INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

ANALYTICAL APPROACH FOR DETECTION OF OUTLIERS IN TIME VARIANT DATA STREAM USING HYBRID TECHNIQUE

MR. MUKESH K. DESHMUKH [1], PROF. A. S. KAPSE [2]

1. M. E. Scholar, Department of CSE, P. R. Patil College of Engineering & Technology, Amravati , M.S. India.
2. Asst. Prof. , Department of CSE, P. R. Patil College of Engineering & Technology, Amravati , M.S. India.

Accepted Date: 15/03/2016; Published Date: 01/05/2016

Abstract*:* Data mining provides us a broad research area as most of the work is concentrated around the organization and operation related to data. But prior to taking data into consideration, all of us should ensure that the data should be free from any error-prone contents. So, there need to be some techniques for ensuring that data provided to the system should be well arranged and bug free. Data may contains some objects behaves differently than rest of elements in dataset, are termed as **Outliers**, and the process of detecting such elements is called as **Outlier Detection**. Most of the existing methods to detect outliers useful only on predefined datasets. The proposed work discussed about approach to detect outliers in stream data. In this work, we discussed a customized algorithm, a combination of partitioned based and distance based algorithm, for detecting outliers in continuous time varying data stream, which provide us a more precise and accurate results than individual traditional algorithm. The main contribution of this work is to reduce the false detection rate and tries to improve the outlier detection accuracy from streamed data. **[2] [9]**

Keywords: Cluster-based, Distance-based, Data Mining, Dynamic data Stream, K-Means, Outlier Detection.

Corresponding Author: MR. MUKESH K. DESHMUKH

Access Online On:

www.ijpret.com

How to Cite This Article:

Mukesh K. Deshmukh, IJPRET, 2016; Volume 4 (9): 1227-1235

PAPER-QR CODE

INTRODUCTION

Data Mining provides us a very Broad spectrum in the research area as data management gained a widespread attention in every field. As data is a very crucial part of any organization, no organization allows to handle data in inconvenient manner. So many researchers day by day tries to designed new techniques which are used to handle the data in a proper and convenient manner and to provide the error-free data for the further processing. Outlier detection is a process proposed by data mining community researchers to provide a more efficient and error-free data. Day by day, the researchers tried to improve the technique used by the for outlier detection as the form of data is changed from a conventional form to the real time data. However, Earlier techniques used for outlier detection is more suitable for the disk resident data sets or static data set values where all datasets are available in advance and algorithm can operate in multiple phases.[1] Outliers are candidates that should not obey the normal behavior as follows by rest of elements in that particular data chunk. Outliers Detection is defined as finding an anomalous patterns which does not obey the properties follows by normal data elements. [5][16][17]

Up till now, we are only dealing with the data which is readily available with us as and when required for our study purpose. It is also a case whenever data is continuously flowing in nature, and if we want to process that data, then it has to be done on a real time basis. As the data is dynamic in nature, and if it is not processed immediately or stored, then it is lost forever. [5] Earlier work for outlier detection focused only on detection rate of outliers elements while keeps ignoring the an important issue related to data stream mining like, low memory requirements and processing with high speed algorithms to keep pace with high volume, unbounded data streams. [13]

By analyzing all requirements, researchers reached to the point, that there is need of more advanced method than only partitioned based and distance based methods. In this paper, new technique is discussed knows as Windowing technique, which is highly helpful in managing data which is dynamic in nature. Section 2 provides surveys of related work in the domain of outlier detection. Section 3 provides the details of technique used for outlier detection. Section 4, provides experimental results analysis and paper will concluded in section 5.

1.1 Motivation:

Analyzing incoming data properly and to deal with real time data are the prior fundamental task of data stream mining. In addition with fundamental task of data mining, outlier detection can also be considered as important task in data analysis i.e., mining useful and interesting information from a huge amount of data. Data stream is likely continuous flow of data in and out. To deal with problem of processing streamed data, efficient outlier detection method needs to be incorporated. Unfortunately, Traditional outlier detection techniques are not find more suitable for handing dynamic nature of data. [9]

2. RELATED WORK

Many authors round the globe, already provided the numbers of literature based on Outlier detection methods. Many of them discussed distance based, density based, partition based techniques, Sliding Window Based Outlier Detection and much more new concepts in order to provide solutions for the above mentioned problems in real time streaming data.

