International Journal of Latest Engineering and Management Research (IJLEMR)
ISSN: 2455-4847
www.ijlemr.com || Volume 02 - Issue 08 || August 2017 || PP. 23-30

A Survey on Outlier Detection Techniques in Dynamic Data Stream

Ramesh Kumar B[1], Aljinu Khadar K V[2]
[1]*Assistant Professor, Department of Computer Science, SreeNarayana Guru College, Coimbatore, Tamil Nadu, India*
[2]*M.Phil Scholar, Department of Computer Science, SreeNarayana Guru College, Coimbatore, Tamil Nadu, India*

Abstract: Outlier detection has significant importance in the data mining domain. Applications which contain streaming data flow may have many abnormal or outlier data and these applications require efficient outlier detection techniques to detect and analyze these abnormal patterns. Outlier detection is the process of detecting patterns in the data which do not adhere to the normal behavior or data. These patterns are known by several terms such as anomalies, outliers, noise or inconsistent data. Detecting and analyzing the abnormal data like outliers is a wide research area with tremendous applications. Finding and selecting appropriate detection technique is mandatory. This survey presents the tools and techniques used for detecting outliers in data streams and attempts to classify the problem in outlier detection methods over the data stream. The review of detection techniques gives an insight into the further research opportunities in this domain.
Keywords: Outliers, Data Mining, Data Streams, AnomalyDetection, Clustering, Classification

I. Introduction

The development of information technology and network with huge set of databases resulted in the necessity of effective and active analysis of great amount of heterogeneous structured information. Data mining technique play an important role in collecting, processing and analyzing these types of complex and heterogeneous data. One prime step in obtaining a deep analysis is the detection of outliers. Outlier detection refers to the problem of finding patterns in data that do not fall under any particular category [1]. This involves the process of identifying data objects that do not relate to the remaining objects in the data set. These outlier patterns are often referred to as anomalies or outliers. Outlier detection methods are used in numerous application domains like finance data analysis, clinical trials, sensor data analysis, network intrusion, etc. Detecting outliers is a significant part of data analysis, which avoids misspecification, biased parameter estimation, and incorrect results. The process of detecting outliers should be done prior to analysis and modeling. Outlier detection is widely used in several research areas like statistics, data mining, sensor networks, environmental science, distributed systems, spatiotemporal mining, etc. It has been studied on a large variety of data types such as high-dimensional data, uncertain data, stream data, graph data, time series data, spatial data, and spatiotemporal data.

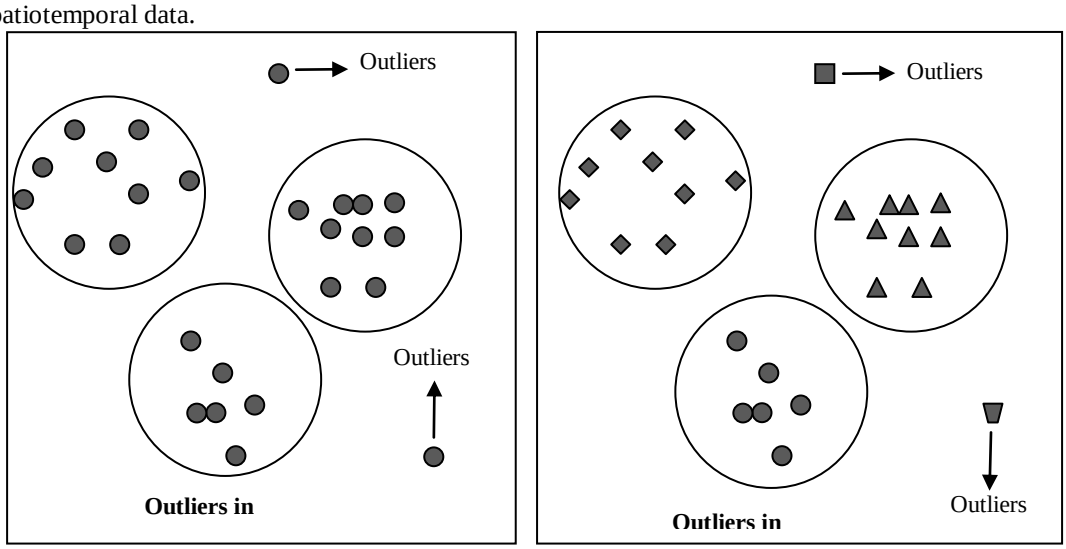

Fig 1.0 Outlier Detection in Clustering and Classification Processes.

