International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor :6.887*
*Volume 5 Issue XII December 2017- Available at www.ijraset.com*

# A Review on Outlier Detection Techniques

Ms. D. R.Gupta[1], Prof. Dr. S. M. Kamalapur[2]
[1]*M.E. Student, Department of Computer Engineering K. K. Wagh Institute of Engineering Education & Research, Nashik Savitribai Phule Pune University, Maharashtra, India.*
[2]*Associate Professor, Department of Computer Engineering K. K. Wagh Institute of Engineering Education & Research, Nashik Savitribai Phule Pune University, Maharashtra, India.*

***Abstract: The outlier is unexpected behavior of data. Outlier detection is important in various domains like fraud detection, intrusion detection, activity monitoring, etc. Data is generated continuously on large scale in many applications. There is need to detect outlier from static data as well as streaming data. There are basic two types of outlier: global outlier and local outlier. This work aims to study various local and global outlier detection techniques for static and streaming data. The works also focuses on various local and global outlier detection techniques which are efficient in terms of time and memory.***
***Keywords: Global outlier, local outlier, outlier Detection, streaming data***

## I. INTRODUCTION

Outlier is nothing but the unexpected behavior of data. Outlier represents extreme or irregular entries in dataset. Outlier detection is applicable in variety of domains such as fraud detection occurred in insurance sector, bank transaction, credit card, intrusion detection in network analysis and cyber security, activity monitoring such as human gait analysis, enemy activity monitoring in military surveillance, etc. Outlier detection and noise removal are closely related topics as they deal with unexpected data in dataset but these two are different techniques. Noise is unwanted data and creates obstruction in the data analysis activity. Analyst removes the noise before data analysis is performed. In many applications data is generated continuously in large scale. To detect outlier on such continuous/ streaming data is a challenging task. Such high rate streaming data generates bulk amount of data. For outlier detection process initially data need to be loaded in main memory. Streaming data is unbounded large amount of data and it is impractical to load whole data in the memory. With limited memory access devices such as wireless sensor networks, raspberry pi devices, etc. real time applications are designed. There is a need to detect rare event among the captured data. Hence memory efficient solution is required for such devices. Paper is organized as follows: section I introduces the outlier detection techniques and methods. Section II gives the literature review. Section III concludes the paper.
In the below sections we are going to discuss about related work done for the proposed research area. We refer some existing research paper for completing this task. It is given as follow:

## II. RELATED WORK

This section describes the work done in the outlier detection techniques and methods. Details of these techniques and methods are as follows:

### *A. Distribution Based Approach*

This is the statistical approach which derives statistical inference form the given input dataset. This technique extracts the unseen instances belonging to the dataset. This technique includes two approaches.

1) *Parametric technique:* In this technique underlying knowledge of data distribution is required which is not available in each case.
2) *Non Parametric Technique:* This technique does not assume the underlying data distribution. Yamanishi and Takeuchi [3] proposed non-parametric technique that applied on non-stationary time series data. It dynamically analyses the data and defines the change point incrementally based on probability density function.

This technique can be able to find global outlier form data.

### *B. Clustering Based Approach*

Clustering based outlier detection [5] [4] technique is applied first and then points those are far away from cluster centroid are declared as an outlier. The main aim of such system is to build clusters. This technique finds the global outlier.

### *C. Distance Based Outlier Detection*

In this technique distance between data points is calculated. Outliers are those points whose distance is higher than other points in the dataset. The distance based outliers are broadly classified in two categories:

1) *Global outlier:* In Global outlier technique whole dataset T is considered and outlier point O is detected if at least P points have higher distance from O in whole dataset T.

Knox and Ng [1] proposed partitioning based outlier detection technique. This algorithm is suitable for low dimensional dataset and generates best results for dimensionality count is less than equal to four.

Neighbor-based pattern detects [9] global outlier on streaming data. It uses sliding window technique to detect incrementally neighbor based patterns. This technique proposes Abstract-M algorithm based on view prediction technique.

Outlier is detected over high speed streaming data with limited memory resources [10]. To work with limited memory resources, global outlier uses classification technique to reduce the dataset size using data editing approach. But this is supervised approach and it requires training data in editing phase with normal entries.

Lifespan-Aware Probing Operation -LEAP technique [11] is proposed for high dimensional streaming data. It optimizes the search space. Lifespan-Aware Probing Operation -LEAP technique is faster than the Neighbor-based pattern detection technique [9] and data editing technique [10].

All these techniques are distance based techniques and are not be able to detect non-homogeneous densities in data and outlier in non-homogeneous densities.

To overcome this problem, combined approach is proposed [8][12]. These techniques combines the strategies used in cluster based and distance based approaches. These techniques work on streaming data. To make the algorithm memory efficient, these techniques preserves the summary of each data stream in terms of candidate outlier and cluster information in memory rather than the complete stream data points.

2) *Local Outlier:* In local outlier detection technique local density of each data point is calculated. The data points having lower local density as compared to the other data points are considered as local outliers. Local density of a point is the reachable distance from other neighboring points.

