International Journal of Advanced Computer Science and Information Technology (IJACSIT)
Vol. 2, No. 3, 2013, Page: 25-34, ISSN: 2296-1739

www.elvedit.com

Outlier Detection in Stream Data by Clustering Method

Authors

Hossein Moradi Koupaie *moradyhsnm@yahoo.com*
Advanced Informatics School, Universiti Teknologi Malaysia *Kuala Lumpur, Malaysia*

Suhaimi Ibrahim *suhaimiibrahim@utm.my*
Advanced Informatics School, Universiti Teknologi Malaysia *Kuala Lumpur, Malaysia*

Javad Hosseinkhani *jhkhani@gmail.com*
Department of Computer Engineering/ Islamic Azad University, Kerman Branch *Kerman, Iran*

Abstract

The fundamental and active research problem in a lot of fields is outlier detection. It is involved many applications. A lot of these methods based on distance measure. But for stream data these methods are not efficient. Most of the previous work on outlier detection declares online outlier and these have less accuracy and it may be lead to a wrong decision. moreover the exiting work on outlier detection in data stream declare a point as an outlier/inlier as soon as it arrive due to limited memory resources as compared to the huge data stream, to declare an outlier as it arrive often can lead us to a wrong decision, because of dynamic nature of the incoming data. The aim of this study is to present an algorithm to detect outlier in stream data by clustering method that concentrate to find real outlier in period of time. It is considered some outlier that has received in previous time and find out real outlier in stream data. The accuracy of this method is more than other methods.

Key Words

Outlier Detection, Stream Data, Clustering Method, Efficient Algorithm.

I. INTRODUCTION

Data mining, in general, deals with the discovery of non-trivial, hidden and interesting knowledge from different types of data. With the development of information technologies, the number of databases, as well as their dimension and complexity grow rapidly. It is necessary what we need automated analysis of great amount of information. A data stream is an ordered sequence of objects X1,..,Xn. The main difference between a traditional database and a data stream management system (DSMS) is that instead of relations, we have unbounded data streams. Applications, such as fraud detection, network flow monitoring, telecommunications, data management, etc., where the data arrival is continuous and it is either unnecessary or impractical to store all incoming objects.

Anomaly detection deals with detecting data elements from a data set which is different from all the other data elements in a set. Anomalies can arise due to different reasons such as mechanical faults, other changes in the system, fraudulent behavior, instrument error, human error or natural deviation. Usually anomalous observations are more interesting and need excess examination. It is not easy to define exactly what is an anomaly or an outlier. Traditional methods for outlier detection can produce good results on stored static dataset. Traditional data mining methods cannot be applied to streaming data efficiently as these methods are suitable for the environment where the entire dataset is already available and algorithm can operate in more than one pass. A general framework for mining data streams need small constant time per record along with the minimum memory requirement, using at most one scan of data. As the nature of data stream is unbounded the problem of mining outlier in data streams is often performed based on certain times intervals, usually called windows [6, 7 and 17].

It is very difficult to collect labeled data for data mining and also new concept may come to existent and others may get outdated in streaming data. In such a scenario unsupervised data mining approaches are more feasible as compare to supervised approaches. Unsupervised outlier detection approaches do not require class labels of objects and can detect unforeseen outlying cases. Clustering based outlier mining methods are unsupervised in nature; do not require knowledge of data in advance. Density based clustering methods can produce outlying objects along with normal cluster. Partitioning based clustering methods can be used for distance based outlier detection [20].

Recently, some data mining methods [21, 22] for a data stream have been actively introduced. In the traditional database environment there is a lot of research towards outlier/ anomaly detection. Among those Distance based approach first proposed by Knorr and Ng [23] and later this approach is further extended in different scenarios. All these techniques are highly dependent on the parameters provided by the users which is difficult in increasing dimensionality, and the wrong choice can effect the results. Later LOF [24] and its extension [25] although useful one, the computation of LOF values for every data objects require a large number of nearest neighbor searches.

