Detection of Outliers in Data Stream Using Clustering Method

SREEVIDYA S S
M.Tech Student
Department of Computer Science and Engineering
Sarabhai Institute of Science and Technology
Vellanad, Trivandrum, India

***Abstract*- Outlier detection is an active area for research in data set mining community. An outlier is a pattern which is dissimilar with respect to the rest of the patterns in the data set. Detecting outliers and analyzing large data sets can lead to discovery of unexpected knowledge in area such as fraud detection, telecommunication, web logs, and web document, etc. A lot of outlier detection methods were exists and most of them are based on distance measure. To declare an outlier as it arrives often can lead us to a wrong decision, because of dynamic nature of incoming data. Outlier detection in streaming data is very challenging because streaming data cannot be scanned multiple times. Most of the existing methods detect outliers only on predefined datasets. The proposed work uses an algorithm to detect outliers in stream data by clustering methods and which is concentrate to find outliers dynamically. In this work Dynamic Threshold Optimization algorithm is used for detecting outliers dynamically. The main contribution of this work is to reduce the false detection rate and improve the outlier detection accuracy.**

***Keywords*-Outliers, data mining, data stream, fraud detection**

I. INTRODUCTION

Outlier Detection over streaming data is active research area from data mining that aims to detect object which have different behavior, exceptional than normal object [1]. Identifying outliers within data lead to the discovery of useful and meaningful knowledge or improve data analysis for further discover within numerous applications domains, it also helps to avoid a wrong conclusion.

Outlier is also called as anomaly due to behavior of object with respect to other data elements. The term outlier is originated from Statistics [2]. Outliers are patterns in data that do not conform to a well defined notion of normal behavior. Figure 1 represents outliers in a 2D data set. Most observations lie in N1and N2 regions. Here N1 and N2 are the normal regions. Points that are far away from the regions are outliers. O1, O2 and O3 are the outliers in this figure.

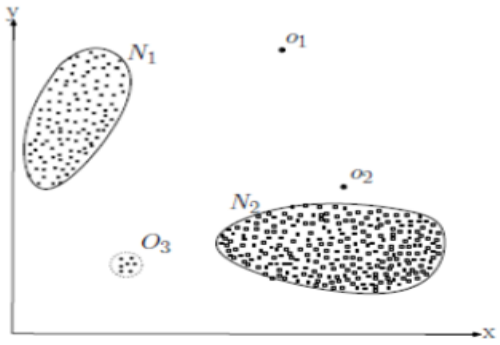


Fig 1: example of outliers in a 2-D data set

A lot of work for outlier detection has been done in data mining [7] community using conventional methods. These conventional methods are more suitable over static data set. Such methods can be used for streaming data but these methods are not able to process data with single pass. As dimensionality increase traditional method takes high computing time and cannot provide an efficient result over analysis of streaming data. Effective outlier detection requires the construction of a model that accurately represents the data. A large number of techniques have been developed for building models for outlier and anomaly detection. However, the real world data set, data stream presents a range of difficulties that bound the effectiveness of the techniques. The assumed behavior of outliers is they were different from other members of cluster or they does not belong to any cluster, or belong to very small clusters, or forced to belong a cluster [3]. The clustering techniques are highly helpful to detect the outliers they are called cluster based outlier detection.

In this paper a outlier detection algorithm is proposed to detect outliers dynamically using clustering method. The proposed system helps to detect outliers in stream data as it arrives and also this system helps to reduce the false detection rate.

The remaining paper is organized as follows: Section 2 discusses related work and the existing methods for outlier detection. In Section 3, the proposed scheme is described. Section 4 presents

ISSN: 2278 – 7798

the system architecture and section 5 concludes the paper.

II. RELATED WORK

Outlier detection aims to find patterns in data that do not conform to expected behavior. Over the years, a large number of techniques have been developed for building models for outlier detection. The clustering techniques are highly helpful to detect the outlier and they are called cluster based outlier detection. The clustering based techniques involve a clustering step which partitions the data into groups which contain similar objects. A number of clustering algorithms have been introduced in recent years for data streams [4]. Density based clustering method can produce outlying objects along with normal cluster. The clustering based detections are in unsupervised in nature [8], so they do not require knowledge of data in advance. In Deviation based approach identify outliers by inspecting the characteristics of objects and consider an object that deviate these features declared as an outlier**.** Also there exist a number of methods for tracking changes of clustering structures from non stationary data sequences.

