www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 4 Issue 11 Nov 2015, Page No. 14946-14949

A Survey paper on An Effective Analytical Approaches for Detecting Outlier in Continuous Time Variant Data Stream.

Mr. Raghav M. Purankar [1], Prof. Pragati Patil [2]

[*1] M.E Scholar , Deptt of CSE , Abha Gaikwad Patil College of Engineering and Technology
R T M Nagpur University , Nagpur , Maharashtra, India .

E-mail : raghavpurankar88@gmail.com

[*2] Department of CSE , Abha Gaikwad Patil College of Engineering and Technology
R T M Nagpur University , Nagpur , Maharashtra, India .

E-mail : pragatimit@gmail.com

ABSTRACT

Outlier detection and is an important branch of data mining. Data mining is extensively studied field of research area; where most of the work is focused on the information discovery. A data stream is a massive sequence of data objects continuously generated at much faster rate. There are various approaches and methods are used for outlier detection. Some of them use K-Means algorithm for outlier detection in data streams which help to create a similar group or cluster of data points. The K-means algorithm is the best known partitioned clustering algorithm. As we know that streaming data often fails to scan the multiple items and also the new concepts may keep evolving in coming data over time hence the outlier detection plays the challenging role in the streaming data. The irrelevant attributes can be termed as noisy attribute at the time of working with the data streams objects and such attributes imposes the challenge. In high dimensional data the number of attributes associated with the dataset is very large and it makes the dataset unmanageable. Clustering is a data stream mining task which is very useful to gain insight of data and data characteristics. Clustering is also used as a pre-processing step in over all mining process for an example clustering is used for outlier detection and for building and development of Hybrid approach. Purpose of this paper is to review of Hybrid approach of outlier detection with others approach which uses K-Means algorithm for clustering dataset with some other techniques like Euclidean distance approach. Various application domains of outlier detection are discussed in this paper.

KEYWORDS : *Outlier Detection, Euclidean Distance, K-Means, Dataset, Information Discovery.*

1. INTRODUCTION

Outlier detection has been a very important concept in the realm of data analysis. Recently, several application domains have realized the direct mapping between outliers in data and real world anomalies, that are of great interest to an analyst. Outlier detection refers to the problem of finding patterns in data that do not conform to expected normal behavior. Data streams are potentially unlimited chain of data objects, they are temporally ordered. It is not possible to store whole data stream because of its large volume. As the time going on the new concepts or data objects are added in the data streams. This new concepts are requires algorithm for processing the data streams The algorithm require for continuously update the models of data streams for adapting the changes. Collecting the labeled data for the data mining is the very difficult task, and also adding the new concept may come to existent and others may get outdated in streaming data. The clustering based outlier detection methods which are
existing give the equal importance to the relevant ant irrelevant attributes. This behavior gives them the poor performance on real world data. A data stream is an unremitting, immediate, stream flow of sequence of items and it is not possible to control the order in which data item arrive, or not possible to store these entire data items. Data stream clustering is a well-known task in mining data stream, clustering is known as grouping related objects into a cluster. Applications of outlier detection are web logs, fraud detection. Clustering is the process of partitioning or grouping a given set of patterns into disjoint clusters. This is done such that patterns in the same cluster are alike and patterns belonging to two different clusters are different.

DOI: 10.18535/Ijecs/v4i1.11

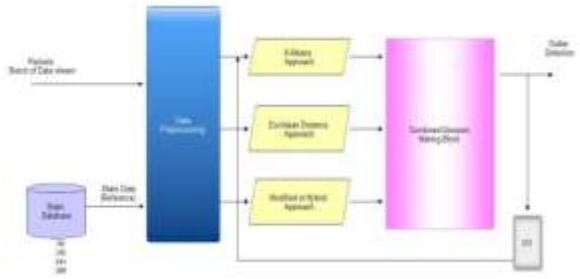


Fig 1 : Basic overview of the System

2. IMPORTANT CONCEPTS OUTLIER DETECTION ON DATA STREAM :

A. Outlier

The Outlier is also referred as anomalous objects due to different behavior with respect to other data elements. The Term outlier originates from Statistical domain [10]. Various definitions have been proposed for outlier in data mining. Here we are following two definitions as a classical definitions for outlier in data mining. First definition is "*An outlier is an observation, which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism*"[18]. Second 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*" [17].