The existing methods are either nearest neighbor based, which calculate the pair wise distances in a specified neighborhood, which doesn't suit in the environment of data stream, or model based, which are specific for intrusion detection in network based data flow. Where a profile or model is developed first and then incoming data is compared with the model to detect abnormal data. Few researchers applied same techniques and methods for streaming data, but most of the methods are computationally expensive if applied to unbounded data streams [19].

Clustering-based methods assume that the normal data objects belong to large and dense clusters, whereas outliers belong to small or sparse clusters or do not belong to any clusters. Clustering-based methods detect outliers by examining the relationship between objects and clusters. Intuitively, an outlier is an object that belongs to a small and remote cluster, or does not belong to any cluster. This leads to three general methods to clustering-based outlier detection, consider an object [8, 18].

- Does the object belong to any cluster? If not, then it is identified as an outlier.
- Is there a large distance between the object and the cluster to which it is closest? If yes, it is an outlier.
- Is the object part of a small or sparse cluster? If yes, then all the objects in that cluster are outliers.

II. RELATED WORK

There is a lot of literature on the anomaly detection problem which describes a variety of approaches like, Distance- Based outlier Detection is proposed by Knorr and Ng [23]. Given parameters k and R, an object is a distance-based outlier if less than k objects in the input data set lie within distance R from it. Further extended by Ramaswamy et al in [27] having idea, in order to rank the outliers, introduced the following definition: given k and n, an object o is an outlier if no more than n-1 other objects in the dataset have higher value for Dk than o, where Dk(o) denotes the distance of the k^{th} nearest neighbor of o. This concept is future extended in [28, 29, 30], where each data point is ranked by the sum of distances from its k nearest neighbors. Distance based outliers are not suitable if the clusters have different densities so to overcome the shortcoming of distance based outliers Breunig et al proposed a concept of LOF [24] which are the objects outlying relative to their local neighborhoods, with respect to densities of the neighborhood.

This concept is useful but to compute the LOF value large number of k-nearest neighbor searches makes it computationally expensive. On the bases of observing density distribution from the data, Aggarwal and YU [31] proposed a technique for outlier detection. The basic idea in their definition is, a point is an outlier, if in some lower dimensional projection it is present in a local region of abnormally low density. This method is also an efficient method for high dimensional data set. Some Clustering-Based outlier detection techniques are proposed [32, 33]. This technique work in two basic steps, fixed width clustering with w(radius) and after the first phase

the next step is sorting of clusters produced in the first step. Points in the smaller clusters are declared as outliers. Clustering-Based techniques are further extended by proposing the concept of cluster based local outlier, in which a measure for identifying the outlierness of each data object is defined. Deviation-based techniques identify outliers by inspecting the characteristics of objects and consider an object that deviates these features, declared as an outlier [33]. Distance-based methods previously discussed are designed to work in the batch framework, under the assumption that the whole data set is stored in secondary memory and multiple passes over the data can be accomplished.

Hence, they are not suitable for data streams. While the majority of the approaches to detect anomalies in data mining consider the batch framework, some researchers have attempted to address the problem of online outlier detection. The replicator neutral network (RNN) based technique is introduced to detect outliers by Harkins, et al. [34]. Although there is lot of research on outlier detection, but there is little research in the direction of outlier detection in dynamic data streams. This area still needs lot of attention because the existing methods are not appropriate in the stream environment.

Elahi, et al. [19] have presented a cluster-based outlier detection in data stream, which pay more attention to the points detected as outliers and give them a chance of survival in the next incoming chunk of data, rather declare them outlier by observing the current chunk. It only keep the most suitable candidate outliers. As in data stream it can't keep the entire stream in some physical memory so it discard the region which is safe and do not contain outliers, and free memory for the next generation of data to be processed effectively. The experimental results show that the approach outperforms some existing methods on identifying meaningful outliers over DataStream.

Elahi, et al. has been suggested Efficient Clustering-Based Outlier Detection Algorithm for Dynamic Data Stream. This algorithm depends on the clustering based approach that splits the stream to clusters and chunks. In the stable number of clusters, each chunk using k-mean. It also retains the applicant outliers and means value of every cluster for the next fixed number of steam chunks as a replacement for keeping only the summary of information that is utilized in clustering data stream to assured that the discovered candidate outliers are real. It is better to decide outlines for data stream objects by utilizing the mean values of the current chunk of stream with the mean value of the clusters of previous chunk [1, 15].