A statistical test based algorithm [5] for dynamic clustering exists, which estimates a GMM [9] in an online manner and then conducts a statistical test to determine whether a new cluster is identical to the old one or not [6]. If it is new cluster, it is merged to the old one, otherwise it is recognized as a cluster which has newly emerged. In this framework a Dynamic Model Selection (DMS) is used for tracking changes of statistical models from non stationary data. It can be applied to tracking of changes of clustering structures. Also the conventional approaches for topic detection have mainly been concerned with the frequencies of words.

Majority of approaches to detect anomalies in data mining consider the batch framework, some researchers have attempted to address the problem of online outlier detection. Traditional methods for outlier detection can produce good results on stored static data set. The distance based clustering is not suitable if clusters have different densities. The presence of noisy attributes conceals real clustering structure of data and hence leads to lower outlier detection rate and higher false alarm rate.

Although there are lot of research on outlier detection, but there is little research in direction of outlier detection in dynamic data streams. This area needs lot of attention because the existing methods are not appropriate in the stream environment. In this work a dynamic link anomaly based technique is introduced, which clusters the data streams dynamically and detect the anomalies.

III. PROPOSED WORK

The proposed link anomaly detection technique detects outlier in data stream dynamically by clustering method. It uses probability model and dynamic link optimization algorithm for finding outliers. The accuracy of this method is more than other methods. The advantage of the proposed scheme over existing method is dynamic detection of outliers in data stream using clustering method is more efficient. The proposed link anomaly based change point detection is highly scalable. All the methods used in proposed work require only linear time against the length of analyzed time period.

The proposed scheme is constructed in following four steps

i) Training dataset
ii) Clustering of class
iii) Link anomaly score
iv) Dynamic link optimization

A. TRAINING DATASET

In this module, a training set is created by filtering the data randomly. Training set consists of several classes. In which each class consists of several clusters. The data with similar properties are grouped into clusters. The clusters in each class have some similar features. Data from various areas were used for clustering. For getting good results, the filtering of data is done multiple times. The training set should be close as possible to actual data. In addition, the cluster structure was formed in a way that new clusters can be added to the structure online.

The training set consists of several classes such as sports, education, entertainment etc. Each class consists of several clusters and data with similar properties are organized in each cluster. For example, consider the sports class, which consists of clusters named cricket, football, hockey etc. Each cluster consists of data which have similar features. User has the option to add clusters and features to the set. Many classes and features are added to the training set for the proper functioning of the set.

Thus in training set there are several classes and each class consists of several clusters. Also the training set consists a repository for storing outliers. Outlier is an object which is different from normal object.

B. CLUSTERING OF CLASS

In this module related objects are grouped into clusters. Clustering based techniques involve a clustering step which partitions the data into groups which contains similar objects. Clustering is used to improve the efficiency of the result by making groups of the data. The goal of a clustering algorithm is to

ISSN: 2278 – 7798

group objects into meaningful subclasses. Clustering can be used to generate class labels for a group of data. For clustering data, in this work k-means algorithm and a probability model is used. Clustering refers to task of assigning class label c to an unlabeled data x. It is performed by calculating probability distribution over class assignments P(c|x). Using Baye's rule

$$P(c|x) \propto P(c)\, P(x|c) \qquad (1)$$

where P(c|x) is prior and P(x|c)---conditional probability of data estimated from training set .

1)Probability model

The probability model introduced here is used to capture the features of data and how to train the model. Probability specifies how the concerned properties of the observed data can be generated from the models. Probability model is the most widely used "tools" for modeling the uncertainty with theoretical backbone and widespread application and acceptance. Given two parameters k and V, k is the data given by user and V is the set of data which have mentioned behavior.

Formally, consider the following joint probability distribution:

$$P(k, \mathrm{V} \mid \theta, \{\pi_v\}) = P(k \mid \theta) \prod_{v \in V} \pi_v \qquad (2)$$

The probability P(k|θ) is defined as a geometric distribution with parameter θ as follows:

$$P(k \mid \theta) = (1-\theta)^k \theta \qquad (3)$$

The probability of given data v Є V is denoted by π_v (where the sum of π_v must be 1, $= \sum_v \pi_v = 1$).