The outlier is the part of both clustering and classification processes, which performs the detection of the abnormal object shown in Fig 1.0. The significance of outlier detection lies in the fact that outlier in data can be improved to useful information in an extensive array of the application domain. For example, an abnormality in the traffic pattern in a network indicates that the computer is hacked and is sending out sensitive data to an unauthorized destination. In public health data, the outlier detection techniques are widely used to detect abnormal patterns in patient medical records which may represent symptoms of a new disease. Wellbeing information and exception discovery systems are broadly used to identify strange examples in patient healing report which may speak to the manifestation of another disorder. Similarly, discordant observation in credit card transaction data could indicate misuse or theft of credit card.

1.1 Types of Outliers

The outliers can be categorized into three categories, namely, global outliers, contextual or conditional outliers and collective outliers

1.1.1 Global or point Outliers: In a given data set, a data object is a global outlier, if it deviates significantly from the rest of the data set. If an individual data instance is considered as anomalous with respect to the rest of data, then the instance is termed as a point or global outlier. Global outliers are sometimes called point anomalies and are the simplest type of outliers. Most outlier detection methods are aimed at finding global outliers. This is the simplest type of outlier and is the focus of major research on outlier detection. This type of outlier remains distinct from other data points by representing its outlier point. They are detected by analyzing point metrics which indicate the extent to which an individual data gets deviated from the other data in the data set.

1.1.2 Contextual outlier: another type of outlier is the contextual outlier, in which a data object deviates significantly with respect to a specific context of the object. Contextual outliers are also known as conditional outliers because they are conditional on the selected context. So, in contextual outlier detection, the context has to be specified as part of the problem definition. In general, in contextual outlier detection, the attributes of the data objects in question are divided into two groups such as contextual attributes and behavioral attributes. The contextual attributes of a data object define the object context. If we consider temperature data as an example, we may use date and location as contextual attributes and behavioral attributes may be the temperature, humidity, and pressure which define the object's characteristics, and are used to evaluate whether the object is an outlier in the context to which it belongs.

1.1.3 Collective Outliers: In collective outliers, a subset of data objects forms a collective outlier if the objects as a whole deviate significantly from the entire data set. Importantly, the individual data objects may not be outliers.

In the types of outliers, the global outliers are simplest than others. The data set can have multiple types of outliers. Moreover, an object may belong to more than one type of outlier. In business, different outliers may be used in various applications or for different purposes. Context outlier detection requires background information to determine contextual attributes and contexts. Collective outlier detection requires background information to model the relationship among objects to find groups of outliers.

II. Literature Review on Outlier Detection Methods

Outlier detection methods are of different types and are classified as statistical, clustering, distance, density and sliding window based outlier detection methods shown in fig 2.0.

2.1 Statistical based methods

Distribution based approach deals with statistical based methods that are based on the probabilistic data model. A probabilistic model can be either a priori given or automatically constructed using given data. Authors in [2] illustrated that object will be treated as an outlier if it does not suit the probabilistic model. In the probabilistic model, the techniques proposed to vary in terms of their complexity. The simple statistical techniques for novelty detection can be based on statistical hypothesis tests, which are equivalent to discordancy tests in the statistical outlier detection [3]. These techniques determine whether attest sample was generated from the same distribution as the "normal" data or not, and are usually employed to detect outliers.

The approach proposed in [4] considered a weighted sum of the distances from the k nearest neighbors to each data point and classifies those points which have the largest weighted sumsas outliers. The k nearest neighbors of each point are found by line arising the search space using a Hilbert space curve. This work is built upon previous techniques that prune the search space for nearest neighbors [5]. Then later, it partitions the data

space into a grid of hyper cubes of fixed sizes. If a hypercube contains many data points, such points are likely to be normal. Conversely, if a test point lies in a hypercube that contains very few examples, the test point is likely to be an outlier.

