Review: Outlier Detection using Clustering Based Techniques in Data Stream

Prashant V. Chauhan[1], Vijay K. Vyas[2], Darshan P. Upadhyay[3]

[1,2,3]*Assistant Professor, Department of Information Technology, VVP Engineering College, Rajkot, India*

---***---

Abstract - Outlier detection is a critical data mining task used to detect and, where necessary, delete abnormal occurrences from dataset or tries to uncover useful anomalous and uneven patterns contained in large datasets. Various approaches are used to assign data points in different clusters based on different parameters. In this paper, we present a review of various existing approaches based on clustering of data points for detecting outliers from data set. The survey will cover the traditional outlier detection methods for static datasets, low dimensional datasets and recent developments that deal with outlier detection problems for dynamic data stream and high-dimensional datasets. In a comparative analysis, we highlight their benefits and drawbacks.

Key Words: Outlier detection, Clustering based outlier detection, D-Stream, CORM, Hy-CARCE

1. INTRODUCTION

Due to its frequent use in a variety of applications, outlier detection is still a crucial and vast study area in data mining. Researchers can gain crucial information that aids in better data judgements by recognizing outliers [1,2].

A wide number of real-world applications in business, engineering, security, and other fields today depend on outlier detection, sometimes referred to as anomaly detection in certain literature. For credit card firms, for instance, outlier detection can assist in identifying suspicious fraudulent transactions. It can be used to detect irregular brain signals that could point to the earliest stages of brain cancer [10].

In the past few years, a lot of study has been done on outlier detection because of its intrinsic usefulness in a variety of fields. There are several outlier identification methods that have been put out that make use of various processes and algorithms. The main contemporary outlier detection techniques are covered in this study. We will discuss the key types of outlier identification techniques and assess each of their benefits and drawbacks rigorously [2].

2. FUNDAMENTAL CONCEPTS

This section depicts some basic terms of the outlier detection in data streams.

Outlier: In any dataset there are many data points, when we arrange those data points based on some characteristics in different clusters, some data points cannot fall in any cluster as they (having abnormal behavior / are different in normal behavior from other data points / distinctly different from other data points. That type of data points is known as outliers, and data points which falls within clusters are known as inliers. Inlier data points have normal expected behavior. Malicious behaviour, instrumental error, setup error, changes in the environment, human error, and catastrophe are all possible causes of outliers in a dataset. Data point of any dataset are distributed as shown in Figure 1. After distribution of data points three clusters are formed named C1, C2, C3 and four outlier data points are formed namely O1, O2, O3 and O4 [2,3].

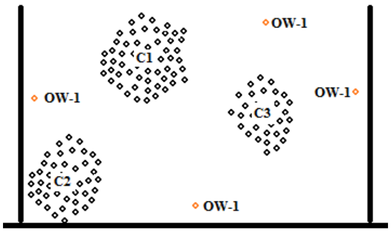

Fig -1: Clusters and outliers after data point distribution.

Data Stream Mining: Data stream mining is the process of looking at the stream of data and identifying valuable patterns to make decisions. Data streams can be described as massive and continuous, unbounded or ordered sequences of information which are arriving at a rapid pace and with a changing distribution. Some example of data stream are sensor data, web browsers activity logs, and logs of financial transaction systems [2].

Outlier Detection in Data Stream Mining: Outlier detection is the process of detecting data points which shows abnormal behaviour in a dataset. Collection of information of outliers from streaming data has become very important requirement in many applications that generate stream data. Outlier detection is required in data stream applications such as the detection of fraud, dieses detection of patient, detection of weather changes, detection of an abnormality within a computer network, and so on.

3. EXISTING APPROACHES FOR OUTLIER DETECTION

To detect outliers from stream data, many techniques are developed based on different approaches. In this section, we discuss outlier detection techniques that are used for outlier detection from data stream. These techniques are classified into below categories [1,9].

