The Hybrid Approach for Handling and Detecting Outliers from Dynamic Data Stream.

Mr. Raghav M Purankar, Prof. Pragati Patil

Abstract— The Outlier detection is currently area of active research in data set mining community. In this article we propose hybrid approach to capture outliers in dynamic data stream. We apply k-mean algorithm which Partition the data set into number of chunks or clusters. Each chunk contains set of data. Once cluster are formed, centroid of each cluster are calculated. The points which are lying near the centroid of the cluster are not probable candidate outlier and we can prune out such points from each cluster. Next distance based technique is used to find the distance from centroid to candidate outlier. For that threshold value is set. If this distance is greater than threshold value then it will declare as outlier otherwise as a real object. In proposed approach, two techniques are combining to efficiently find the outlier from the data set. This hybrid approach takes less computational cost. Proposed algorithm efficiently prune of the safe cells and save huge number of extra calculations.

Index Terms— Minimum Centroid, Cluster based, Data Stream, Distance based, K-means, Outlier Detection, Pdist.

1 Introduction

The Detection and removal of data elements with inconsistent behaviour is recently an important and active research problem in many fields and involved in various use. Most of the present procedure is based on distance amount. Because of progressive nature of the incoming data, declare a set often can lead us to a wrong decision. However, earlier research for the problem of set detection is suitable for disk on a long-term datasets where the entire dataset is available in advance. But, outlier detection over data set is a demanding task because data is continuously updated. Finding outlier set in a collection of patterns is a very well-known problem in the data mining field. An outlier set is a collection of outliers which is different with respect to the rest of the patterns in the dataset. In some cases being present of outliers are will affect the results drawn out of the examinations and hence need to be removed in advance. There are different reasons for outlier detection such as due to rare normal events showing entirely different features etc. Outlier set is the data point that does not comply with the rules and to the normal points specified in the data set. Finding points with abnormal behaviour among the data points is the basic idea to find out an outlier. Outlier detection also called Anomaly detection deals with detecting data elements from a data set which is different from all the other data elements in a set. Anomalous data points can occur because of different reasons such as mechanical faults, other changes occur throughout the system, error occurs in equipment, fraudulent behaviour, human error or natural causes and effects. In general data point with anomalous behaviour are more interesting and need excess examination.manuscript.

When In this proposed work, we are mentioning about clustering method that will identify candidate outliers, the points which may contain outliers also called as Temporary Outliers). Temporary outliers may contain inliers, in order to filter the outlier, next apply distance based for all candidate outliers, which is used to identify a point to be an outlier or not. The main objective is to find the outlier from the static data set using cluster based method and distance based method and then to apply the hybrid approach of both the methods for dynamic data stream. Efficient detection of outliers minimizes the risk of making wrong decisions based on error contained data, and aids in identifying, preventing, and repairing from the effects of anomalous behaviour of data elements. Additionally, many data mining and machine learning algorithms and techniques for statistical analysis may not give proper results in the presence of outliers. Correct and well organized removal of outliers may greatly increase the analysis of statistical data, data mining algorithms and techniques. Detecting and removing such outliers as a before processing each step for other techniques is known as data cleaning. The large and infinite amount of data streams in this field may lead to false representation of outliers or anomalies. The serious and major issue arises when we segmenting the data into the number of blocks also called as clustering. For the detection of the outlier first the data must be segmented into the number of blocks and after that each block is compared with another one for getting the candidate set of outlier. Outlier set detection as a branch of data mining has many important applications and justifies more attention from data mining communities. Moreover due to rapid stream evolution, data element property can change over time. This introduces to main problem first as only one scan is possible to process data points and secondly due time constraints data is considered as an evolutionary stream.

1.1 Objectives

The The detection of anomalous points or outlier detection is an extremely important task in a various kinds of application domains. To accomplish the above task and for system development the following objectives are identified:

1. To propose the detail System Architecture of outlier detection.
2. To identify the safe region which are pruned out from the current chunk.
3. To Design and implement a distance based outlier detection approach on each clusters with respect to centre of cluster.

International Journal of Scientific & Engineering Research Volume 7, Issue 11, December-2016
ISSN 2229-5518

2 Literature Review

With increase in dimensions of high streaming data, sophisticated approach and algorithms need to be proposed to handle the process of outlier detection over dynamic data streams. This section discusses a variety of existing approaches proposed in earlier literature for outlier detection. Traditional methods for outlier detection can be effectively apply and gives greater accuracy of outlier detection on statically downloaded dataset. Traditional and existing data mining approaches cannot be applicable directly to streaming data efficiently as these methods are suitable where the entire dataset is statically available and algorithm can operate in more than multiple passes. The evolution of data streams led to the change in the characteristics of the data streams like dimensionality, In data streaming, a point which is an outlier in current phase may become an inliers in the next phase [2]. So in data stream, it is incorrect to declare the outlier at the very initial stage since the data stream is continuous flowing in nature, so outliers predicted may be treated as inliers in the next phase of data stream.