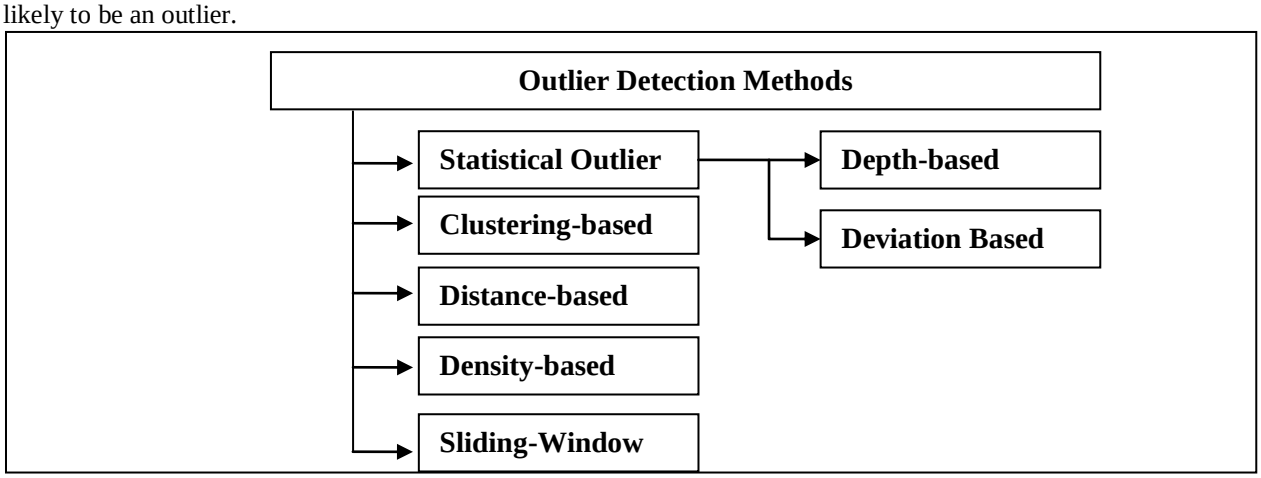

Fig 2.0 Outlier Detection Methods

Distance based approaches are known to face the local density problem created by the various degrees of cluster density that exist in a dataset. In order to solve the problem, density based approaches have been proposed. The basic idea of density based approaches is that the density around an outlier remarkably varies from that around its neighbors [6]. The density of an objects neighborhood is correlated with that of its neighbor's neighborhood. If there is a significant difference between the densities, the object can be considered as an outlier. To implement this idea, several outlier detection methods have been developed recently. The detection methods estimate the density around an object in different ways.

In paper [7] authors developed the local outlier factor (LOF), which is amongst the most commonly used method for outlier detection. LOF is influenced by variations like local correlation integral (LOCI) [8], Local distance based outlier factor (LDOF), and local outlier probabilities (LOOP).Many of these statistical tests, such as the one frequently used by Grubb's (1969), assume a Gaussian distribution for the training data and work only with univariate continuous data, although variants of these tests have been proposed to handle multivariate data sets. For example, a simple statistical scheme for outlier detection is based on the use of the box-plot rule.

In paper [9] authors proposed a variant of the Grubb's test for multivariate data. The Grubb's test computes the distance of the test data points from the estimated sample mean and declares any point with a distance above a certain threshold to be an outlier. This requires a threshold parameter to determine the length of the tail that includes the outliers (and which is often associated with a distance of three standards deviations from the estimated mean).

2.2 Distance based methods

Distance based techniques define as an outlier if its locality (or proximity) is sparsely populated. In distance based approaches, the distances between an object and its nearest neighbors are determined and then used to estimate the outliers of an object. Basically, the distance-based approaches assume that outliers are far apart from their neighbor objects[6]. Any appropriate distance measure can be used such as Euclidean distance, Mahalanobis distance, or some other measure of dissimilarity. Usually, the type of the variables affects the choice of distance measure. Several well-known methods based on this idea are discussed here.