Distance based outlier detection:

Distance based approach determines the outlierness of a data point by calculating and analyzing the distances between its neighbor data points [9]. The most widely utilized definition of an outlier detection using distance is based on the idea of the local neighborhood, the k-nearest-neighbor of data points and the conventional distance threshold value. Distance between neighboring data points can calculated using manhattan distance, Euclidean distance metrics. For data stream application with categorical data, data normalization is performed to normalize various scales of data feature and after that outlier detection is carried out [11].

Density based outlier detection:

Density based approach uses more complex mechanism than distance based outlier detection approach. In Density based outlier detection, outliers are detected/identified by calculates density around the data points. Data point which is having density much lower than to their neighbor are identified as outliers, whereas data points having density near to their neighbors is called as inliner [12]. Means outliers are appeared in region of data points where density is low, and inliner are appeared in region of data points where density is high.

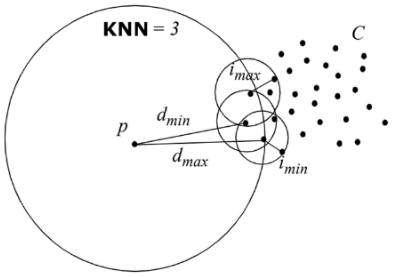

Fig -2: Nearest neighbors with LOF reachability distance.

To evaluate density of any data point Local Outlier Factor (LOF) is used. This LOF score of a data point describes ratio of local density of that data points and local density of their nearest neighbor data point. If this ratio is higher than this data point is declared as outlier [9].

Clustering based outlier detection:

Primary aim of clustering based approach is to generate clusters of data points based on their behaviour. Inlier Data points which are having similar behaviour are grouped into cluster. Outlier data points can't lie into any cluster as they are far away from the center point of nearest cluster or they belong to small or sparse cluster [9].

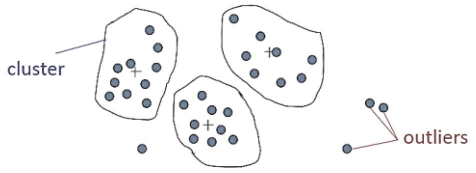

Fig -3: Outlier detection using clustering techniques.

Clustering based algorithms inherently define outliers as background noise in clusters. As shown in below figure there are three clusters and four outlier data points. Any of these four outlier data points does not lies in any cluster nor they are near to any clusters, thus they are detected as outliers.

Statistical based outlier detection:

Data distribution model or probability model are used in statistical based outlier detection approaches.

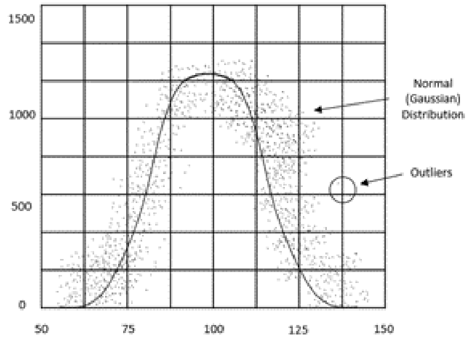

Fig -4: Outlier detection using statistical techniques.

After modelling data points using distribution or probability model, some data points are declared as outlier as they do not align with or are not in line with the distribution model. Statistical based outlier detection techniques are divided into two categories i.e. parametric methods and non-parametric methods. In parametric methods, it assumes distribution of given dataset in advance, and then it estimates attributes of the distribution model from given dataset. Non parametric methods does not make any prior assumption for dataset distribution model.

Frequent pattern mining based outlier detection:

Frequent pattern mining based outlier detection approaches uses unsupervised method, in which they analysis the data patterns and model data points with normal behaviour. As many data points possesses the common features of the dataset they unlikely to be outliers. Based on minimum threshold parameter, outliers are the data points which have value less than minimum threshold or they have small number of frequent patterns [9].

Classification based outlier detection:

Classification based approaches classifies data points into different classes thus provide more precise analysis of dataset. This approach uses training data for learning classification model and then classify newer data points into the learned classes. Classification based outlier detection approach has a common assumption that a classifier has learnt and based on that learning, that can differentiate between normal and outlier classes [9].