Following are the different approaches to find local out liers Local outlier factor (LOF) is assigned to each data point in a dataset [2]. This technique combine the strategy of distance based and cluster based outlier detection technique. LOF is calculated using k-nearest neighbor distance and Reach ability distance of point from nearest k points.

This approach is applicable for static data environment. This is unsupervised technique and identifies outliers with higher accuracy.

LOF is applicable for numeric data set. LOF is applied on categorical dataset as well as on numerical dataset. LOF proposes K-LOF algorithm to find outlier at the center rather than at the cluster borders.

Incremental Local outlier factor (ILOF) is the incremental approach of LOF algorithm [7]. This technique is proposed for outlier detection in data streams. In existing LOF techniques complete dataset is required for processing. To overcome the drawback of existing techniques and to reduce the computational overhead, Incremental LOF - iLOF technique evaluates k nearest neighbor and evaluates its local outlier factor based on k-nearest neighbors and updates the k-nn value and local outlier factor of those k points.

There is no need to update LOF of each point and hence this technique is efficient. But for this processing whole data need to be loaded in the memory. As streaming data is unbounded, it is practically impossible to load complete data in memory. Hence this approach is applicable only for limited size dataset.

Memory efficient incremental local outlier (M ILOF) is unsupervised, memory efficient technique [13]. To identify outlier on limited memory resources MiLOF technique is proposed. The processing of data is distributed in three phases. : Summarization, Merging and revised insertion. To fulfill to the memory constraint requirement, this algorithm create the summary of existing data points. Rather than keeping the complete data in memory summary of data point is loaded. LOF of each upcoming data point is calculated with respect to other streaming data points and the previously generated summary. In merging phase, the system generates cluster of incoming data points using C-means algorithm these clusters are then incorporated to the existing cluster or new clusters are created. In revised insertion LOF values of data points are updated.

MiLOF technique reduces the computational time as well as memory and is suitable for low memory devices such as wireless sensor network.

This technique handles the high dimensional data for processing. The performance of the system varies with respect to the dataset size, data dimension and number of cluster generated during processing.

The above section summarizes the related work of outlier detection technique finds the data points in dataset having behaviour different than other data points in the dataset. Local outlier can be identified using various techniques such as distribution based, clustering based, distance-based. The focus of the research is on local outlier factor. Local outlier factor (LOF) is based on density

approach. To overcome this problem, memory efficient incremental outlier detection over data stream-(MiLOF) is proposed in the literature. This technique preserves the summary of previous data points in each iteration and finds local outliers.

## III. CONCLUSION

Various outlier detection techniques are discussed in this paper. These techniques are mainly classified in three categories: distribution based, cluster based and distance based. Local outlier can be detected using distance based technique. To achieve more accurate results distance based and cluster based technique are merged together. To provide memory efficient solution for finding local outlier over streaming data MiLOF is the solution. The performance of MiLOF varies with respect to the input data. There is need to evaluate the system performance under various circumstances such as: change is dimension count, cluster creation count and dataset size.

## IV. ACKNOWLEDGMENT

Authors would like to thanks Prof. Dr. K. N. Nandurkar, Principal and Prof. Dr. S. S. Sane, Head of Department of Computer Engineering, K.K.W.I.E.E.R., Nashik for their kind support and suggestions. We would also like to extend our sincere thanks to all the faculty members of the department of computer engineering and colleagues for their help.

## REFERENCES

[1] E. M. Knox and R. T. Ng, 1998, Algorithms for mining distance-based outliers in large datasets.
[2] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, 2000 LOF: Identifying density-based local outliers.
[3] K. Yamanishi and J.-I. Takeuchi, 2002, A unifying framework for detecting outliers and change points from non-stationary time series data.
[4] S. Guha, A. Meyerson, N. Mishra, R. Motwani, and L. O'Callaghan, 2003, Clustering data streams: Theory and practice.
[5] F. Cao, M. Ester, W. Qian, and A. Zhou, 2006, Density-based clustering over an evolving data stream with noise.
[6] J. X. Yu, W. Qian, H. Lu, and A. Zhou, 2006, Finding centric local outliers in categorical/numerical spaces.
[7] D. Pokrajac, A. Lazarevic, and L. J. Latecki, 2007, Incremental local outlier detection for data streams.
[8] M. Elahi, K. Li, W. Nisar, X. Lv, and H. Wang, 2008, Efficient clustering based outlier detection algorithm for dynamic data stream.
[9] D. Yang, E. A. Rundensteiner, and M. O. Ward, 2009, Neighbor-based pattern detection for windows over streaming data
[10] V. Niennattrakul, E. Keogh, and C. A. Ratanamahatana, 2010, Data editing techniques to allow the application of distance-based outlier detection to streams
[11] L. Cao, D. Yang, Q. Wang, Y. Yu, J. Wang, and E. A. Rundensteiner, 2014, Scalable distance-based outlier detection over high-volume data streams.
[12] M. Moshtaghi, 2014, Streaming analysis in wireless sensor networks.
[13] Mahsa Salehi, Christopher Leckie, James C. Bezdek, Tharshan Vaithianathan and Xuyun Zhang, 2016, Fast Memory Efficient Local Outlier Detection in Data Streams.