In paper [10] authors presented a technique using which we can simultaneously cluster and discover outliers in data. This approach is the generalization of K-means approach and hence it is NP-Hard. It is an iterative approach and it converges to local optima. This algorithm is not suitable for all similarity measures. However, the number of outliers cannot be determined automatically.

Authors in [11] proposed a general framework for handling the three major classes of distance based outliers inclusive of the long established distance threshold based and the nearest neighbor based definitions in streaming environments is proposed. Two novel optimization principles to achieve scalable outlier detection are proposed, and they are minimal probing and lifespan-aware prioritization. This method is proven to be superlative for determining the outlier status of data points. But modern distributed multi-core clusters of machines are not used to its full advantage to improve scalability.

Orca is one of the most successful algorithms for the improvement of distance based outlier detection. It is based on a nested loop with randomization and a simple pruning rule. Orca-based outlier detection on a

multi-core CPU is proposed in [12]. Data parallelism and a multithread model are utilized in the proposed parallelization model. Here outlier score tables and cutoff values are shared for pruning among worker threads. A cache of the cutoff value for each worker thread is made and outlier-score tables are managed hierarchically. The proposed model can not work well for block partition, but it worked well for round-robin partition.

Authors in [13] applied a novel perspective on clustering high dimensional data. Here instead of trying to avoid the curse of dimensionality, dimensionality is embraced. It is shown that for high-dimensional data clustering, hubness is a good measure of point centrality. This paper states that hubs can be used effectively as cluster prototypes. Cluster based prototype method is proposed which proves to be better than K-means. This method provides better inter cluster separation in high dimensional data. The major drawback of this system is that it detects only hyper spherical clusters, just as K-Means. The work proposed in [14] shows the role of hubness in high dimensional data. They provide Anti-Hub method using reverse nearest neighbor counts for outlier detection. This method can efficiently find outliers in high dimensional data. But accuracy may be sacrificed to obtain efficiency sometimes.

Authors in [15] proposed distance based unsupervised method for outlier detection is proposed. It uses iterative random sampling. This method takes inspiration from the simple notion that outliers are not as easily selected as inliers in blind random sampling. Therefore selected objects are given more outlierness scores. A new measure called observability factor is developed using this idea. Moreover, the entropy of scores is proposed to provide the heuristic guideline to find the best size of the nearest neighborhood. But the performance of this method deteriorates for highest entropy values. Overall, it finds outliers effectively and can be used with the combination of other methods for better results.

The author in [16] proposed an approach for selecting meaningful feature subspace and conducting anomaly detection in the corresponding subspace projection. This approach aims to maintain the detection accuracy in high-dimensional circumstances. The suggested approach determines the angle between all pairs of two lines for one specific anomaly candidate: the first line is connected to the pertinent data point and the center of its adjacent points; the other line is one of the axis-parallel lines. Those dimensions which have a comparatively small angle with the first line are then chosen to constitute the axis-parallel subspace for the candidate. Then, a normalized Mahalanobis distance is introduced to measure the local outliers of an object in the subspace projection.

In [17] authors proposed an outlier detection approach to address data with imperfect labels and incorporate limited abnormal examples into learning is proposed. To deal with data with imperfect labels, likelihood values for each input data are introduced which denotes the degree of membership of an example concerning the normal and abnormal classes respectively. The proposed approach works in two steps. In the first step, a pseudo training data set by computing likelihood values of each example based on its local behavior is generated. Kernel k-means clustering method and kernel LOF based method to compute the likelihood values are presented. Then the generated likelihood values and limited anomalous examples are integrated into SVDD learning framework to construct a more accurate classifier for global outlier detection. This combines the local and global outlier detection concepts. By integrating these two, existing method explicitly handles data with imperfect labels and enhances the performance of outlier detection.

2.3 Density based

Density based algorithms possess quite a few significant advantages for data clustering such as i) the ability to detect arbitrarily shaped clusters, ii) the ability to handle noise and iii) they require just one time to scan raw data. Apart from that, such algorithms do not require prior knowledge of the number of clusters (k), unlike k-means algorithms where the number of clusters is provided in advance.DBSCAN [18], GDBSCAN and DENCLUEare all density based clustering algorithms that can be used to detect any arbitrarily shaped clusters. However, they are unsuitable for processing clusters in data streams.