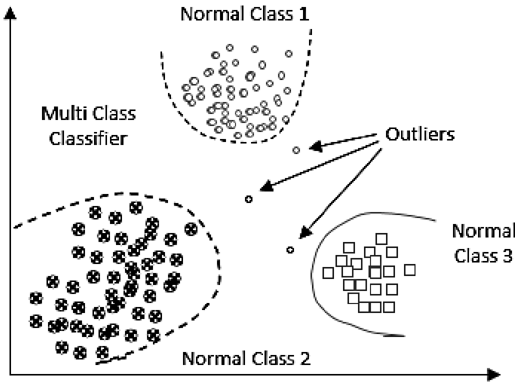

Fig -5: Outlier detection using Classification based techniques.

Angle based outlier detection:

Outliers are detected by considering angle variations between data points. Angles variations between each target instance and the remaining data points are calculated and based on this identify outliers in angel based outlier detection.

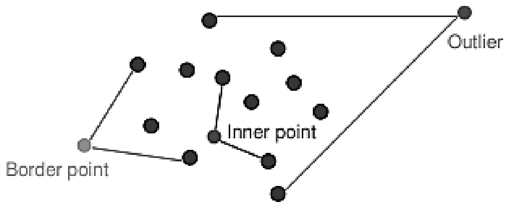

Fig -6: Angles variations of datapoints. Outliers have smaller angle variances whereas inliers have larger angle variances.

Angle variance of an outlier is lower than that of the normal points. This is useful to reduce effect of "curse of dimensionality" on mining high-dimensional data [9].

4. CLUSTERING BASED OUTLIER DETECTION

To distribute data items into multiple clusters which have similar characteristics based on some parameters clustering techniques are used. Some data items fall into the clusters and some data items are not fall into cluster or far away from the center of nearest clusters. To identify outliers using clustering techniques, it works in two steps first it grouped data points with similar characteristics into similar clusters and then differentiate outlier data points based on deviations of clustering result. Except clustering by cluster membership, there is two other methods based on which cluster is formed are distance to cluster center and second is cluster cardinality.

D-Stream [4]

A grid based approach D-Stream is introduced by Chen et al. for clustering data points and detect outliers. This algorithm forms grids of data points and then form clusters. If the grid is sporadic grid, then it is labelled as outlier. To from clusters grid are categorized in three categories as dense grid, sparse grid and transitional grid according to grid density. Clusters are generated by connecting dense grid and sparse grid nearer to that dense grid. Each data points have associated with a density value, and sum of density value of all these data points residing into the grid is considered as greed density. A threshold value is used to identify the grids with low density, supposing that outliers are mapped to grids with few data points.

D-Stream works in two phases. First phase is online phase, in this it reads data points and then categorized data points into an associated density grid, and update characteristic coefficient of that grid. After first phase, second phase is offline phase in which adjust clusters dynamically at every time stamp. Then cluster generated initially are updated periodically after removal of sporadic grids.

Cluster based OutlieR Miner(CORM) [5]

CORM is working based on k-means clustering algorithm and is working on sliding window model for process data chunks of data stream. Two phase procedure is used to clustering data points by maintaining actual cluster center for the current sliding window data points and updated cluster centers defined based on previous cluster center and current cluster center. Outliers are detected based on the distance between data points and updated cluster centers.

Elahi et al. [5] also added that to maintain quality of outlier detection, this algorithm use threshold parameter which is given by user. In each step outlier score for potential outlier points are calculated and based on that outlier are declared. Main advantage of this approach is less memory consumption.

Working of CORM: First data chunks are input one by one, then divide this data set into k different clusters. Each cluster represented by their center point. If no clusters are formed, then appropriate clusters center are chosen and all data chunks are assigned to their nearer cluster otherwise data chunks are assigned to formed nearer cluster. After inserting data points into different clusters their mean value is calculated based on new cluster mean and previous cluster mean found during processing of L number of data chunks. Where L is user defined parameter used to confirm any data point as outlier.

Outliers generated during previous L number of data chunks are also clustered with current data chunks. Final Outliers are declared at stage L If its value passed to threshold value given by user.