Elahi, M. KunLi, Nisar, W. XinjieLv, HonganWang, in the text about a clustering based method, it split the stream into chunks and for cluster each chunk is given input to the k mean algorithm for fixing the number of cluster. In this method the author take the temporary outlier for consideration and the average value of each cluster is found out for the next incoming data streaming chunks, in order to ensure that the detected outlier are the real outliers[3].

Luis Torgo, Carlos soares, et al., proposed a method by considering the hierarchical clustering approach to execute the task of outlier detection. The method is verified on the statistical dataset and the foreign trade transaction data, in which the data is collected from the statistics institute.

Safal V Bhosale et al proposes article for streaming the data proposed a clustering based unsupervised outlier detection approach. For taking the advantages of both that is density based and distance based outlier detection, the proposed system combine both approaches density based and clustering based method. The mining tasks uses weighted k-means clustering[4].

T. Divya, Dr. T. Christopher et al have analysed the performance of CURE with K-Means and CURE with CLARANS clustering based algorithm for outlier detection. In order to consider the best clustering algorithm for outlier detection some of the performance measures are considered such as Accuracy of outlier detection, Detection rate[5].

- *Mr. Raghav Purankar is currently pursuing masters degree program in Computer science and engineering in Nagpur University, India, E-mail: raghavpurankar88@gmail.com*
- *Prof. Pragati Patil is currently working in Computer science and engineeringdepartment AGPCE college in Nagpur University, India, E-mail: pragatimit@gmail.com*

3 Proposed Implementation Architecture

The Proposed implementation architecture can be described with the help of following steps:

1. Partition the data set into number of chunks and each chunk contain set of data.
2. Over each chunk, apply clustering algorithm to find out candidate or temporary outliers and safe region.
3. Apply distance based outlier detection approach over clusters by finding centroid of cluster.
4. Give a chance to the temporary or candidate outlier to combine in next incoming set, and allow it for appropriate number of set chunks, and then declare candidate outliers as real outliers or inliers.

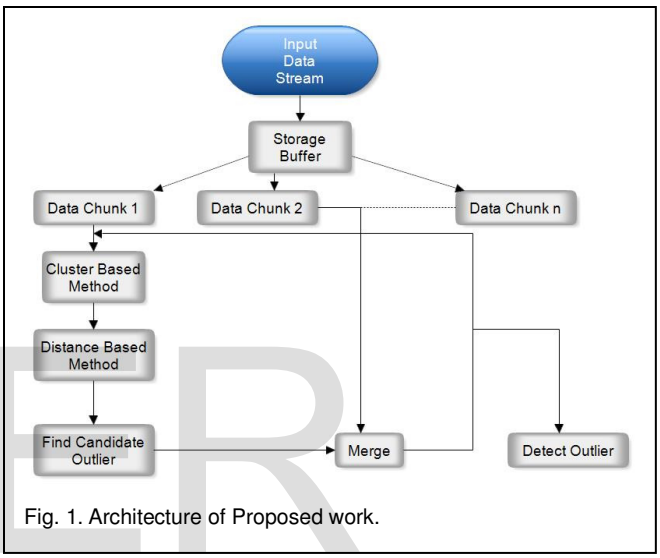


Fig. 1. Architecture of Proposed work.

If it is assumed that, the number of outliers in any dataset is expected to be extremely small as compared to the normal data. So, it is not considerable to apply the traditional outlier detection algorithms over the entire data set, especially in case of data set this method can become highly expensive as well as can often led us to wrong decision in finding most outstanding outliers. Proposed method only declares these points as candidate outliers and compares them with next incoming data set chunk to make sure that these are real outliers.

In the first phase, the cluster based method is applied; safe data will be pruned out while the candidate cells will still be kept there for further processing. Safe region are pruned out from the current chunks so that next chunks can allow for storing. Then in second phase, distance based strategy is applied over the candidate to efficiently figure out the outliers while discard rest of data. Proposed algorithmic approach efficiently find out the safe cells and save huge number of extra calculations.