B. Basic Motivation for Outlier Detection in Streaming data.

outlier detection can be viewed as fundamental task in data analysis i.e., extracting useful and interesting information from a large amount of data [8]. More ever due to rapid stream evolution, data element property can be considered as time varient, here we are not able to scan the data second time. In order to deal with problem of processing streaming data efficient outlier detection method need to be used. Also it required more attention from data mining community. The Constraints of dataset and the nature of data make design of an appropriate outlier detection technique more challenging. Traditional outlier detection techniques might not be suitable for handing dynamic nature of data. Some of the following aspects are shown in table which motivates us that we require efficient outlier detection method over streaming data.

C. Challenges and important Features of Streaming data

- **Streaming Data in Distributed Manner-**

Data coming from various distributed application may dynamically according to time constraints. Computational probability is challenging or difficult task because of such Dynamic nature.

- **Large Size with High Speed-**

Data streams are of infinite size and continuously flow at very high speed of data. Due to this processing and storage of data streams is computationally expensive. It is found that traditional outlier detection methods do not perform well to process large amount of distributed data streams in an online enviorment.

- **Uncertainty and Noisy Attribute in Streaming data-**

In most application domain that we don't have a sufficient operational data because of uncertainty and missing data attributes. Due to insufficient information for data that may lead us for wrong prediction.

- **Single Scan of Streaming data-**

Due to infinite volume of data stream it is impossible to store entire data stream like traditional data sets that can be stored in memory. This leads to the only single access of streaming data points. So clustering technique should be such that it can store the summaries of data for further analysis in single scan of streaming data.

- **Robustness to Outliers-**

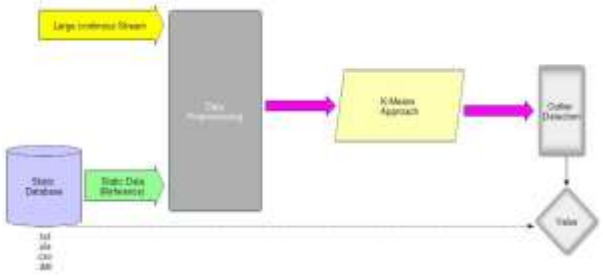
A data stream clustering must be able to identify the outlying data points because outliers can distort the complete clustering structure of data.

3. PROPOSED WORK

3.1. Proposed K-Means Clustering Algorithm

Generating cluster: *K*-means clustering is a partitioning method .Initially, cluster the entire dataset into k cluster using K-mean clustering and calculate centroid of each cluster. K-mean Clustering: Given *k*, the *k-means* algorithm is implemented in following four steps:

a) Select k observations from data matrix X at random
b) Calculate distance with each instances (with respect to randomly selected instances)
c) Assign each instance to the cluster with the nearest seed
d) Go back to Step b, stop when no instance to move group.



3.2. Distance based Algorithm

This method is highly dependent on parameter provided by the users and computationally expensive when applied unbounded data set. With the development of information technologies, the number of databases, as well as their dimensions and complexity grow rapidly. With high dimensional dataset calculate distance with each instances will increase the computational cost. We are comparing distance based method with proposed method.
Pairwise distance computes the Euclidean distance between pairs of objects in n-by-p data matrix X. Rows of X

correspond to observations; columns correspond to variables. y is a row vector of length n(n–1)/2, corresponding to pairs of observations in X. The distances are arranged in the order (2,1), (3,1), ..., (n,1), (3,2), ..., (n,2), ..., (n,n–1)). y is commonly used as a dissimilarity matrix in clustering or multidimensional scaling.

