ISSN:0975-9646

A Survey: Outlier Detection in Streaming Data Using Clustering Approached

Safal V Bhosale
CSE, MIT, Aurangabad

Abstract — **In the public field like network intrusion detection, credit card fraud detection, stock market analysis .The theory of outlier detection is used for detecting the outlying in the medical data. As we know that streaming data 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 and such attribute magnify the challenge. In this paper, we propose the method for streaming data, the method is known as unsupervised outlier detection method. Clustering is one of the unsupervised data mining task, hence this method is based on the theory of clustering and label data is not required in this method. In the proposed method for taking the advantage of both the density based outlier detection and density based outlier detection the scheme combined the application of both density based clustering and partitioned based clustering. Depending upon their respective relevance in mining task the proposed scheme assign weights to attribute. This weighted attributes are very helpful to reduce or remove the effect of noisy attributes. For concept evaluation the proposed scheme is incremental and adaptive for keeping the view challenges of streaming data. Our proposed approach outperforms other existing approach in terms of outlier detection rate, and the increasing percentage of outlier.**

Keywords: Irrelevant attributes, Streaming Data, Unsupervised Outlier detection.

I. INTRODUCTION

In our day to day life streaming data was generated by the many applications such as online transactions, remote sensors, medical systems, real time surveillance. The two words data streams and streaming data are synonymous to each other. 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 incredible volume. As the time going on the new concepts 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.

The outlier detection in the data mining task is also known as outlier mining. An outlier is an object which does not comply with the behavior of normal data objects. Outliers are more interesting than normal cases, the application which proves this statement true are network intrusion detection, fault diagnosis in machines, fraud in credit card detection, detecting outlier cases in wireless sensor network data.

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. Comparing with the supervised approaches unsupervised data mining approaches are more feasible. Clustering based outlier detection method does not required knowledge of data in advance because the clustering based outliers are unsupervised in nature. There are two clustering methods which are density based and partitioned based clustering. The first clustering method can produce outlying objects along with normal cluster. Second clustering method used for distance based outlier detection.

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 which is having the attribute in noisy. This will happen because we know that performance of clustering outlier detection method is depend upon the quality of clusters discovered. The real clustering structure of data is conceals with the presence of noisy attribute and hence leads to lower outlier detection rate and higher false alarm rate.

In this paper for streaming the data we proposed a clustering based unsupervised outlier detection scheme. For taking the advantage of both the density based and distance based outlier detection the proposed system combine the both density based and partitioned clustering method. The proposed scheme assigns weights to the attributes. The weight is assign depending on their respective relevance in mining tasks using weighted k-means clustering.

Rest of the paper is organized in the following manner: In Section II we discussed about the related work done on the outlier detection method. In section III we discussed about the background and problem ideation and in section IV the proposed approached. In section V we discussed about the conclusion and at the last references are given.

II. LITERATURE SURVEY

In the research community the outlier detection is the very interesting topic [1], [5], [2], [8], [9], [10], [11]. Ramasway proposed a method which is known as distance based outlier detection method [8]. According to this method, the parameter which are given as l and m, an object is an outlier if no more than n-1 other object in the dataset have higher value for Dl than the object a. Dl (a) denotes the distance of nth nearest object of a. this idea is further experimented, in which each data point is ranked by the sum of distance from its n nearest neighbors. After that Breunig introduced the notion of the local outlier factor which is termed ad LOF. In this proposed method it captures the relative degree of outlierness of an object. All

the above described methods are not suitable for outlier detection in the data streams because they are either distance based or the nearest based distance. The new definition of outlier which was presented by The et al which was named as cluster based local outlier, this method provides importance to the local data behavior. After that Duan et al proposed an algorithm which is named as cluster based outlier detection algorithm. This algorithm detects both the single point outlier and cluster based outlier.

All the methods which we are discussed above are used for storing static data sets. These methods are not applicable in the data stream environment. For resolving this problem the exact-STORM and approx-STORM algorithm are presented for detecting the distance based outliers the notion of sliding window the outlier detection techniques are proposed. The proposed techniques are very much depending upon the selection of window size. Sadik was proposed the method which is known as distance based outlier detection for data streams. This method detects outliers which are based on the two parameters which are radius neighbor density and minimum neighbor density. This method is not handling the concept of evolution in the streaming data. Elahi represented the cluster based outlier minera method. This is a clustering based approach for the outlier detection which is based on the k-mean. This method divides data stream in the chunks for the further processing. Yogita has proposed a framework for outlier detection in evolving data streams by weighting attributes in clustering. This method is the clustering based framework that assigns weights to all attributes depending on their respective relevance clustering.

Let us see some literature with their proposed work.

[12], Proposed the clustering algorithm named as CURE, the algorithm is used for detecting an outliers. The algorithm achieved by presenting point per clusters, the method allows the algorithm to adjust the geometry of the non spherical shapes and the reduction helps to decrease the result of outliers. The mixture of random sampling and partitioning and the experimental results verify that the quality of clusters produced by the algorithm is much better than those found by existing algorithms. Moreover, the authors expressed the partitioning and random sampling enable CURE to not only outperform existing algorithms but also to scale well for large databases without sacrificing the quality of cluster.