An extension of the DBSCAN known as the incremental-DBSCAN [19] was developed. This method can proficiently add and remove points incrementally in data warehousing. It is capable of detecting arbitrarily shaped clusters but requires parameters tuning. For static data sets, the OPTICS algorithm is the solution for density-based clustering algorithms that are dependent on parameters. It contains two concepts for organizing points: i)the core distance and ii) the reachability distance.

Another improvement of the DBSCAN algorithm known as LDBSCAN [20] has also been proposed. This algorithm uses the concept of local density-based clustering. It is able to detect density based local outliers and noise. However, this algorithm does not work well in data streams. A two-phase scheme density-based algorithm known as Den Stream has been developed to cluster evolving data streams. In the first phase, this algorithm uses the fading window model to create a synopsis of the data. Then, in the second phase, the synopsis of the data stored from the first phase is utilized to provide the clustering result. This algorithm can handle arbitrary shaped clusters, but due to the numerous time vector calculations, it has high time complexity.

2.4 Cluster based

An improvement of the DenStream algorithm is rDenStream [21], which is a three-phase clustering algorithm. In this algorithm, previously discarded unimportant clusters are stored in a transitory memory. This approach ensures that this data has the chance to form clusters and increase the clustering accuracy.rDenStream can handle a huge number of outliers and its first two phases are comparable to those of DenStream but it has an additional phase known as theretrospect. This phase allows the algorithm to learn from the discarded data to increase its accuracy. From an experimental comparison, rDenStream outperforms DenStream in the initial phase. However, this algorithm needs more time and memory as compared to the DenStream, because it processes and saves the historical buffer before performing the detection.

The D-Stream algorithm [22] has also been proposed, which is capable of making automatic and dynamic adjustments to the data clusters without user specification with regard to the target time horizon and number of clusters. D-stream algorithm generates map the new incoming data into different segmented grids. A decay factor is used with the density of each data point in order to determine which data are recent and which are less important. The D-Stream algorithm is incapable of processing high dimensional data; however, the DenStream algorithm has no difficulty in processing such data. Additionally, D-Stream and DenStream have been found to outperform CluStream. Similar to D-Stream, MR-Stream creates cell partitions in the data space. Whenever a dimension is divided in half, a single cell goes through another division to form 2d sub cells, where d is the dimension of the data set. The division process can be set to a maximum limit by a user-defined parameter. The segmented cells are put into a quad tree structure, this allows for data clusters to be created at different resolution levels. The MR-Stream algorithm allocates all new data into the appropriate cells at every time stamp interval during the online phase and also updates the summarized data. In a comparison between MR-Stream and D-Stream, MR-Stream showed better performance. For clustering-based approaches, they always conduct clustering based techniques on the samples of data to characterize the local data behavior. In general, the sub-clusters contain significantly fewer data points than other clusters, are considered as outliers. In the work, the clustering techniques iteratively detect outliers to multidimensional data analysis in subspace. As clustering based methods are unsupervised without requiring any labeled training data, the performance of unsupervised outlier detection is limited.Another density-based clustering algorithm for streaming data is the DSCLUalgorithm. DSCLU uses micro clusters to detect suitable clusters, focusing on localizing dominant micro clusters on the basis of their neighbors' weight. It is able to detect clusters in multi-density environments. OPCluStream is another density-based algorithm for clustering data streams (Wang et al.2012). This algorithm utilizes a tree topology for organizing points and directional pointers to link all related points together. This algorithm is able to detect arbitrary and overlapping clusters.

Authors in [23] presented a cluster based method for outlier detection. A new global outlier factor and a local outlier factor and an efficient outlier detection algorithm are developed. This method can be used with the traditional distance based outlier detection methods. But it is suggested to use this method as a compliment with other methods of outlier detection. Later authors propose two algorithms namely Distance Based outlier detection and Cluster Based outlier algorithm for identifying and eradicating outliers using an outlier score are proposed. By cleaning the dataset and clustering based on similarity, one can remove outliers on the key attribute subset rather than on the full dimensional attributes of the dataset. This work suggests (based on the results obtained) that cluster based approaches produce better accuracy as compared to distance based methods.