Ren, et al. (2009) suggested another approach, which is based on "Clustering-Based Approaches" that is Efficient Outlier Detection Algorithm for Heterogeneous Data Streams, which divide the stream in chunks. After that, each chunk is clustered and the equivalent clusters are kept in same cluster situations. The amount of adjacent cluster situations and the illustration degree are calculated to create the final outlier references such as potential outliers. The researchers believe that experimental findings release that its model has better scalability and higher investigation accuracy [3, 14].

Furthermore, the aim of this study is to recommend a cluster based partitioning algorithm that can split the stream in candidate and safe region. In order to accomplish precise results for finding most outstanding outliers, LOF (Local outlier Factor) algorithm was utilized through all these dividers distinctly by some insignificant improvement of LOF evaluation throughout candidate region in second phase. Many researches on different dataset approve that in comparison to the other developed LOF, this technique is able to track improved outliers with low computational cost than the direct LOF [2,10].

On the other hand, there is a few researches have been done on the introduction an algorithm. Through the first stage, this algorithm split the stream to chunks and creates initial clusters in each chunk. K-nearest neighbor approach for outlier detection is used for each cluster in the second stage. For each element of data, this algorithm decreases the nearest neighbor searches. It detects outliers by the development of data stream in the best way. According to the authors, techniques can discover appropriate outliers with low computational cost than the other exiting [1, 11 and 13].

III. PROPOSED WORK

In high-dimensional data streams, there are new techniques which are called Stream Projected Ouliter deTector (SPOT) that are managing outlier detection problem and it is somehow unique such as use a new window-based time model and degenerations cell summaries to get data from the data stream. Sparse Subspace Template (SST) also makes it unique. It is a set of top that bare gotten subspaces by supervised and unsupervised learning procedures, and it is made in SPOT to detect expected outliers well. In unsupervised learning, a real search method for discovering outlying subspaces from training data is Multi-Objective Genetic Algorithm (MOGA).

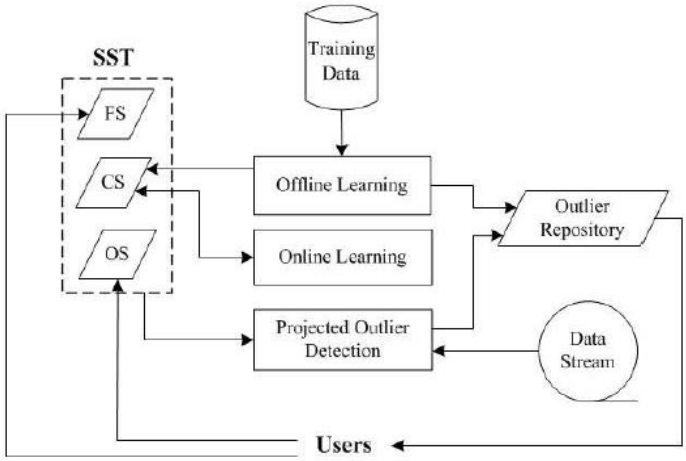

FIGURE 1: AN OVERVIEW OF SPOT [4]

Lastly, SST can perform online self-evolution to handle dynamics of data streams that offers some parts on the technical and motivational challenges of detecting outliers; it can be from high-dimensional data streams. Moreover, it is providing the maps for system demonstration of SPOT. An Overview of SPOT has shown in Figure 1 [4].

The k-means algorithm is the best known partitioned clustering algorithm. It is perhaps also the most widely used among all clustering algorithms due to its simplicity and efficiency. Given a set of data points and the required number of k clusters (k is specified by the user), this algorithm iteratively partitions the data into k clusters based on a distance function [5].