Surekha Peshtwar et. al [1] discussed about the hybrid approach over the dynamic streamed data which is highly useful in real time work environment. According to them, division of streamed data into chunks at that particular time instance is done. Over each data chunk, hybrid approach, which consist of Cluster based and distance based approach, is applied which resulting into clustering of data points. Due to application of distance based approach, by defining threshold for each cluster, pruning out points which lying outside the radius. In this way, the data points which lie outside the threshold will considered as final candidate outlier. Rajendra Pamula et.al [4] presented a clustering based method to capture outliers. Here they apply K-means clustering algorithm to divide the data set into clusters. They concluded that the points which are lying near the centroid of the cluster are not probable candidates for outlier and can pruned out such points from each cluster. Based on the outlier score obtained, they declare the top n points with the highest score as outlier elements. They claimed that the experimental results using real data set demonstrate that even though the numbers of computations are less, their proposed method performs better than the existing outlier detection methods. Sreevidya S S [2] converse about the link anomaly detection technique which detects outlier in data stream dynamically by using clustering method. They used probability model and Dynamic link optimization algorithm for detecting outliers. Through their research they claimed that the accuracy of this method is more than other methods presented by other authors. All the methods used by them in their proposed work require only linear time against the length of analyzed time period.

3. PROPOSED METHODLOGY

3.1. System Architecture

Data Stream is evolutionary in nature so new concept came into existence and old concept is becomes obsolete. Due to this characteristic of data stream it is difficult to collect labeled data for processing. So for detecting outliers from data streams, unsupervised approach is more suitable compared to supervised approach. Cluster-based algorithms are unsupervised in nature, hence it does not require knowledge of data in advance. In this approach, K-Means algorithm is used for clustering data stream. First stream data of specified size is entered into the window. [5] [17]

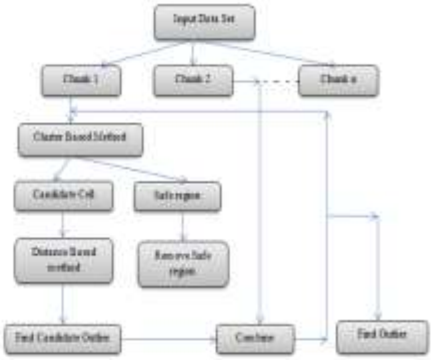


Fig 1: Outlier detection over dynamic data stream [1]

Proposed system discussed about windowing technique for Outlier detection, which gives working steps as follows:

- Partition the incoming streamed data into no. of chunks and each chunk contain set of real time data.

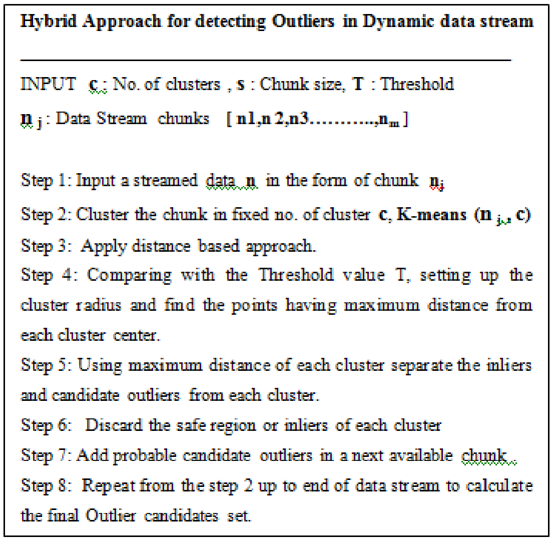
- Over each data chunk, apply clustering method to figure out probable candidate outliers and safe region. • Apply distance based outlier detection algorithm over clusters with respect to centroid of cluster.

- Find out the most probable outlier candidate and add it into next available data chunk ready for processing in next round.