[13] conversed about a clustering based method, it split the stream into chunks and for cluster each chunk use the k mean algorithm for fixing the number of cluster. In this method the author take the outlier of candidate and the average value of each cluster for the next fixed number of data streaming chunks, to make sure that the identified candidate are the genuine outliers. The average is used in the cluster of previous streaming chunks and the current chunk of average values are taken to be the deliberation [14], discussed about the three alternative of the k means algorithm to the cluster binary data streams. Variants are on line K means, Scalable Means, on line K means and Incremental K means proposed a variant introduced that finds higher quality solution in less time. All variants were compared with real and synthetic data sets. The proposed Incremental K Means variant is faster than the already quite fast Scalable K means and finds solution of comparable quality. The K means variants are compared with respect to quality of speed and results. The proposed algorithms can be used to check the transactions.

[15], discussed about the algorithm of k means for clustering of data streams and recognition of outliers. The method which has been used for outlier detection is based on distance as well as on time, on which they appear in the cluster. The author receives into account the choice of k centers and variable size of buckets with the help of which space can be efficiently use at the time of clustering. Most established algorithms makes very difficult problem in clustering by falling their quality for a better competence.

[16], Proposed an unsupervised outlier detection method for the streaming data. The scheme is based on clustering as clustering is an unsupervised data mining task and it does not need labeled data. In this scheme both densities based and partitioning clustering method are mixed to take benefit of both density and distance based outlier detection. It assigns weights to attributes depending upon their respective relevance in mining task

[17], Proposed a method for the purpose of hierarchical clustering method to execute the task of outlier detection. The method is tested on the official statistic data and the foreign trade transaction data, in which the data is collected from the statistics institute. In this method author discuss about the outlier ranking method and its results.

[18], discuss the k means based data stream clustering algorithm termed as BICO, the algorithm evaluate the high quality solution in a short time, it also evaluate a summary K of the data with the demonstrable quality. The comparison of the algorithm with the popular algorithm BIRCH and Mac Queen Algorithm is done.

[19], discuss the number of methods and procedures for the detection of outliers, also the working of outlier with the distinguished among the uniform versus multivariate methods and parametric versus non parametric procedure.

III. BACKGROUND AND PROBLEM FORMULATION

The algorithm used for the clustering the current data chunk is DBSCAN. This algorithm does not require number of clusters and can find the arbitrary shape clusters. The output which is obtained from the DBSCAN algorithm is the set of clusters and outlying objects. The outlying object is considered s candidate outliers and it is used as input to outlier detection for further verification of their outlying nature. Small size clusters of DBSCAN may be group of outliers or these may be portion of a cluster that yet to be come in next data chunk and has been split over to chunks. These objects are treated as the candidate outliers and feed to outlier detection module. For clustering the current data chunk weights of previous phase are taken and that are then updated in weighted k-mean clustering module. DBSCAN parameters using equation following equations

$$\text{Epsilon} = \frac{\sum_{i=1}^{k} Avg\ Intra\ (Ci)}{k}$$

Avg Intra (Ci) $=\frac{\sum_{i=1}^{n}\sum_{j\neq i}^{n} Dist(Oi,Oj)}{2\times n}$

Min Pts = Avg No. of objects with a distance of epsilon from an object in cluster of smallest density'

Where Density (Ci) = $\frac{No.of\ Objects\ (Ci)}{Radius\ (Ci)}$

In the above equation k is the number of clusters, n is the total number of objects comprises all clusters. Ci represent ith cluster and Oi represent ith object. Dist () is the distance between two objects. We have taken Euclidian distance in our experiment for implementation. Based on these equation parameters of DBSCAN are updated in MinPts and epsilon updation selection.

IV. PROPOSED APPROACH

In this section the proposed scheme is presented as follows As we know that the data stream is limitless string of data. It is impossible to store the complete data stream, for that reason we split the data stream into data chunk of the same size. User decide the size of the chunk, this size is depend upon the nature of the data. In this method the buffer system is used i.e. when the current chunk is processed at that time the incoming data is stored in the buffer and after some period of time used as a data chunk. After storing required statistics, processed data chunk is deleted at the end of the processing iteration to empty space for next incoming data chunk.