Authors in [24] discussed two clustering algorithms namely BIRCH with K-means and CURE with K-means which are used for clustering the data items and finding the outliers in data streams. To analyze the experimental result, two performance factors are used such as clustering accuracy and outlier detection accuracy. In this paper, the proposed CURE with K-means algorithm has given good performance results when compared with the algorithm BIRCH with the k-means clustering algorithm. They discussed data stream clustering algorithms, which are highly used for detecting the outlier efficiently. In the latter part of the paper, authors discussed data stream clustering algorithms which are highly used for detecting the outlier efficiently. This paper has focused on two clustering algorithm namely CURE with K-means and CURE with CLARANS. Distinct sizes and types of datasets and two performance factors are considered for analysis. The performance metrics are clustering accuracy and outlier detection accuracy. In this paper authors proposed CURE with CLARANS clustering algorithm's performance is more accurate than the existing algorithm CURE with K-means. Authors discussed data stream are a new emerging research area in data mining. In this paper is to perform the clustering process in data streams and to detect the outliers in high dimensional data using the existing clustering algorithms like K-means, CLARA, CLARANS, and CURE. Finally the technique CURE clustering algorithm yields the best performance compared to other algorithms is proved in the experiment.

In paper [25] authors 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 probable candidates for outlier and we can be 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.

In paper [26] discussed various approaches 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. This paper reviewed different approaches to outlier detection used for K-Means algorithm for clustering dataset with some other. Clustering-based approaches to distance-based novelty detection include methods such as the k-means clustering. In this general type of methods, the "normal" class is characterized by a small number of prototype points in the data space. The minimum distance from a test point to the nearest prototype is often used to quantify abnormality. The methods use different approaches to obtain the prototype locations.The k-means clustering algorithm is perhaps the most popular method of clustering structured data due to its simplicity of implementation.

A new definition of a cluster-based local outlier is proposed in the literature, which takes in to account both the size of a point's cluster and the distance between the point and its closest cluster. Each point is associated with a cluster-based local outlier factor, which is used to determine the likelihood of the point is an outlier. This approach partitions the data into clusters using a squeezer algorithm, which makes a single pass over the data set and produces initial clustering results. The outlier factor is then computed for each point, and those points which have the largest factors are considered outliers. This approach is linearly scalable with respect to the number of data points and was found to work well with large data sets. Another technique that addressed computational efficiency was proposed in which an efficient indexing technique called CD-trees were used to partition data into clusters. Those points belonging to sparse clusters are declared anomalies.

Authors in [27] proposed a model based outlier detection approaches. Among them, support vector data description (SVDD) proposed has been demonstrated empirically to be capable of detecting outliers in various domains. SVDD conducts a small sphere around the normal data and utilizes the constructed sphere to detect an unknown sample as normal or outlier. The most interesting property of SVDD is that it can transform the original data into a feature space via kernel function and effectively detect global outliers for high-dimensional data. However, its performance is affected by the noise involved in the input data.

Another type of clustering is the model-based method that runs a hypothesized model for every cluster and determines which data will fit the model perfectly. The COBWEB algorithm is one such model-based algorithm and it is an incremental conceptual method to cluster data. This method uses the tree structure generated by a category function. It generates a hierarchical clustering in the form of a classification tree. In this form, each node keeps a notion and has a probabilistic description of that notion which summarizes the objects classified under the nodes. COBWEB can detect outliers but because it utilizes the tree structure, it has a limitation in terms of the capacity of the leaves.

CluDistream is an algorithm that has been developed in [28] based on the expectation maximization technique for clustering streaming data. This algorithm is only capable of dealing with the clustering problem in a landmark window with the expectation maximization executed at every node of the distributed network. Nonetheless, CluDistream has been found to have significant results, particularly when it is implemented in distributed stream environments where transmitted data can either be noisy or missing.