AnyOut [6]

AnyOut uses a tree structure created as a result of hierarchical clustering in order to represent incoming data in new window and determine outlier scores for that data points. AnyOut uses a tree structure called ClusTree [13]. It was originally designed for the parameter-free hierarchical clustering of the data streams.

ClusTree's tree nodes compactly represent a cluster using cluster features. This is a tuple that contains the number of data points in the cluster as well as the linear sum and squared sum. ClusTree can update the cluster features whenever new data points are added to the window. ClusTree has buffer entries that permit for the anytime insertion and updating of cluster features. AnyOut emphasizes its real-time characteristic. It outputs a more precise score if it is given more time. This is done using a top-down outlier evaluation strategy. The data point searches for the closest cluster at each level. The result for current data point must be returned when the next data point received for processing, the outlier score is calculated based on the relationship between the most recent cluster found in the tree.

There are two ways to calculate an outlier score. The first, called the mean outlier score is the distance between the mean of entries in the cluster and the data point. Another is the Gaussian probability density is used to calculate the outlier score.

Hy-CARCE [7]

A weighted ensemble framework was developed to detect outliers within data streams. It uses a clustering algorithm that models the normal patterns in data instances from previous windows. The authors address the situation where data streams swings between different states. Each state could potentially contain multiple data point distributions. The proposed framework consists of three components. The Hy-CARCE clustering algorithm is used to first cluster all of the data points inside the current window.

A hyper ellipsoidal clustering algorithm without predetermined cluster numbers is called HyCARCE. As the created "clustering model," HyCARCE generates a set of cluster boundaries. Each window is clustered, and the memory-based clustering models are used to calculate the ensemble score for the incoming data points. In next step ensemble weight is obtained from the similarity between two clustering models. Focal distance between two hyperellipsoids is used for calculate ensemble weight. To be more precise, the distance between every pair of hyper ellipsoid borders, each from a distinct clustering model, is first estimated for two clustering models. Next, beginning from the boundary pair with the smallest distance, pairs of boundaries are chosen out greedily. The similarity between two clustering models is ultimately calculated as the reciprocal of the sum of the distances between the generated pairings.

In third step ensemble model is used in the framework to determine an outlier score for a data point of new window based on its association with earlier clustering models. For each clustering model in the history, the algorithm specifically determines if a data point is a member of any cluster by determining whether the Mahalanobis distance between the data point and the cluster hyperellipsoid exceeds a predetermined threshold. For each prior clustering model, the check generates a binary score. The weighted total of those binary values is the final outlier score, is determined by the similarity between the current clustering model and the matching previous clustering model.

HyCARCE with Gaussian Cluster [8]

A different method for detecting outliers in data streams based on the HyCARCE clustering algorithm. They use Gaussian clusters to characterize typical data patterns, which differs from the previous described strategy in [8], and they generate the outlier score based on the Gaussian probability density function of a cluster. Moreover, the suggested strategy takes into account and manages recently formed clusters.

The first step of the suggested method is to analyses the data points in a new window and see whether any Gaussian clusters already in existence may explain the underlying distribution of some of the data points. In order to accomplish this, they developed two criteria. In first criteria the number of data points in the new window fitting the cluster must not be too small, which is tested by the Cumulative Binomial Probability (CBP) function, and in second criteria the data points must spread out the cluster, which is tested by converting the data points into conventional Gaussian distributions, followed by a spherical coordinate system.

After deleting the data points that can be explained by the models in use in the first step, the second stage involves applying CBP to identify probable developing Gaussian clusters. HyCARCE clustering is used to cluster these data points and save the new model if the outcome is favorable. The largest value among the probability of a data point being seen under each of the Gaussian clusters is the score of a data point.

5. ADVANTAGES AND DIS-ADVANTAGES OF CLUSTERING BASED OUTLIER DETECTION

Advantages: Because clustering data stream are unsupervised, they are a good option and highly helpful for outlier detection in data streams. Further new points may be added, and after learning from the clusters, they can be checked for outliers. They can adjust to an incremental mode as a result. Also, the data distribution is more suited for incremental mode because it doesn't require prior information.