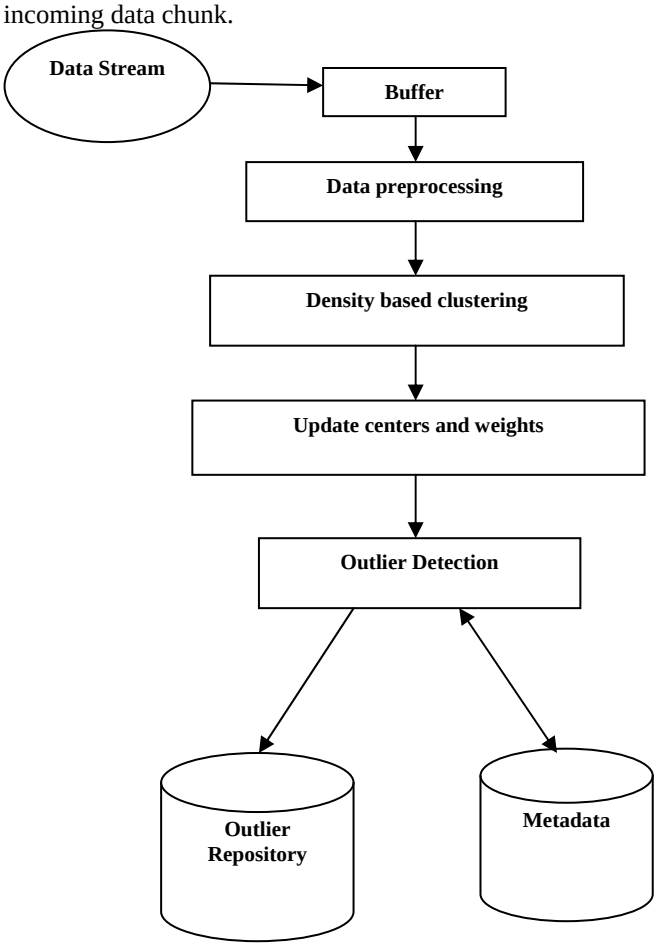

Figure 1. Block diagram of unsupervised outlier detection in streaming data using weighted clustering

V. CONCLUSION

In this paper we proposed the clustering based unsupervised outlier detection method for the streaming data. The method applied both the density based clustering method and partition based clustering method for detection of individual and group of outliers. The weighted k-means clustering method is used for assigning the weights to the attribute. The method is depending upon their respective relevance in clustering. The proposed method is incremental and dynamic in nature for facing the challenges of data stream at processing. Streaming data is process in the form of data chunks and candidate outliers are checked over multiple consecutive data chunks before declaring them as outliers or inliers. At the time of processing the data chunks which are necessary and the chunkis discarded to free up memory for next chunk. In the future we will try to implement advance level of our method for the data types like categorical and mixed.

REFERENCE

[1] F. Angiulli and Fassetti, 'detecting distance based outliers in data streams', n Proceedings of the sixteenth ACM conference on information and knowledge management.

[2] M. S. Sadik and L. Gruenwald, DBOD-DS: Distance based outlier detection for data streams. Springer, 2011.

[3] L. Duan, L. Xu, Y. Liu, and J. Lee, 'Cluster based outlier detection', Europian journal of scientific research.

[4] J. Z. Huang, M. K. Ng, H. Rong, and Z. Li, 'Automated variable weighting in k-means type clustering', IEEE Tran pattern Anal. Match Intell.

[5] S. Subramaniam, T. Palpanas, D. Papadopoulos, V. Kalogeraki, and D. Gunopulos, 'online outlier detection in sensor data using non-parametric models', in proceeding of the 32nd international conferenceon very large data base.

[6] F. T. Liu, K. M. Ting, and Z.-H. Zhou, 'isolation based anomaly detection', ACM trans. Knowledge discov Data.

[7] A. Frank and A. Asuncion, 'UCI machine learning repository', 2010.

[8] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," in Proceedings of the 2000 ACM SIGMOD international conference on Management of data, ser. SIGMOD '00. New York, NY, USA: ACM, 2000, pp. 427–438.

[9] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in Proceedings of the 2000 ACM SIGMOD international conference on Management of data, ser. SIGMOD '00. New York, NY, USA: ACM, 2000, pp. 93–104.

[10] Z. He, X. Xu, and S. Deng, "Discovering cluster based local outliers," *Pattern Recognition Letters*, vol. 2003, pp. 9–10, 2003.

[11] M. Elahi, K. Li, W. Nisar, X. Lv, and H. Wang, "Efficient clustering-based outlier detection algorithm for dynamic data stream," in Proceedings of the 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery – Volume 05, ser. FSKD '08. Washington, DC, USA: IEEE Computer Society, 2008, pp. 298–304.

[12] Sharma, M. Toshniwal, D," Pre clustering algorithm for anomaly detection and clustering that uses variable size buckets", Published in Recent Advances in Information Technology (RAIT), 1st International Conference on 15 - 17 March 2012.

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

[14] Carlos Ordonez, "Clustering Binary Data Streams with K-means", proceedings of ACM international conference on management of data, sigmod 1998.

[15] Madjid Khalilian, Norwati Mustapha , " Data Stream clustering-Challenges and issues", Proceedings of the International Multi Conference of Engineers and Computer Scientists, Hong Kong ,Vol I, pp.17 - 19,March 2010

[16] Sudipto Guha, Rajeev Rastogi, Kyuseok Shim, "CURE: an efficient clustering algorithm for large databases" ACM LIBRARY, 1999.

[17] Luis Torgo, Carlos soares, "Resource-bounded Outlier Detection using Clustering Methods", proceedings of the conference on data mining for business applications, 2010.
[18] Hendrik Fichtenberger, Marc Gillé, Melanie Schmidt, in Algorithms –ESA 2013 , Volume 8125, 2013, pp 481-492.
[19] Irad Ben-Gal, "Outlier Detection", Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers, Kluwer, Academic Publishers, 2005.