- Likewise, process all data chunks to find out the final outlier set. [1]

So, it is highly inefficient to apply the traditional outlier detection algorithms over the entire data set, especially in case of high amount of data this method can become highly expensive as well as can often led us to wrong decision. 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. [1] Proposed approach do not suffer from high computational cost, which is always a problem when distance based approaches are applied over high speed huge volume of data streams. Proposed algorithm efficiently done pruning of the safe cells and saves much number of extra calculations.

3.2 Proposed Algorithm: [1]



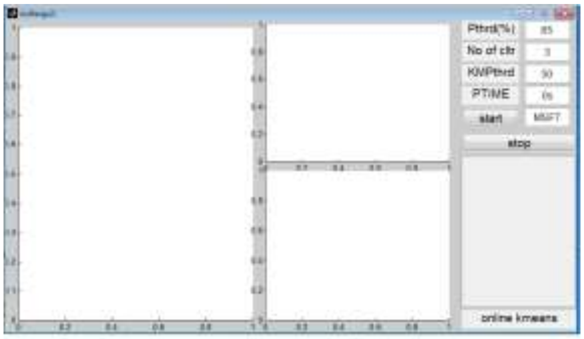
```
Hybrid Approach for detecting Outliers in Dynamic data stream

INPUT c : No. of clusters , s : Chunk size, T : Threshold
n j : Data Stream chunks [ n1,n 2,n3..........,nm ]

Step 1: Input a streamed data n in the form of chunk nj
Step 2: Cluster the chunk in fixed no. of cluster c, K-means (n j, , c)
Step 3: Apply distance based approach.
Step 4: Comparing with the Threshold value T, setting up the cluster radius and find the points having maximum distance from each cluster center.
Step 5: Using maximum distance of each cluster separate the inliers and candidate outliers from each cluster.
Step 6: Discard the safe region or inliers of each cluster
Step 7: Add probable candidate outliers in a next available chunk.
Step 8: Repeat from the step 2 up to end of data stream to calculate the final Outlier candidates set.
```

4. EXPERIMENTAL RESULTS

The experimental results using dynamic data stream demonstrate that even though the number of computations are less, proposed method performs better analysis than the existing detection method. In order to carry out experimental analysis, we use MATLAB tools for implementing our algorithms. Our Experiments was conducted in MATLAB Version 7.11.0.584 (R2010b) on various Static data sets and Dynamic streaming data.



Pthrd(%) 85
No of cltr 3
KMPthrd 50
PTIME 05
MSFT

Fig 2: GUI for processing Dynamic data Stream

While performing result analysis for Dynamic data stream, we used stock data of Microsoft Inc. which is available at Yahoo Finance website. In order to process real values data, we were applying our customized or hybrid approach over this data stream. As the data is continuously flowing nature, somehow we have to divide that data and store in some set. In first phase, we divide streamed data into chunk which is of variable size. Over each data chunk, we apply our hybrid approach. As in hybrid approach, we have to provide the number of cluster required, here we input e.g. c=3. So the incoming data chunk is divided into 3 clusters. Here we assumed that the data element lying near to the centroid is the real data, so we need to prune out this safe region records. By removing safe region candidate we find out maximum distance element by applying threshold value to maximum distant element for each cluster and finds outside points. Once the threshold has find, the data which lie within that threshold limit will consider as "real" data. So we will prune out that real data and the data which lie outside the threshold will considered as candidate outlier. But these outlier candidates still contain some real data and outlier value both. So, in order to get more filtered data, we need to apply Distance based approach. In this, we will find the distance of each element with respect the remaining point and then counts the number of objects which lies within some threshold value. [1]

