www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 5 Issue 1 January 2016, Page No. 15453-15456

A Survey On Outlier Detection Technique In Streaming Data Using Data Clustering Approach

Mr. Mukesh K. Deshmukh ^{1}***, Prof. A. S. Kapse*** ^{2}

*1 M.E Scholar, Department of CSE, P. R. Patil College of Engg. & Technology, S.G.B. Amravati University, Amravati (Maharashtra) – India .
E-mail : deshmukh.mukesh24@gmail.com
&
2 Department of CSE , P. R. Patil College of Engg. & Technology, S.G.B. Amravati University, Amravati (Maharashtra) – India.
E-mail : arvind.kapse@yahoo.com

ABSTRACT:

Data mining is a highly researched area in the today's world as data is crucial part of many application, due to which many researchers express their interest in this domain. As there arises a need to process large dataset which imposes different challenges for researchers. To have a data which is free from a noisy attributes , known as a filtered data , is of much important to gain accuracy in a result sets. For that , finding and eliminate the noisy objects has gained a much more importance. An object that does not follow the footprints of usual data object is called **outliers**. Outlier detection process is used in numerous applications like fraud detection, intrusion detection system, tracking environmental activities, healthcare diagnosis. Numbers of approaches are used in the process of detection of outlier. Most approaches focuses to use Cluster-based and Distance based approach (i.e. using K- Means algorithm and Euclidian distance) for outlier detection in data sets which help them to create a group of similar elements or cluster of data points. Clustering techniques are highly useful for grouping similar data items from data sets and after that by applying distance based calculations, detection of outlier is done, so they are called cluster-based outlier detection. K- Means and Euclidian distance are the most common and popular algorithm for clustering and outlier detection process due to its simplicity and efficiency. Different application areas of outlier detection are discussed in this paper.

Keywords *- Cluster-based, Dataset, Distance-based, Dynamic data Stream, K-Means, Outlier Detection.*

1. INTRODUCTION

Outlier detection is currently an active and important research problem facing by the many data mining researchers and involved in number of applications . Due to time variant nature of the incoming data; declaring an outlier often can lead us to a wrong conclusion. But However, earlier research done over a mentioned problem of outlier detection is more suitable for static data sets where the entire dataset is readily available and algorithms can operates over multiple passes. But, outlier detection over dynamic data set is a very challenging task because data is continuously updated and flowing.

Finding outliers from a collection of data is a very well-known problem in the domain of data mining. An object that does not follow the footprints of usual data object is called **outliers**. In other words, outlier is a pattern which is not similar with respect to the rest of the patterns in the dataset. Depending on the application domain, outliers are of particular interest. Detecting outliers may lead to the discovery of truly

unexpected behavior and help avoid wrong conclusions etc. Which gives us a filtered and cleared data to operate on.
In particular, distance-based techniques use the distance function for relating each pair of objects of the data set. Distance-based definitions [7] represent an useful tool for data analysis [8].

1.1 Objectives :

Main objective is the Detection of Outlier in a Static data sets as well as in and continuously data stream which gives us filtered data with which we can carry out experiments on application consisting a huge data sets .

1.2 Motivation :

Analyzing datasets and Clustering analysis are the prior fundamental task of data 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 data flow in and out. To deal with problem of processing streaming data, efficient outlier detection method need to be used. Unfortunately, Traditional outlier detection techniques are not find suitable for handing dynamic nature of data .

2. LITERATURE SURVEY

[A] In the paper, ***Rajendra Pamula, Jatindra Kumar Deka, Sukumar Nandi, "An Outlier Detection Method based on Clustering", 2011 Second International Conference on Emerging Applications of Information Technology .*** **[3]**
author discussed a clustering based method to capture outliers. Here we, apply K-means clustering algorithm to divide the data set into clusters. The points which are lying near the centroid of the cluster are not a probable candidates for outlier and we can pruned out such points from each cluster. Based on the outlier score obtained, we declare the top *n* points with the highest score as outliers. The experimental results using real data set demonstrate that even though the number of computations is less, the proposed method performs better than the existing outlier detection methods.

[B] In the paper, ***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 Engineering College, Ghaziabad, India [5]***

Author discussed, Various approaches used to achieve the mentioned goal, 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. So they are called cluster based outlier detection. Purpose of this paper is to review of different approaches of outlier detection which is used for K-Means algorithm for clustering dataset with some other m

3. PROPOSED WORK :

3.1 Proposed Methodology using K-means Approach

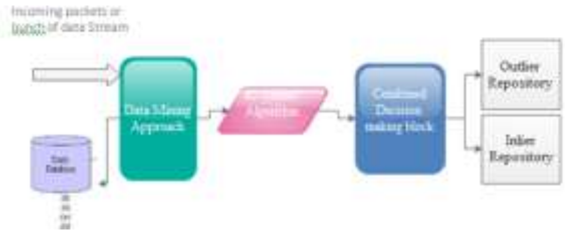


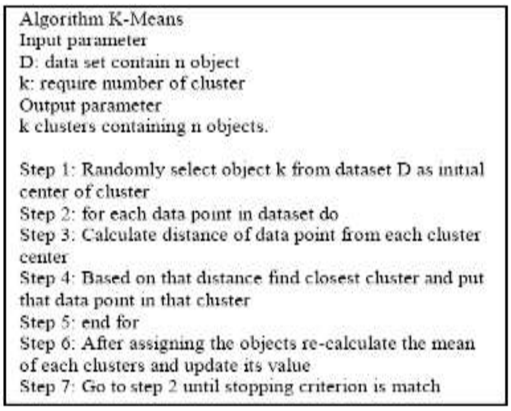
Fig.1 Proposed Outlier detection System using k-means algorithm