K-mean Algorithm

Algorithm *k*-means (*k*, *D*)
1 choose *k* data points as the initial cancroids (cluster centres)
2 **repeat**
3 **for** each data point **x** ∈ *D* do
4 compute the distance from **x** to each centred;
5 assign **x** to the closest centred // a centred represents a cluster
6 **end for**
7 re-compute the centred using the current cluster memberships
8 **until** the stopping criterion is met

FIGURE 3: THE K-MEANS ALGORITHM

The algorithm that is proposed uses incremental k-means clustering because of stream data. At first stream data in a window with a specify size entered. After that it is continued with a parallel process. One way (process1) is to find online outlier per time and other way (process2) to find real outlier (offline outlier) with more accuracy. In procces1, it is clustered data in window with k-means algorithm and some cluster that is small or far from of other cluster introduce as an online outlier. And then this process repeat for next window. In process2, at first previous outlier from n previous window add to current window. Next data from current window is clustered by k-mean algorithm. And some cluster that is small or far from other clusters introduce as a real outlier or offline outlier. At last this process repeat for next window. The important parameter in this process is n. n should determine correct. It should be considered the type of data. For example if stream data is about home electronic, it is considered the time that is entered data. Or if stream data is about verifying credit card of in bank, it should considered the time and who the owner of card is and so on.