2.4 Sliding window Based

The SWEM algorithm has proposed in [29] to cluster the data streams in a temporal sliding window. This used the time based sliding window technique and the expectation maximization technique for clustering. This algorithm consists of two phases; in the first phase by scanning the data, it creates a synopsis of that data as micro components. After that, in the second phase this data summary is utilized to create global data clusters. This two-step structure is designed to deal with the limited memory and a single-scan processing problem of the data stream.This algorithm is able to detect noise and handle the missing data properly. SWEM when compared to the CluStream algorithm was found to show better performance in terms of time complexity and quality of clusters.

2.5 Other techniques

In paper [30] authors used SSODPU algorithm which is semi supervised. It deals with the problem of detecting outliers with only a few labeled positive examples. There are two main steps in this algorithm: Initially, some of the reliable negative examples will be extracted using K-NN. And the second step is the fuzzy clustering including both positive and negative example of an outlier. Here outliers are detected on the basis of new labeled examples. The limitation of this method is that accuracy is not up to the mark. In the semi-supervised method, the labeled and unlabeled data are used to detect the outliers. The semi-supervised

approaches are followed by the researchers to detect the outliers with several algorithms. The authors presented a fuzzy rough c-means clustering to detect the outliers. Authors proposed an outlier detection system. In this system, the normal instances are used to build the ensemble feature to detect the anomaly from the received instances. Authors used the entropy measure to detect the outliers. Initially, the steadfast negative samples are extracted from unlabeled and positive data, and then the outliers are detected based on the entropy score to remove the outliers. Also, authors) presented a score based outlier detection using stochastic network method. A semi-supervised cluster was also proposed in the literature to detect the outliers from the digital mammograms.

III. Conclusion

A huge set of techniques have been proposed for outlier detection; however, most of them have some intrinsic restrictions and issues. Even though outlier detection looks simple, it is certainly a highly challenging and complicated task. It is very difficult to define the normal behavior or a normal region and the definition of an outlier is highly dependent on the domain and the application. It is very hard to enumerate every possible normal behavior in a dataset for any particular application. Due to the above challenges, the outlier detection is an emerging research area. This survey explores a prior knowledge about the different outlier detection techniques. This concludes that the individual methods are not capable over uncertain and dynamic streaming data. So hybrid and mixed approaches should be designed for the selected application and data.

References

[1]. Aggarwal, Charu C., and Philip S. Yu. "Outlier detection for high dimensional data." *ACM Sigmod Record*. Vol. 30. No. 2. ACM, 2001.
[2]. Petrovskiy, M. I. "Outlier detection algorithms in data mining systems." *Programming and Computer Software* 29.4 (2003): 228-237.
[3]. Barnett, Vic, and Toby Lewis. *Outliers in statistical data*. Vol. 3. No. 1. New York: Wiley, 1994.
[4]. Angiulli, Fabrizio, and Clara Pizzuti. "Fast outlier detection in high dimensional spaces." *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer, Berlin, Heidelberg, 2002.
[5]. Ghoting, Amol, Srinivasan Parthasarathy, and Matthew Eric Otey. "Fast mining of distance-based outliers in high-dimensional datasets." *Data Mining and Knowledge Discovery* 16.3 (2008): 349-364.
[6]. Chandola, Varun, Arindam Banerjee, and Vipin Kumar. "Anomaly detection: A survey." *ACM computing surveys (CSUR)* 41.3 (2009): 15.
[7]. Breunig, Markus M., Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. "LOF: identifying density-based local outliers." In *ACM sigmod record*, vol. 29, no. 2, pp. 93-104. ACM, 2000.
[8]. Papadimitriou, Spiros, Hiroyuki Kitagawa, Phillip B. Gibbons, and Christos Faloutsos. "Loci: Fast outlier detection using the local correlation integral." In *Data Engineering, 2003. Proceedings. 19th International Conference on*, pp. 315-326. IEEE, 2003.
[9]. Aggarwal, Charu C., and Philip S. Yu. "Outlier detection with uncertain data." *Proceedings of the 2008 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics, 2008.
[10]. Chawla, Sanjay, and Aristides Gionis. "k-means–: A unified approach to clustering and outlier detection." Proceedings of the 2013 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2013.
[11]. Yu, Y., Cao, L., Rundensteiner, E. A., & Wang, Q. (2014, August). Detecting moving object outliers in massive-scale trajectory streams. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 422-431). ACM.
[12]. Oku, Junki, Keiichi Tamura, and Hajime Kitakami. "Parallel processing for distance-based outlier detection on a multi-core CPU." *Computational Intelligence and Applications (IWCIA), 2014 IEEE 7th International Workshop on*. IEEE, 2014.
[13]. Tomasev, N., Radovanovic, M., Mladenic, D., & Ivanovic, M. (2014). The role of hubness in clustering high-dimensional data. *IEEE Transactions on Knowledge and Data Engineering*, *26*(3), 739-751.
[14]. Radovanovic, Milos, Alexandros Nanopoulos, and Mirjana Ivanovic. "Reverse nearest neighbors in unsupervised distance-based outlier detection." *IEEE transactions on knowledge and data engineering* 27.5 (2015): 1369-1382.
[15]. Ha, Jihyun, Seulgi Seok, and Jong-Seok Lee. "A precise ranking method for outlier detection." *Information Sciences* 324 (2015): 88-107.
[16]. Zhang, Liangwei, Jing Lin, and Ramin Karim. "An angle-based subspace anomaly detection approach to high-dimensional data: With an application to industrial fault detection." *Reliability Engineering & System Safety* 142 (2015): 482-497.

