International Journal of Data Mining Techniques and Applications
Volume 5, Issue 1, June 2016, Page No.96-101
ISSN: 2278-2419

A Survey on Cluster Based Outlier Detection Techniques in Data Stream

S.Anitha[1], Mary Metilda[2]
[1]Research Scholar, Bharathiar University, Coimbatore, Tamil Nadu, India
[2]Asst. Prof., Queen Mary's College, Chennai, Tamil Nadu, India
E-mail: anitasenthil@gmail.com, metilda_dgvc@yahoo.co.in

Abstract: In recent days, Data Mining (DM) is an emerging area of computational intelligence that provides new techniques, algorithms and tools for processing large volumes of data. Clustering is the most popular data mining technique today. Clustering used to separate a dataset into groups that finds intra-group similarity and inter-group similarity. Outlier detection (Anomaly) is to find small groups of data objects that are different when compared with rest of data. The outlier detection is an essential part of mining in data stream. Data Stream (DS) used to mine continuous arrival of high speed data Items. It plays an important role in the fields of telecommunication services, E-Commerce, Tracking customer behaviors and Medical analysis. Detecting outliers over data stream is an active research area. This survey presents the overview of fundamental outlier detection approaches and various types of outlier detection methods in data stream.

Keywords: Clustering, Outlier Detection, Anomaly Detection, Data Stream,

I. INTRODUCTION

Detection refers to the process of finding patterns in data that do not conform to expected normal behavior. [2] Outlier detection is an (or Anomaly) essential problem for many Applications such as credit card fraud detection, insurance, risk analysis, weather prediction, Medical diagnosis, network intrusion for cyber security, detecting novelties in images and military surveillance for enemy activities and many other research areas. Hawkins [2] defines an outlier as an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism. An Outlier is defined as "an observation (or subset of observation) which appears to be inconsistent with the remainder of that set of data". Aggarwal states that "outlier may be considered as noise points lying outside of a set of defined clusters or outliers may be considered as the points that lie outside of the set of clusters but also are separated from the noise" [1].

Outlier detection (or anomaly) has been resulted to be directly involved in lot of domains. Outlier detection has been used to detect and remove unwanted data instance from large dataset. These unusual patterns are often known as outliers, anomalies, exceptions, aberrations, surprises in different domains of applications.[3] In data mining, outlier detection methods are non parametric that designed to manage large databases from high dimensional spaces. The separation of outliers for improving the quality of the data and reducing the impact of wrong values in the process of research. The importance of isolating outliers can improve the quality of stored data. Based on the observations, an outlier detection algorithms are applied various machine learning categories; the three fundamental approaches are supervised outlier detection techniques, Semi-supervised outlier detection techniques and unsupervised outlier detection techniques. In supervised techniques, classification is an essential machine learning concept. The primary aim of supervised approach is to learn a set of labeled data instance (training) and then classify an unseen instance into one of the class (testing). Except the 2 classes, the entire portions outside the classes, represented as outlier. Various types of classification algorithms are used for detecting outlier such as neural networks, Bayesian networks support vector machines (SVM), decision trees and regression models etc. These techniques are used to classify a new observations as normal (or) outliers. In Semi-supervised outlier detection techniques, some applications have used only trained data for normal class or only the abnormal classes. These techniques represented as semi supervised method. In this method, the one class classifier learns a limit around the normal objects and specifies any test objects outside this limit as an outlier. [5] Both supervised and semi supervised methods, some time unseen data objects are declared as an outlier, in such cases, a threshold is required to specify the particular data objects as an outlier. [Chow et al 1970]The generalization and rejection problems have been used to solve this problem. [Jain et al 1988]

In unsupervised method, cluster analysis a popular machine learning technique to group similar data objects into cluster. In this type determining the outliers with no prior knowledge of data, outlier maybe detected by clustering, where values which are similar are organized into groups or clusters. The values that fall outside of the set of cluster may be considered as outliers.[Lourcioco et al 2004]. Many data mining algorithms try to minimize the influence of outliers or eliminate them all together [4].