Generating cluster : *K*-means clustering is a partition based m methodology . Initially, cluster the whole dataset into k clusters 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) Randomly select object *k* from dataset D [5] as initial centroid of cluster.
b) For each data point in dataset, Calculate distance of that data point from center of each cluster.

c) Based on the distance, put that data point in the nearby cluster.

d) Go back to Step b, algorithm stop when no more data points to move or stopping criterion match.



```
Algorithm K-Means
Input parameter
D: data set contain n object
k: require number of cluster
Output parameter
k clusters containing n objects.

Step 1: Randomly select object k from dataset D as initial
center of cluster
Step 2: for each data point in dataset do
Step 3: Calculate distance of data point from each cluster
center
Step 4: Based on that distance find closest cluster and put
that data point in that cluster
Step 5: end for
Step 6: After assigning the objects re-calculate the mean
of each clusters and update its value
Step 7: Go to step 2 until stopping criterion is match
```

Fig.2 : K-Means Algorithm [6]

3.2 Distance based Approach :

Distance based approach mainly uses Euclidian distance calculation. With high dimensional dataset, calculate distance with each instances will increase the computational cost. We are comparing distance based method with proposed method. [16]

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

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 data 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.

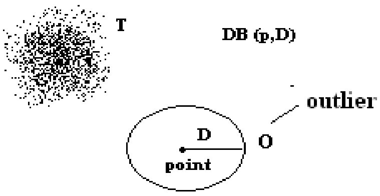


Fig. 3 Basic model of distance based method

4. DISCUSSION :

4.1 Outline

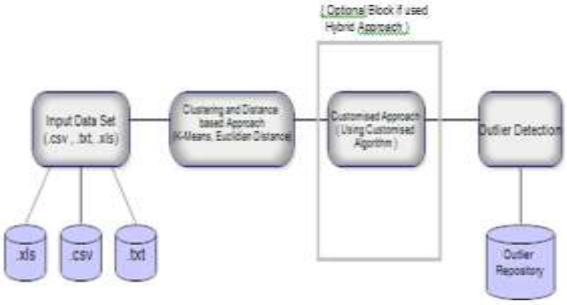

(Optional Block if used Hybrid Approach)
Input Data Set (.csv , .txt .xls)
Clustering and Distance based Approach (K-Means, Euclidian Distance)
Customised Approach (Using Customised Algorithm)
Outlier Detection
.xls
.csv
.txt
Outlier Repository

Fig. 4 : System Overview

In this paper, we discussed methodology implemented, cluster-based and distance based method to capture outliers.[22] Here, We apply K-means clustering algorithm to divide the data set into number of clusters. The data points which are lying near the centroid of the cluster are not probable candidate for outlier and we can prune out such points from each cluster. In Next step, we calculate a distance based outlier score for remaining points. Based on the outlier score we declare the top *n* points with the highest outlier score as outliers. The experimental results using real time data set demonstrate that even though the number of computations is less, hence the proposed method performs better than the existing method.

5. CONCLUSION :

Data streams are dynamic flow of data and these are emerging from many real world applications. Data stream mining is one of great challenges faced by the data mining community. This papers aims to detection of outliers, which are nothing but finding objects which behaved differently than the rest of objects in that dataset. The proposed survey paper provides idea about previous work done regarding outlier detection and gives us motivation to deals with the dynamic streaming data challenge. Some techniques require a priori knowledge about data distribution in dataset. Assumption based method can work quite well if our prior assumption made about data is correct.

6. REFERENCES :

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

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

[3] *Rajendra Pamula, Jatindra Kumar Deka, Sukumar Nandi* , "An Outlier Detection Method based on Clustering", *2011 Second International Conference on Emerging Applications of Information Technology .*

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

[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 Engineering College, Ghaziabad, India* @ 2015 IEEE

[6] *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.*

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

[8] *S. Ramaswamy, R. Rastogi, and K. Shim. Efficient algorithms for mining outliers from large data sets. pages 427–438, 2000.*

[9] Han, Jiawei, and Micheline Kamber. Data Mining, Southeast Asia Edition: Concepts and Techniques . Morgan kaufmann, 2006.

[10] *Neeraj Chugh, Mitali Chugh, Alok Agarwal et al "*Outlier Detection in Streaming Data A research Perspective*", International Journal of Science, Engineering and Technology Research (IJSETR)Volume 4, Issue 3, March 2015.*

[11] *T. Divya, Dr. T. Christopher et al* "A Study of Clustering Based Algorithm for Outlier Detection in Data streams", *International Journal Of Advanced Networking and Applications (IJANA), ISSN 0975-0282, March 2015.*

[12] *Safal V Bhosale et al* "Outlier Detection in Straming data Using Clustering Approached " , *International Journal of Computer Science and Information Technologies(IJCSIT), Vol. 5 (5) , 2014, 6050-6053.*

[13] *Nancy Lekhi, Manish Mahajan* "Outlier Reduction using Hybrid Approach in Data Mining", *International Journal of Modern Education and Computer Science , May 2015, 5, 43-49.*

[14] *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".*

[15] *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.

[16] *Mr. Raghav M. Purankar, Prof. Pragati Patil "*A Survey paper on An Effective Analytical Approaches for Detecting Outlier in Continuous Time Variant Data Stream " , *International Journal Of Engineering And Computer Science ISSN: 2319-7242, Volume 4 Issue 11 Nov 2015, Page No. 14946-14949*