Disadvantages: Outliers in clustering contexts are binary, meaning there is no quantitative way to know if an item is an outlier or not. They are also renowned for not being able to go back; as a result, they are unable to reverse previous actions.

The majority of clustering techniques rely on and depend on the users to predetermine the number of clusters, which is a challenging undertaking. Arbitrary form clusters in clustering algorithms might also make it difficult to identify the precise clusters of the data. As the form of the clusters must be determined beforehand, the majority of known clustering methods need multiple. To assume several clusters in advance in a data stream situation is quite difficult [14].

6. CONCLUSIONS

In data stream mining, outlier detection is one of the more difficult tasks. As a result, several strategies have been put out to address this issue. It might be difficult to decide which algorithm would work best in a certain situation. In practice, it depends on several factors related to the type of the input and the desired outcomes. In this work, we have described many methods for finding outliers in data streams based on various established methodologies. Although not all of the methodologies for outlier detection in data streams are included in this paper, many of them are. So, further work may be done by including different strategies and addressing additional issues.

REFERENCES

1. Shiblee Sadik and Le Gruenwald, “Research Issues in Outlier Detection for Data Streams,” SIGKDD Explorations Volume 15, Issue 1, pp. 33-40, 2012.
2. P. Chauhan and M. 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, pp. 580-585, 2015.
3. M. Shukla, Y. P. Kosta and P. Chauhan, "Analysis and evaluation of outlier detection algorithms in data streams," 2015 International Conference on Computer, Communication and Control (IC4), pp. 1-8, 2015.
4. Yixin Chen and Li Tu. 2007. Density-based clustering for real-time stream data. In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 133–142.
5. Manzoor Elahi et al. 2008. Efficient clustering-based outlier detection algorithm for dynamic data stream. In Proceedings of the 5th IEEE International Conference on Fuzzy Systems and Knowledge Discovery. 298–304.
6. Ira Assent, Philipp Kranen, Corinna Baldauf, and Thomas Seidl. 2012. Anyout: Anytime outlier detection on streaming data. In Proceedings of the International Conference on Database Systems for Advanced Applications. Springer, 228–242.
7. Mahsa Salehi, Christopher A. Leckie, Masud Moshtaghi, and Tharshan Vaithianathan. 2014. A relevance weighted ensemble model for anomaly detection in switching data streams. In Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining 2014. 461–473.
8. Milad Chenaghlou et al. 2017. An efficient method for anomaly detection in non-stationary data streams. In Proceedings of the IEEE Global Communications Conference. 1–6.
9. R Souiden, I., Brahmi, Z., Toumi, H. (2017). A Survey on Outlier Detection in the Context of Stream Mining: Review of Existing Approaches and Recommadations. In: Madureira, A., Abraham, A., Gamboa, D., Novais, P. (eds) Intelligent Systems Design and Applications. ISDA 2016. Advances in Intelligent Systems and Computing, vol 557.
10. Progress in Outlier Detection H. Wang, M. J. Bah and M. Hammad, "Progress in Outlier Detection Techniques: A Survey," in IEEE Access, vol. 7, pp. 107964-108000, 2019.
11. Advancements of Outlier Detection a Survey) Ji Zhang. 2013. Advancements of outlier detection: A survey. ICST Trans. Scal. Inf. Syst. 13, 1 (2013), 1–26.
12. Tran L, Fan L, Shahabi C (2016) Distance-based outlier detection in data streams. PVLDB 9(12):1089–1100.
13. Philipp Kranen, Ira Assent, Corinna Baldauf, and Thomas Seidl. 2009. Self-adaptive anytime stream clustering. In Proceedings of the 2009 9th IEEE International Conference on Data Mining. IEEE, 249–258.
14. Azzedine Boukerche, Lining Zheng, and Omar Alfandi. 2020. Outlier Detection: Methods, Models, and Classification. ACM Comput. Surv. 53, 3, Article 55 (May 2021).