Euclidean distance :

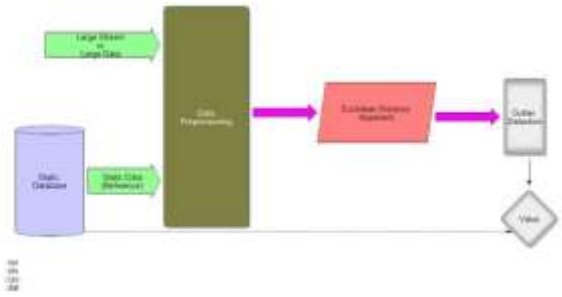
$$d_{rs}^2 = (x_r - x_s)(x_r - x_s)'$$

Where,

$$\bar{x}_r = \frac{1}{n}\sum_j x_{rj} \text{ and}$$

$$\bar{x}_s = \frac{1}{n}\sum_j x_{sj}$$

1) Calculate pairwise distance that is computing the Euclidean distance between pairs of object.
2) Take square distance. Calculate maximum values from square distance values
3) Take threshold from user.
4) If distance > threshold value that will be the outliers.

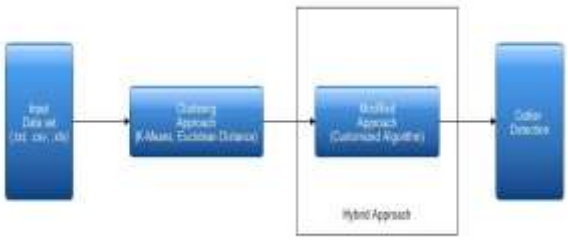


Input Data Set: Collecting dataset from UCI Machine learning repository [19].

Cluster Based Approach:
Clustering is an important concept for outlier analysis. Cluster based approach is here act as filter for noisy data. Statistical outlier detection techniques formulate the model using distribution of data point available for processing. Detection model is formulated to fit the data with reference to distribution of data.

Distance Based Approach:
Distance based technique is used to find out maximum distance value for each cluster. If this distance is greater than some threshold value then it will declare as outlier||else as an inliers. Threshold is given by user.

Hybrid Approach for Outlier Detection:
Outlier detection is a process of finding objects that are inconsistent or having different behavior with respect to the remaining data or which are distant apart from their cluster centroids.



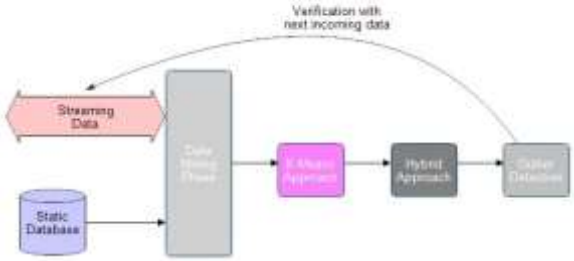
4. OUTLIER DETECTION OVER STREAM DATA

Outlier detection is a primary step in many data-mining applications. It refers to the problem of finding patterns in data that do not conform to expected normal behavior or anomalous behavior. These anomalous patterns are often referred to as outliers, anomalies, discordant observations, exceptions, faults, defects, aberrations, noise, errors, damage, surprise, novelty, peculiarities or contaminants in different application domains.
Outlier detection technique has following major ingredients.
1. Nature of data, nature of outliers, and other constraints and assumptions that collectively constitute the problem formulation.

2. Application domain in which the technique is applied. Some of the techniques are developed in a more generic fashion others directly target a particular application domain.

3. The concept and ideas can be applied from one or more knowledge domains.

Data are continuously coming in a streaming environment with a very fast rate and changing data distribution (change of data distribution is known as concept drift) [9], and thus, any fixed data distribution is not adequate to capture the dynamic behavior of data streams.



5. DISCUSSION

Effective outlier detection requires the developing of a model that accurately represents the data. Over the years, a large number of techniques have been developed for building such models for outlier and anomaly detection. To present effectiveness for outlier detection that require a handle following drawback of existing outlier detection techniques. We provide a review of existing outlier detection scheme with respective data mining and address some drawbacks:

a) Distance- based method:

- Operate on whole data. Cannot give number of clusters.
- Computation time will increases.
- Give only one value as most expected outlier.

b) Clustering and Distance based method:

- Reduce the size of database that will reduces computation time.
- To each cluster user can give certain radius to find outliers.
- Can group the data in to number of clusters.