Clustered based outlier techniques belong to large and dense cluster, while outlier form very small cluster. Comparing with the supervised approaches, unsupervised data mining approaches are more feasible. There are two clustering methods: density based and partitioned based clustering. The density based method can produce outlying objects along with normal cluster. Partitioning techniques divides the object in multiple partitions; every single partition is called as cluster. The objects with in single clusters are of similar characteristics where the objects of different cluster have dissimilar characteristics in terms of dataset attributes distance measure is one of the feature space used to identify similarity or dissimilarity of patterns between data objects k-Mean, k-Medoid and CLARAN are partitioning algorithms [4]. The partition based clustering method is used for distance based outlier detection. There are many techniques that are used to

International Journal of Data Mining Techniques and Applications
Volume 5, Issue 1, June 2016, Page No.96-101
ISSN: 2278-2419

represents the outliers. Two steps are involved in outlier detection when the user analyzes it. First identifies outliers around the data set using set of inliers. Second, data request are analyzed and identified as outliers when attributes are different from attribute of inliers [4].

Applications of cluster analysis are Economic data, classification, Pattern Recognition, Image Processing, text mining etc. Single algorithm is not efficient to track problems from different area. In this research work, some algorithms are presented that are based on distance between two objects. The purpose of the study is to minimize the distance of each and every object from the centre of the cluster to which the object belongs. Clustering technique can group the data in to number of clusters. It reduces the size of database that will reduce computation time. To each cluster user can give certain radius to find outliers in data streams. A wide range of streaming data, such as network flows, sensor data have been generated. Analyzing and mining streams of data are interesting challenges for developer [1, 2]. The dynamic nature of evolving data streams, the role of outliers and clusters are often exchanged, and consequently new clusters often emerge, while old clusters fade out. It becomes more complex when noise exists. Generally, outliers do not belong to any cluster and belong to very small clusters but they can force to belong to a cluster where they are very different from other data member. This paper is organized as follows, Chapter II describes about the different methods and approaches on outlier detection. Chapter III discusses the various clustering methods in outlier detection techniques. Chapter IV provides a compact survey of existing outlier detection techniques using k-Means and k-Medoids in partitioning cluster method. Finally chapter V presents a conclusion and further enhancements.

II. METHODS AND APPROACHES OF OUTLIER DETECTION

In recent days many approaches are used to detect outliers over streaming data such as statistical distribution, Distance-based, Depth-based, Density-based outlier detection .In statistical distribution based approach; many tests are performed for single attributes. As in one-dimensional procedures, the distribution mean (measuring the location) and the variance-covariance (measuring the shape) are the two most commonly used statistics for data analysis in the presence of outliers by Rousseeuw and Leory in 1987. The use of robust estimates of the multidimensional distribution parameters can often improve the performance of outlier detection [9].Statistical outlier detection model generate a distribution for the given data set. It is represented by a multidimensional data where some attributes are discrete variables (e.g. IP address, etc.) while others are continuous ones (time, duration, source bytes, etc.). The survey of Kenji Yamanishiet al in [6] proposed that Gaussian mixture model used for continuous data. Here a Gaussian mixture model takes a form of a linear combination of a finite number of Gaussian distributions.In statistical method, the data distribution may be unknown. It requires knowledge about parameters of the data set, such as the data distribution. A distance-based approach was constructed to overcome the problem arise from statistical approach. Distance-based methods were originally proposed by Knorr and Ng [10]. The notion of outliers studied here is defined as follows: An object 0 in a dataset T is a DB (p, D)-outlier if at least fraction p of the objects in T lies greater than distance D from 0. Where DB (p, D) - Distance-Based outlier (detected using parameters p and D). We use the term DB(p, D)-outlier as shorthand notation for a Distance-Based outlier (detected using parameters p and D). [3] Where the values of p and D are decided by the user.It is suitable for situations where the observed distribution does not fit any standard distribution. it is well-defined for k-dimensional datasets for any value of k. [mahalanobi, 1936] ,[silvia cateni and Valentina colla constructed Various distance matrix were used for outlying degree.