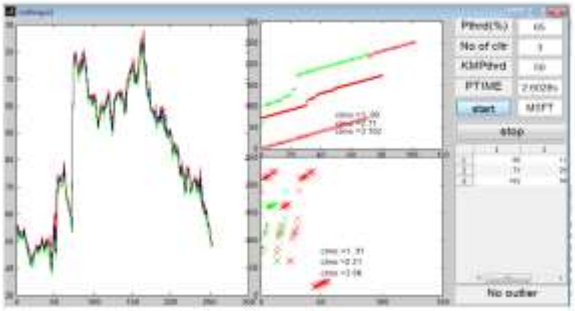
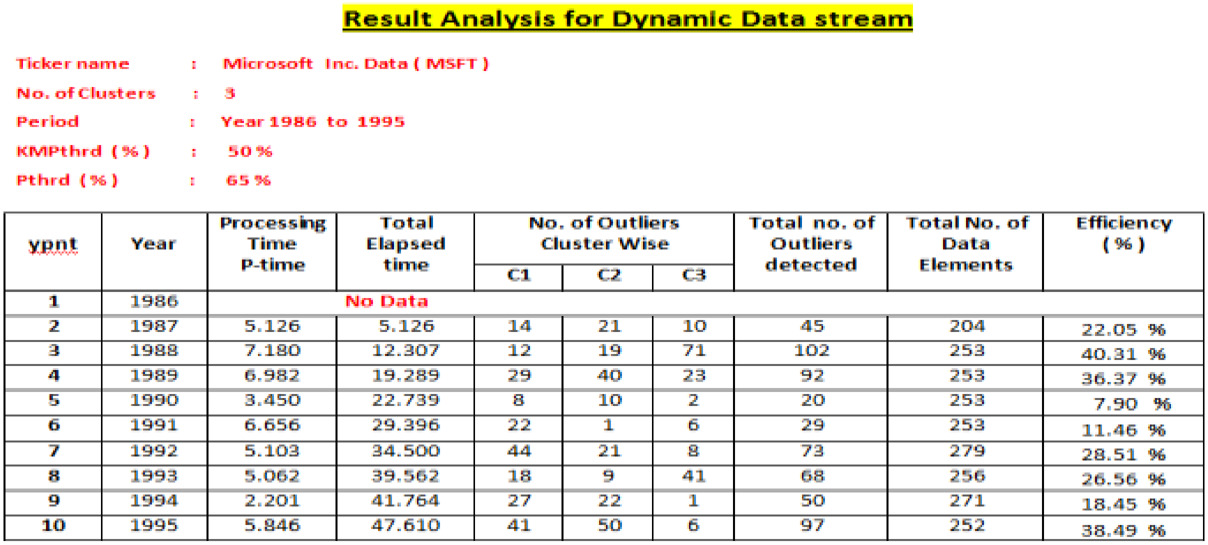


Fig 3: Outlier candidates marked using proposed approach [1]

In figure 3, it shows three subplots, in first panel, complete stock data is shown graphically. In second subplot, it shows the result of cluster based approach for 3 clusters with the no. of elements in each cluster. We marked candidate outlier value which we get as final outlier after applying the distance based approach which is marked within third subplot.

4.2 Analysis with Dynamic Data Stream processing



Result Analysis for Dynamic Data stream

Ticker name	:	Microsoft Inc. Data (MSFT)
No. of Clusters	:	3
Period	:	Year 1986 to 1995
KMPthrd (%)	:	50 %
Pthrd (%)	:	65 %

ypnt	Year	Processing Time P-time	Total Elapsed time	No. of Outliers Cluster Wise			Total no. of Outliers detected	Total No. of Data Elements	Efficiency (%)
				C1	C2	C3			
1	1986	No Data							
2	1987	5.126	5.126	14	21	10	45	204	22.05 %
3	1988	7.180	12.307	12	19	71	102	253	40.31 %
4	1989	6.982	19.289	29	40	23	92	253	36.37 %
5	1990	3.450	22.739	8	10	2	20	253	7.90 %
6	1991	6.656	29.396	22	1	6	29	253	11.46 %
7	1992	5.103	34.500	44	21	8	73	279	28.51 %
8	1993	5.062	39.562	18	9	41	68	256	26.56 %
9	1994	2.201	41.764	27	22	1	50	271	18.45 %
10	1995	5.846	47.610	41	50	6	97	252	38.49 %