This proposed work presented a solution to the problem of outlier detection. The key idea and main contributions of this work lie in the proposed clustering and distance based technique. Outlier Detection is the task that finds objects that are dissimilar or inconsistent with respect to remaining data. The Clustering algorithm is first performed, and then small clusters are determined and considered as candidate outlier. Other

International Journal of Scientific & Engineering Research Volume 7, Issue 11, December-2016
ISSN 2229-5518

outliers are then determined based on distance measures. In our view, using clustering method here act as a data reduction and distance based method that calculate distance from centroid.

3.1 Techniques Used

1. Dividing Input data stream into Manageable chunks.

As the system is capable of handling large input data stream, but it is highly advisable to divide or to cluster the input dynamic data stream into manageable data chunks so that hybrid approach can be suitably applied over each data chunks individually. Each data chunks may be able to give fewer numbers of outliers which are called candidate outliers because these outliers are not final outlier. These candidate outliers are then combined with next incoming data chunk and again applied the hybrid approach and then find out the outlier from that data chunks. This process goes on repeating till the end of data stream or user explicitly stops execution of hybrid algorithm. The basic idea behind the clustering of large input data stream into manageable chunks is to facilitate the system to prune out the real outliers efficiently and easily. The basic idea behind the clustering or dividing large input data stream into manageable chunks is that it is easy to manage the small data chunk instead of storing entire data stream into the system for further processing, since the large portion of memory is required to store dynamic data stream.

2. Combined hybrid approach over each chunk.

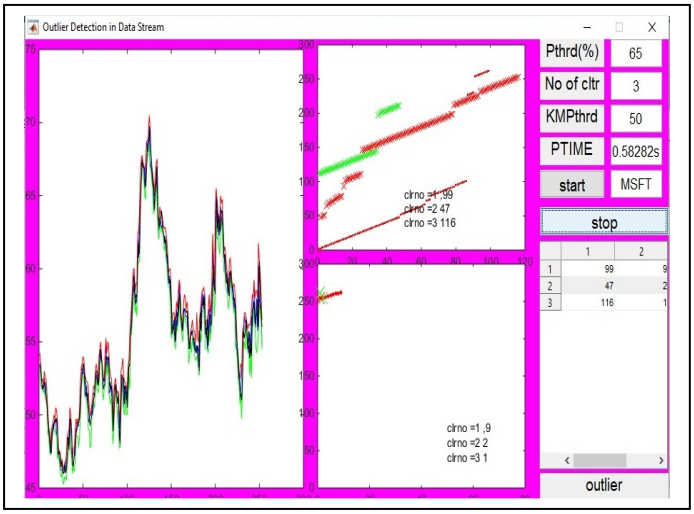
The hybrid approach simply combines the cluster based and distance based approach into one combined algorithm unlike static approach that includes two individual approaches. The clustering of input dynamic data stream into manageable chunks enable the system to apply the hybrid approach over each data chunks and can be considered as single cluster. Hybrid approach is applied over each data chunk to find out the candidate outliers from that data chunk and then these candidate outliers are then combined into next incoming data chunk, again apply the same hybrid approach over that data chunk and again find the outlier. This process is continuing till the end of data stream. The hybrid approach process fetch the data from the Yahoo finance web site after a specific time interval divide that data stream into small data chunks and then hybrid approach is applied to each data chunk. Thus in this concept, it is notable that the hybrid approach is capable of handling data stream and efficiently prune the outliers from the dynamic data stream which is the main goal of the developed system. Traditional algorithms are not capable of handling the data stream, Thus the hybrid approach is capable to overcome the drawbacks of existing system up to certain mark.

4 Proposed Hybrid Method on Dynamic Data Stream

In Hybrid approach of outlier detection, The two approaches that is cluster based approach and distance based approach is used and applying both the approaches as a single module. Threshold value is set as per the user choice. Cluster based algorithm K-means is used initially in order to divide the real time data stream into desired number of clusters, Since the data stream is continuous and infinite sequence of data, the clustering algorithm divides large data stream into manageable chunks each having certain number of data elements. Then apply K-means algorithm on each data chunk and subsequently distance based approach in order to find out the distance of each data element within that cluster from the centroid. Now the threshold value given by user is used for setting the cluster radius and find out the data element those have maximum distance from cluster centroid generally considered as candidate outliers and elements having minimum distance can be considered as inliers. In proposed hybrid approach, the online real time data is being fetch from "Yahoo Finance" website for stock exchange information. The dynamic real time data is fetch from year 1986 to 2013 for a particular stock. The stepwise approach can be described as follows:

1. To fetch the input dynamic data stream and divide it into manageable data chunks.
2. To partition the data stream into specific number of clusters using cluster based approach.
3. To apply distance based approach over each cluster by comparing threshold value.
4. To separate the inliers and candidate outliers
5. To Add probable candidate outliers in the next available data chunk.
6. To Repeat the same procedure from point no.2 till the end of input data stream to find out the set of Final outliers.

