli, Fabio Fassetti, "Detecting Distance-Based Outliers in Streams of Data", ACM, 2007.

[12] Md. Shiblee Sadik, Le Gruenwald, "DBOD-DS: Distance Based Outlier Detection for Data Streams", dexa, 2010.

[13] Dragoljub Pokrajac, Aleksandar Lazarevic, Longin Jan Latecki, " Incremental Local Outlier Detection for Data Streams", IEEE Symposium on Computational Intelligence and Data Mining (CIDM), April 2007.

[14] Seyed Hesamodin Karimian, Manouchehr Kelarestaghi, Sattar Hashemi, "I-IncLOF: Improved Incremental Local Outlier Detection for Data Streams", The 16th CSI International Symposium on Artificial Intelligence and Signal Processing, IEEE, 2012.

[15] Xiaoke SU, Yang LAN, "Sliding Window-based Outlier Detection in Mixed Data Stream", Journal of Computational Information Systems 6:14, 2010.

[16] Yu Xiang, Lei Guohua1, Xu Xiandong Lin Liandong, "A Data Stream Outlier Detection Algorithm Based on Grid", IEEE, 2015.

[17] Amineh Amini and Teh Ying Wah, "Requirements for Clustering Evolving Data Stream", 2nd International Conference on Soft Computing and its Applications (ICSCA'2013) Sept. 25-26, 2013.

[18] Yogita Thakran, Durga Toshniwal, "Unsupervised Outlier Detection in Streaming Data Using Weighted Clustering", IEEE, 2012.

[19] Varun Chandola, Arindam Banerjee, Vipin Kumar, Aomaly Detection: A Survey", ACM Computing Surveys, 2009.