[17]. Breunig, Markus M., et al. "LOF: identifying density-based local outliers." *ACM sigmod record*. Vol. 29. No. 2. ACM, 2000.
[18]. Ester, Martin, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. "A density-based algorithm for discovering clusters in large spatial databases with noise." In *Kdd*, vol. 96, no. 34, pp. 226-231. 1996.
[19]. Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (1998). Density-based clustering in spatial databases: The algorithm gdbscan and its applications. *Data mining and knowledge discovery*, *2*(2), 169-194.
[20]. Duan, Lian, Lida Xu, Feng Guo, Jun Lee, and Baopin Yan. "A local-density based spatial clustering algorithm with noise." *Information Systems* 32, no. 7 (2007): 978-986.
[21]. Amini, Amineh. "An adaptive density-based method for clustering evolving data streams." PhD diss., University of Malaya, 2014.
[22]. Tu, Li, and Yixin Chen. "Stream data clustering based on grid density and attraction." *ACM Transactions on Knowledge Discovery from Data (TKDD)*3.3 (2009): 12.
[23]. Wang, Xiaochun, Xia Li Wang, Yongqiang Ma, and D. Mitchell Wilkes. "A fast MST-inspired kNN-based outlier detection method." *Information Systems* 48 (2015): 89-112.
[24]. Vijayarani, Dr S., and Ms P. Jothi. "Partitioning Clustering Algorithms for Data Stream Outlier Detection." *International Journal of Innovative Research in Computer and Communication Engineering* 2.4 (2014): 3975-3981.
[25]. Pamula, Rajendra, Jatindra Kumar Deka, and Sukumar Nandi. "An outlier detection method based on clustering." *Emerging Applications of Information Technology (EAIT), 2011 Second International Conference on*. IEEE, 2011.
[26]. Deshmukh, Mr Mukesh K., and A. S. Kapse. "A Survey On Outlier Detection Technique In Streaming Data Using Data Clustering Approach."
[27]. Tax, David MJ, and Robert PW Duin. "Support vector data description." *Machine learning* 54.1 (2004): 45-66.
[28]. Zhou, Aoying, et al. "Distributed data stream clustering: A fast EM-based approach." *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*. IEEE, 2007.
[29]. Dang, Xuan Hong, et al. "An EM-Based Algorithm for Clustering Data Streams in Sliding Windows." *DASFAA*. 2009.
[30]. Daneshpazhouh, Armin, and Ashkan Sami. "Entropy-based outlier detection using semi-supervised approach with few positive examples." *Pattern Recognition Letters* 49 (2014): 77-84.