Table 4.1 Table for Dynamic DataStream Processing

5. CONCLUSION AND FUTURE SCOPE

By considering all the pros and cons, related to the existing methods of Outlier detection, we can stated that the traditional methods used for processing static data sets are not suitable for real time streaming data. There is much need for the more improved technique to handle a dynamically streamed data. By taking the help from all the literature, we proposed our modified or customized approach for the detection of outlier in a dynamic data stream. So, after analyzing the proposed approach, we conclude that, the modified approach is very efficient for the processing dynamic streamed data which is less computationally expensive than output of processing of any individual algorithms. For future work, we need to improve our approach to make it more time and space efficient. proposed approach only able deals with numerical data from our datasets. So, future work requires some modifications to deal with data mining problems and make it compliant for textual mining also.

REFERENCES

1. Surekha V Peshawar & Snehlata Dongre, "Outlier Detection Over Data Stream Using Cluster Based Approach And Distance Based Approach", International Conference on Electrical Engg. and Computer Science (ICEECS-2012), Trivendum, May 12th, 2012.

2. SREEVIDYA S S , "Detection of Outliers in Data Stream Using Clustering Method ", International Journal of Science, Engineering and Technology Research (IJSETR) Volume 4, Issue 3, March 2015 .

3. Mr. B. Muruganantham, Ms. Ankita Dubey , "Outlier Detection Using Distributed Mining Technology In Large Database", SSRG International Journal of Computer Science and Engineering (SSRG-IJCSE) – volume 2 issue 2 February 2015.

4. Rajendra Pamula, Jatindra Kumar Deka, Sukumar Nandi, "An Outlier Detection Method based on Clustering " , 2011 Second International Conference on Emerging applications of InformationTechnology. 2011 IEEE.

5. Prashant Chauhan , Madhu Shukla , "A Review on Outlier Detection Techniques on Data Stream by Using Different Approaches of K-Means Algorithm", 2015 International Conference on Advances in Computer Engineering and Applications (ICACEA) , IMS Engg. College, Ghaziabad, India @ 2015 IEEE

6. A. Mala , F. Ramesh Dhanaseelan , "Data Stream mining Algorithms – A Review of issues and Existing approaches", International Journal on Computer Science and Engineering (IJCSE) Vol. 3 No. 7 July 2011.

7. Deevi Radharani , Navya Dhulipala , Tejaswi Pinniboyina, " Outlier Detection for Dynamic Data stream using Weighted K-Means ", International Journal of Engineering Science and Technology (IJEST), Vol. 3 No.10 October 2011.

8. Pooja Thakkar , Jay Vala , Vishal Prajapati, "Survey on Outlier Detection in Data Stream ", International Journal of Computer Applications (0975 – 8887) Volume 136 – No.2, February 2016 .

9. Mr. Mukesh K. Deshmukh , Prof. A. S. Kapse , " A Survey On Outlier Detection Technique In Streaming Data Using Data Clustering Approach", International Journal Of Engineering And Computer Science, Volume 5 Issue 1 January 2016, Page No. 15444- 15447 .

10. M. Gupta, J. Gao, C. C. Aggarwal, and J. Han, “Outlier Detection for Temporal Data", in Proc. of the 13th SIAM International Conference on Data Mining (SDM), 2013. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 1, JANUARY 2013

11. 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.

12. 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

13. Manzoor Elahi, L v Xinjie, M. Wasif Nisar, Hongan Wan, "Distance Based Outliers for Data Streams using Grid Structure" , Information technology Journal 8(2) , 128-137, 2009.

14. 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.

15. 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.

16. Rohini Balkrishna Gurav, Sonali Rangdale, "Hybrid Approach for Outlier Detection in High Dimensional Dataset ", International Journal of Science and Research (IJSR) , Volume 3 Issue 7, July 2014

17. Hossein Moradi Koupaie, Suhaimi Ibrahim , Javad Hosseinkhani , “Outlier Detection in Stream Data by Clustering Method ”, International Journal of Advanced Computer Science and Information Technology (IJACSIT) , Vol. 2, No. 3, 2013, Page: 25-34