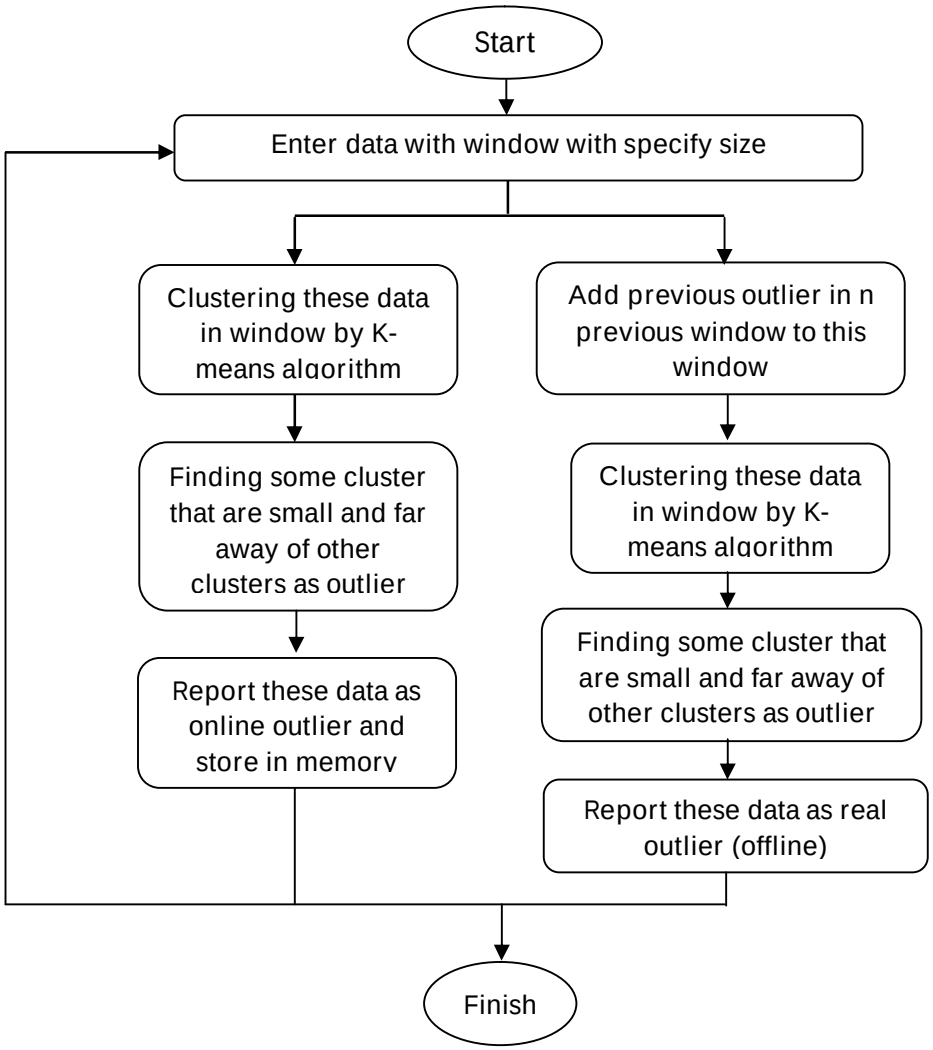

FIGURE 4: ALGORITHM TO DETECTION OUTLIER IN STREAM DATA WITH CLUSTERING

IV. CONCLUSION

It is very difficult to collect labeled data for data mining and also new concept may come to existent and others may get outdated in streaming data. In such a scenario unsupervised data mining approaches are more feasible as compare to supervised approaches. Unsupervised outlier detection approaches do not require class labels of objects and can detect unforeseen outlying cases. Clustering based outlier mining methods are unsupervised in nature; do not require knowledge of data in advance. Density based clustering methods can produce outlying objects along with normal cluster.

In this paper, we present a cluster-based outlier detection in data stream, where we prefer an incremental clustering algorithm to detect real outlier in stream data. It is used of k-mean

algorithm but with a high accuracy to find outlier. In the further work, It could be considered about the size of window in this algorithm and determine some good parameter to find a suitable n(the number of previous window).

REFERENCES

[1] Elahi, M. L., Xinjie ; Nisar, Wasif ; Khan, Imran Ali2 ; Qiao, Ying ; Wang, Hongan (2008). DB-Outlier detection algorithm using divide and conquer approach over dynamic DataStream. International Conference on Computer Science and Software Engineering, CSSE 2008. Wuhan, Hubei, China, IEEE Computer Society, 445 Hoes Lane - P.O.Box 1331, Piscataway, NJ 08855-1331, United States. 4: 438-443.

[2] Elahi, M. K., Li ;Nisar, Wasif2 ; Xinjie, Lv1 ; Hongan, Wang (2009). Detection of local outlier over dynamic data streams using efficient partitioning method. 2009 WRI World Congress on Computer Science and Information Engineering, CSIE 2009. Los Angeles, CA, United states, IEEE Computer Society, 445 Hoes Lane - P.O.Box 1331, Piscataway, NJ 08855-1331, United States. 4: 78-81.

[3] Ren, J. W., Qunhui ; Zhang, Jia ; Hu, Changzhen (2009). Efficient outlier detection algorithm for heterogeneous data streams. 6th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2009.Tianjin, China, IEEE Computer Society, 445 Hoes Lane - P.O.Box 1331, Piscataway, NJ 08855-1331, United States. 5: 259-264

[4] Zhang, J. G., Qigang ; Wang, Hai (2008). SPOT: A system for detecting projected outliers from high-dimensional data streams. 2008 IEEE 24th International Conference on Data Engineering, ICDE'08. Cancun, Mexico, Inst. of Elec. and Elec. Eng. Computer Society, 445 Hoes Lane - P.O.Box 1331, Piscataway, NJ 08855-1331, United States: 1628-1631.

[5] 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.

[6] Babcock, B., Babu, S. , Datar, M. , Motwani, R., & Widom,J. (2002). Models and issues in data stream systems. In *Proc. 2002 ACM Symp. Principles of Database Systems (PODS'02)*, pages 1–16, Madison, WI, June 2002.

[7] Babu, S. and Widom,J., (2001). Continuous queries over data streams.*SIGMOD Record*, 30:109–120.

[8] Charu C. Aggarwal, Philip S. Yu, (2001). Outlier detection for high dimensional data, Proc. of the 2001 ACM SIGMOD int. conf. on Management of data, p.37-46, May 21-24, 2001, Santa Barbara, California, United States.

[9] Babcock, B., Babu, S., Datar, M., Motwani, R. & Widom, J. (2002). Models and issues in data stream systems. In *Proc. 2002 ACM Symp. Principles of Database Systems (PODS'02)*, pages 1–16, Madison, WI, June 2002.

[10] Muthukrishnan, S. (2003). Data streams: algorithms and applications. In *Proc. 2003 Annual ACM-SIAM Symp. Discrete Algorithms (SODA'03)*, pages 413–413, Baltimore, MD,Jan. 2003.

[11] Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. Artificial intelligence, 97(1), 273-324.

[12] R.S. Kulkarni, G. Lugosi, V.S. (1998). Santosh, Learning pattern classification—a survey, IEEE Transaction on Information Theory 44 (6), 2178–2206.