6. CONCLUSION

Data streams are unbounded flow of data and these are emerging from many real world applications. Data stream mining is one of top ten challenges in front of data mining community. This papers aims to detect outliers is the task that finds objects that are dissimilar or inconsistent with respect to remaining data. This survey paper provides motivations to handle the streaming data challenge. Our approach needs to be implemented on more complex datasets.

Some techniques require a priori knowledge about data distribution in dataset such as distribution based methods. Assumption based method can work quite well if prior assumption made about data is correct.

Finally, for further reading, we direct the reader to a recent book on outlier analysis [1] and [7] for a tutorial version of this survey.

7. REFERENCES

[1] C. C. Aggarwal, Outlier Analysis. Springer, 2013.

[2] M. Gupta, J. Gao, C. C. Aggarwal, and J. Han, "Outlier Detection for Temporal Data," in Proc. of the 13th SIAM Intl.Conf. on Data Mining (SDM), 2013.

[3] Yogita and D. Toshniwal, "A framework for outlier detection in evolving data streams by weighting attributes in clustering," in Proceedings of the 2nd International Conference on Communication Computing and Security, India, 2012.

[4] Sadik, S. and Gruenwald, L. 2010. DBOD-DS: Distance Based Outlier Detection for Data Stream. DEXA' 10.

[5] Parneeta Dhaliwal, MPS Bhatia and Priti Bansal,‖ A Cluster-based Approach for Outlier Detection in Dynamic Data Streams (KORM: k-median OutlieR Miner)‖ JOURNAL OF COMPUTING, VOLUME 2, ISSUE 2, FEBRUARY 2010, ISSN: 2151-9617.PAGES 74-80.

[6] Anguiulli, F. and Fassetti, F. 2007. Detecting Distance-Based Outliers in Streams of Data. CIKM' 07. Pages 811 - 820.

[7] F. Angiulli and F. Fassetti, "Detecting Distance-based Outliers in Streams of Data," In Proceedings of CIKM'07, Pages 811-820, November 6-10 2007.

[8] J. Han and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, San Francisco, 2006.

[9] Jiang, N. and Gruenwald, L. 2006. Research issues in Data Stream Association Rule Mining. ACM SIGMOD RECORD, Volume 35, Issue 1. Pages 14 -19.

[10] V. Hodge and J. Austin, A Survey of Outlier Detection Methodologies, Artificial Intelligence Review, Vol. 22, pp. 85-126, 2003

[11] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "A framework for projected clustering of high dimensional data streams," in Proceedings
of the Thirtieth international conference on Very large data bases –Volume 30, ser. VLDB '04. VLDB Endowment, 2004, pp. 852–863.

[12] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "A framework for clustering evolving data streams," in Proceedings of the 29th international conference on Very large data bases - Volume 29, ser.VLDB '03. VLDB Endowment, 2003, pp. 81–92.

[13] C. C. Aggarwal and P. S. Yu., 2001. Outlier detection for high dimensional data. In Proc. 2001 ACM-SIGMOD Int.Conf. Management of Data (SIGMOD'01), pp37-46.

[14] Ramaswamy S., Rastogi R., Kyuseok S.:Efficient Algorithms for Mining Outliers from Large Data Sets,Proc. ACM SIDMOD Int. Conf. on Management of Data, 2000.

[15] Knorr, E.M., Ng, R.T., "Finding Intentional Knowledge of Distance-Based Outliers", Proceedings of the 25th International Conference on Very Large Data Bases, Edinburgh, Scotland, pp.211-222, September 1999.

[16] E. M. Knorr and R. T. Ng. ―Algorithms for mining distance based outliers in large datasets‖ In Proc. 24th Int. Conf. Very Large Data Bases, VLDB, pages 392–403, 1998.

[17] V. Barnett and T. Lewis, Outliers in Statistical Data, New York: John Wiley Sons, 1994.

[18] D.M. Hawkins, Identification of Outliers, London: Chapman and Hall, 1980.

[19] http://archive.ics.uci.edu/ml/