Next, find out the predictive distribution of both training data and new data (referenced data).Suppose if n training data T= $\{(k_1,V_1),\ldots.(k_n,V_n)\}$are given , predictive distribution is learned as follows:

$$P(k, V \mid T) = P(k \mid T) \prod_{v \in V} P(v \mid T) \qquad (4)$$

Accordingly the probability of known data is given as follows:

$$P(v \mid T) = \frac{m_v}{m+\gamma} \quad \text{for v: } m_v \geq 1 \qquad (5)$$

On the other hand, the probability of new data is given as follows:

$$P(\{v : m_v = 0\} \mid T) = \frac{\gamma}{m+\gamma} \qquad (6)$$

where γ=0.5

Thus we get probability for known data and new data. So the data in each clusters is assigned with a probability. By using k-means algorithm, the unorganized data are organized into several clusters and by using probability model assign probabilities to the data in the clusters.

C. LINK ANOMALY SCORE

Anomaly score is computed for each user depending on current data of user and past data. To compute anomaly score of a new data x=(t,u,k,V) compute probability with training set $T_u^{(t)}$, which is collection of data by the user in the time period [t- T, t]. To measure the general trend proposes to aggregate anomaly scores.

A score function that would generate high scores for outliers would assign scores to a point that is inversely proportional to the sum of the strengths of its entire links.

Accordingly, the link anomaly score is defined as follows:

$$s(x) = -\log(P(k \mid T_u^{(t)}) \prod_{v \in V} P(\mathrm{v} \mid T_u^{(t)}))$$

$$= -\log(P(k \mid T_u^{(t)}) - \sum_{v \in V} \log(P(\mathrm{v} \mid T_u^{(t)}) \qquad (7)$$

1)Combining Anomaly scores

The anomaly score is computed for each data depending on the current data given by the user and the behavior of past data $T_u^{(t)}$. To measure the general trend of user ,propose aggregate the anomaly scores obtained for data $x_1,\ldots..,x_n$, using a discretization of window size τ >0 as follows:

$$s_j^{'} = \frac{1}{\tau} \sum_{t_i \in [\tau(j-1), \tau j]} s(x_i) \qquad (8)$$

where x_i= (t,u,k,V) is the data at time t_i by user u_i including k_i behaviors to users V_i.

D. DYNAMIC LINK OPTIMIZATION

One dimensional histogram is used in DTO for the representation of the score distribution. The dynamically optimized threshold ηt (j) at time step j is defined as the least score value such that the tail probability above ηt (j) is no greater than ρ. Here ρ is called as *significance level parameter*.

The details of DTO are summarized as follows: let N_H be a given positive integer. Let {q(h)

ISSN: 2278 – 7798

(h=1,….N_H): $\sum_{h=1}^{N_H} q(h) = 1$ } be a one-dimensional histogram with N_H bins, where h is an index of bins, with a smaller index indicating a bin having a smaller score.
The procedures for updating the histogram and DTO are given in Algorithm.

Algorithm: Dynamic Threshold Optimization (DTO)

Given: $Score_j \mid j = 1, 2, ...\}$**:** scores, N_H: total number of cells, λ_H: estimation parameter, ρ: parameter for threshold, r_H: discounting parameter, M: data size

Initialization: Let $q_1^{(1)}(h)$ (a weighted sufficient statistics) be a uniform distribution

for j=1,…, M-1 **do**

Threshold optimization: Let l be the least index such that $\sum_{h=1}^{l} q^{(j)}(h) \geq 1 - \rho$. The threshold at time j is given as $\eta(j) = a + \frac{b-a}{N_H - 2}(l+1)$

Output: if $Score_j \geq \eta(j)$

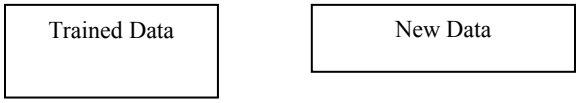
$$q_1^{(j+1)}(h) = \begin{cases} (1-r_H) q_1^{(j)}(h) + r_H & \text{If Score}_j \text{ falls into the h}^{\text{th}} \text{ cell,} \\ (1-r_H) q_1^{(j)}(h) & \text{otherwise} \end{cases}$$

$$q_1^{(j+1)}(h) = (q_1^{(j+1)}(h) + \lambda_H) / (\sum_h q_1^{(j+1)}(h) + N_H \lambda_H).$$

end for

Using Dynamic Threshold algorithm the outliers are detected and they are kept in a repository. The detection of outlier is done by comparing the anomaly score and threshold value. The data which have score less than threshold is organized into the corresponding cluster and the data which have greater score is detected as outlier.