[13] R.S. Kulkarni, M. Vidyasagar. (1997) . Learning decision rules for pattern classification under a family of probability measures, IEEE Transactions on Information Theory 43 (1) 154–166.

[14] Liang, Z., Lia, Y. (2009). Incremental support vector machine learning in the primal and applications. Neurocomputing 72(10-12), 2249–2258.

[15] Zheng, J., Yu, H., Shen, F., Zhao, J. (2010). An Online Incremental Learning Support Vector Machine for Large-scale Data. In: Diamantaras, K., Duch, W., Iliadis, L.S. (eds.) ICANN 2010. LNCS, vol. 6353, pp. 76–81. Springer, Heidelberg.

[16] Liu, X., Zhang, G., Zhan, Y., Zhu, E. (2008). An Incremental Feature Learning Algorithm Based on Least Square Support Vector Machine. In: Preparata, F.P., Wu, X., Yin, J. (eds.) FAW 2008. LNCS, vol. 5059, pp. 330–338. Springer, Heidelberg.

[17] Vapnik, V. (1999). The nature of statistical learning theory. Springer, New York.

[18] Ruping, S. (2002). Incremental learning with support vector machines. Technical Report TR-18, Universitat Dortmund, SFB475.

[19] Elahi, Manzoor, et al. (2008). "Efficient clustering-based outlier detection algorithm for dynamic data stream." *Fuzzy Systems and Knowledge Discovery, 2008. FSKD'08. Fifth International Conference on*. Vol. 5. IEEE.

[20] Thakran, Yogita, and Durga Toshniwal. (2012). "Unsupervised outlier detection in streaming data using weighted clustering." *Intelligent Systems Design and Applications (ISDA), 2012 12th International Conference on*. IEEE.

[21] Angiulli, F. and Fassetti, F. (2007). Detecting distance-based outliers in streams of data. In Proc. of the Sixteenth
ACM Conf. on information and Knowledge Management (Lisbon, Portugal, November 2007). CIKM '07.

[22] Pokrajac, D. Lazarevic, A.Latecki, L.J. (2007). Incremental Local Outlier Detection for Data Streams Computational Intelligence and Data Mining 07. CIDM.

[23] Knorr, E. M., Ng, R. T. (1998). Algorithms for Mining Distance-Based Outliers in Large Datasets, Proc. 24th VLDB.

[24] M.M. Breunig, H.P.Kriegel, R.T. Ng and J.Sander. (2000). LOF: Identifying Density-Based Local Outliers ACM SIGMOD.

[25] Wen Jin and Anthony K. H. Tung and Jiawei Han. (2001). Mining top-n local outliers in large databases. Pages 293-298.

[26] Ramaswamy S., Rastogi R., Kyuseok S. (2000). Efficient Algorithms for Mining Outliers from Large Data Sets, Proc. ACM SIDMOD Int. Conf. on Management of Data.

[27] Fabrizio Angiulli & Clara Pizzuti. (2002). Fast Outlier Detection in High Dimensional Spaces, Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery, p.15-26, August 19-23.

[28] F. Angiulli, S. Basta, and C. Pizzuti. (2006). Distance-based detection and prediction of outliers. IEEE Transaction on Knowledge and Data Engineering, 18(2):145(160, February 2006).

[29] M. F. Jiang, S. S. Tseng, C. M. Su. (2001). Two-phase clustering process for outliers detection. Pattern Recognition Letters, 22(6/7): 691-700.

[30] Charu C. Aggarwal, Philip S. Yu.(2001). Outlier detection for high dimensional data, Proc. of the 2001 ACM SIGMOD int. conf. on Management of data, p.37-46, May 21-24 , Santa Barbara, California, United States.

[31] E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo. (2002) A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data. In Data Mining for Security Applications.

[32] A. Arning, R. Agrawal, P. Raghavan. (1996). A linear method for deviation detection in large databases. In: Proc of KDD'96, 164 169.

[33] S. Harkins, H. He, G. J.Willams, R. A. Baster. (2002). Outlier detection using replicator neural networks. In: Proc of DaWaK'02, 170-180.