The proposed work uses hybrid approach to formulate the cluster and Outlier Detection using dynamic data stream which is continuous and infinite flow of data. The main aim of proposed work is to predict outlier detection on numeric data by using existing cluster based and distance approach and trying to carry out the outlier detection using combined or hybrid approach over the same input data stream [8]. Hybrid approach is used basically in order to reduce the computational cost and resource utilization, thus improves the overall system performance.



Above fig. shows the actual screenshot of how the window will appear when the online data is being fetched and dis-

International Journal of Scientific & Engineering Research Volume 7, Issue 11, December-2016
ISSN 2229-5518

played in panel. Leftmost panel shows the Stock data from "Yahoo Finance" data for specific ticker named as "MSFT" for Microsoft Inc with four values as "Open", "High", "Low" and "Close". Topmost panel shows the Clustering process of each data chunk and underneath panel shows the outliers. The online stock data is to be displayed in tabular format within right most panel. In this process we are providing the timing frame of Five seconds, our system will fetch the real time data after only specific time interval i.e. after 5 seconds each time and repeats the same process as shown in above figure above. In this way we are trying to fetch the online data stream and apply the hybrid approach in our project. Hybrid Approach consists of two combined approaches that is cluster based and distance based approach.

5 Conclusion

The approach here used is hybrid approach to capture outliers from continuously flowing data stream. By considering aspects, relating to the existing approaches for outliers detection, it is concluded that traditional methods were not suitable for handling streaming data. From the studying and taking help from all existing methods, this system proposes the customized or hybrid approach for outlier detection in dynamic data stream. The developed system shows that hybrid approach is capable of handling streaming data. Due to this approach it takes less computational cost. It also prunes of the safe cells and save huge number of extra calculations. So, after analyzing the proposed hybrid approach, we conclude that, the hybrid approach is able to process dynamic data which is less computationally expensive than output of processing of individual algorithms.

References

[1] Manish Gupta, Jing Gao, Charu C. Aggarwal and Jiawei Han, "Outlier Detection for Temporal Data: A Survey", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 1, JANUARY 2013.

[2] Abhishek B. Mankar, Namrata Ghuse, "A Review on Detection of Outliers Over High Dimensional Streaming Data Using Cluster Based Hybrid Approach", International Journal of Science and Research (IJSR), ISSN (Online): 2319-7064, Volume 3 Issue 11, November 2014.

[3] Elahi, M. KunLi, Nisar, W. XinjieLv, HonganWang, "Fuzzy Systems and Knowledge Discovery", Fifth International Conference on Data Mining Vol.5, and Vol.3, pp. 23 - 27, 2002.

[4] Safal V Bhosale A Survey: Outlier Detection in Streaming Data Using Clustering Approached et al "International Journal of Computer Science and Information Technologies(IJCSIT)", Vol. 5 (5) , 2014, 6050-6053.

[5] T. Divya, Dr. T. Christopher, "A Study of Clustering Based Algorithm for Outlier Detection in Data streams", Int. Jnl. Of Advanced Networking and Applications(IJANA),March 2015.

[6] Dr. S. Vijayarani and Ms. P. Jothi, "Detecting Outliers in Data streams using Clustering Algorithms", International Journal of Innovative Research in Computer and Communication Engineering, Volume. 2, Issue 8, October 2013.

[7] Ms. S. D. Pachgade, Ms. S. S. Dhande, "Outlier Detection over Data Set Using Cluster-Based and Distance-Based Approach", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 6, June 2012.)

[8] Prakash Chandore, Prashant Chatur, "Outlier Detection Techniques over Streaming Data in Data Mining: A Research Perspective", International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Volume-2, Issue-1, March 2013.

[9] SREEVIDYA S S, "Detection of Outliers in Data Stream Using Clustering Method", International Journal of Science, Engineering and Technology Research (IJSETR) /2015/2278-7798/Volume 4.

[10] 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.

[11] D.Joice, K. Lakshmi and K. Thilagam, Comparison Of Cluster Based Algorithms For Outlier Detection In High Dimensional Dataset, Karpagam Journal of computer science, Volume 8, issue 3, April 2014.

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

[13] 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.

[14] Niketa V. Kadam, Prof. M.A. Pund, "Joint Approach for Outlier detection", International journalof computer science and Applications, Vol 6, No. 2, Apr 2013.