IV. SYSTEM ARCHITECTURE



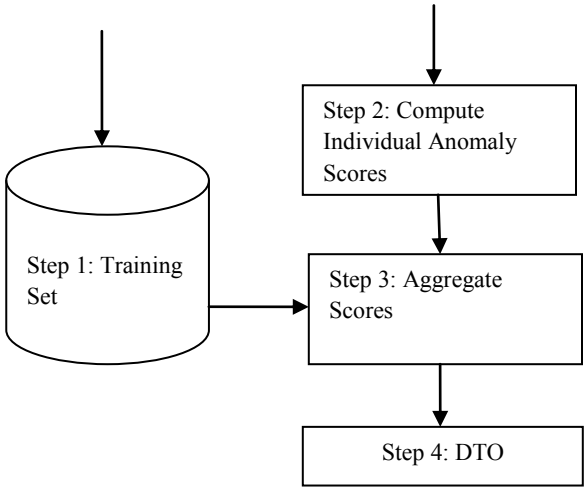


Fig 2: architecture of outlier detection in dynamic data streams using clustering method

In the proposed work, for each new data , use samples within the past time interval of length T for corresponding user for training the model (step 1).Then assign an anomaly score to each data based on learned probability distribution (step 2). The score is then aggregated (step 3). And further fed into SDNML- based change point analysis (step 4). Using DTO the outliers are detected (step 4).

V.RESULTS

In the proposed work detection of outliers is done dynamically. For efficient detection of outliers as soon as it arrives it is needed to make an efficient training set. If the number of training set increases the success ratio also increases. Because while an user click on a link he/she can easily identify the classification result and also detect whether it is an outlier or not. By using Dynamic Threshold Optimization Algorithm easy detection of outlier is possible.

VI. CONCLUSION

An outlier is a pattern which is dissimilar with respect to the rest of the patterns in the data set. A large number of techniques have been proposed in outlier detection area, but most of them have some inherent limitations. The conventional methods detect outlier only based on predefined datasets. To face the challenges of data stream processing the proposed scheme is dynamic in nature. The proposed clustering based outlier detection method gives higher outlier detection rate.

ACKNOWLEDGEMENT

ISSN: 2278 – 7798

This work was supported in part by the Department of Computer Science & Engineering, SIST, Trivandrum. I would like to show my gratitude to Dr. C G Sukumaran Nair (HOD), Associate Professor, Sudha SK and Assistant Professor, Vini Vijayan, for their valuable guidance.

REFERENCES

[1] Bakar, Z. A., Mohemad, R., Ahmad, A., & Deris, M. M.(2006), *"A comparative study for outlier detection techniques in data mining*", In Proc. 2006 IEEE Conf. Cybernetics and Intelligent Systems, pp. 1–6, Bangkok, Thailand.

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

[3] Dr. S. Vijayarani,Ms.P.Jothi, *"An Efficient Clustering Algorithm for Outlier Detection in Data Streams"*, International Journal of Advanced Research in Computer and Communication Engineering Vol. 2,Issue 9,pp. 3657-3665, 2013

[4] Madjid Khalilian, Norwati Mustapha, *"Data Stream Clustering: Challenges and Issues",* IMECS 2010.

[5]M. Song and H. Wang,*"Highly efficient incremental estimation of Gaussian mixture models for online data stream clustering*",Intelligent Computing: Theory and Application, 2005.

[6] So Hirai,Kenji Yamanishi, *"Detecting changes of Clustering Structures Using Normalized Maximum Likelihood Coding"* ,KDD'12,August 2012

[7] P.N. Tan, M. Steinback, and V. Kumar, *Introduction to Data Mining*, Addison Wesley, 2006.

[8]Yogita, Durga Toshniwal*,"Unsupervised Outlier Detection in Streaming data Using Weighted Clustering"*, International Science index Vol.6,No:11,2012

[9]K. Yamanishi et al, 2004. "*On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms"*. In Proceedings of Data Min. Knowledge Discovery. Vol. 8, No. 3, pp 275-300.

ISSN: 2278 – 7798