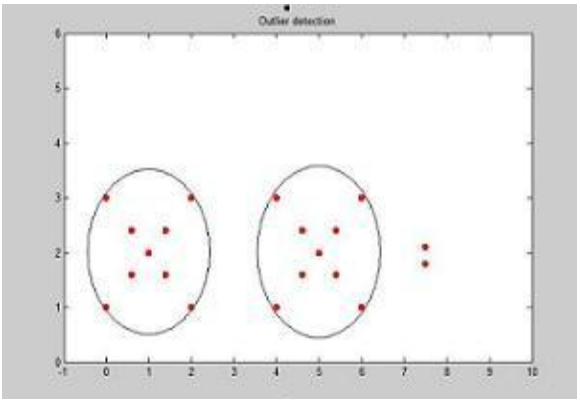

Figure 1: example of Outliers

For example, Mahalanobis distance is defined as equation 1 Where x is a data vector, μ is center of mass of data set and C is the covariance matrix. Distance between each point and its center of mass defined by the mahalanobis distance if the covariant matrix is the identity matrix, the mahalanobis distance become the Euclidean distance, data points located from the centre of mass are declared as outliers[30]. Unlike static data, streaming data are not in fixed length. Data streams may be a time series and multi dimensional [12]. The distance based outlier detection for streaming data can be solved using dynamic cluster maintenance research. Distance based approach deals to Operate on whole data, cannot give number of clusters. Even the Computation time will increases, it gave only one value as most expected outlier. The figer1.0, in a given set of n data points, objects which are significantly differ from other data called outliers.

$$D_M(X) = \sqrt{(X-\mu)^T C^1 (X-\mu)}$$

Bakar, Zuriana Abu, and Rosmayati Mohemad have proposed the performance of control chart, linear regression, and Manhattan distance techniques for outlier detection in data mining were discussed. Experimental results showed that outlier detection technique using control chart is better than the technique modelled from linear regression because the number of outlier data detected by control chart is smaller and better than linear regression. Further, experimental studies showed that Manhattan distance technique is best when compared with the other technique, (distance-based and statistical-based approaches) when the threshold values increased [15]. Depth-based methods are data-driven and avoid strong distributional expectations. They provide visualization of the data set via

International Journal of Data Mining Techniques and Applications
Volume 5, Issue 1, June 2016, Page No.96-101
ISSN: 2278-2419

depth bounded for a low dimensional input space. However, most of the current depth-based methods do not scale up with the dimensionality of the input space. [13] Chen and Yaxin presented survey of a novel statistical depth, the kernelized spatial depth (KSD). Choosing a proper kernel, the KSD can capture the local structure of a data set while the spatial depth fails. They demonstrated that by the half-moon data and the ring-shaped data. Based on the KSD, they proposed a novel outlier detection algorithm, by which an observation with a depth value less than a threshold is declared as an outlier using synthetic data and data sets from real applications. The proposed outlier detector is compared with existing methods. The KSD based outlier detection demonstrates competitive performance on all data sets tested when compared with other methods. Dang, Xin, and Serfling et al, carried out the research work in 2006, Based on depth functions, which order multidimensional data points by "outlyingness" measures and generate outline following the shape of the data set multivariate outlier detection is nonparametric and, with typical choices of depth function, robust. For depth-based outlier identifiers, masking and swamping breakdown points are defined. The values of these robustness measures are constructed for three depth functions, the spatial, the projection, and generalized Tukey[13][31].

Density-based Method can be seen as a non-parametric approach, where clusters are designed as areas of high density (relying on some unknown density-distribution) by Sander, Jörg, Martin Ester in 1998. In parametric approaches that try to approximate the unknown density-distribution generating the data by mixtures of k densities (e.g., Gaussian distributions), density-based clustering methods do not require the number of clusters as input and do not make any specific assumptions concerning the nature of the density distribution. As a result, density-based methods do not readily provide models, or otherwise compressed descriptions for the discovered clusters. A computationally efficient method for density-based clustering on static data sets is, e.g., DBSCAN [Ester, Martin, Hans-Peter Kriegel, 1996]. In density based approach, Outlier detection is done by a density of a particular data point is compared with density of its neighbor. The data points having a low density are declared as outliers. Density based models require the careful settings of several parameters. It requires quadratic time complexity. It may rule out outliers close to some non-outliers patterns that has low density [17].

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 recomputed the centred using the current cluster memberships
8 until the stopping criterion are met;

Figure 2: k-Means clustering algorithm.

Markus M. Breunig, Hans-Peter Kriege et al, have approached local outlier factor (LOF) used to measured the local outliers, is a ratio of local density of this point and local density of its nearest neighbor. LOF value of data point is high is declared as outlier. [Sander, Jörg, Martin Ester and et al, 1998] In this research, clustering algorithm GDBSCAN generalizing the density-based algorithm DBSCAN (Ester et al., 1996) in two important ways. GDBSCAN can cluster point objects as well as spatially extended objects according to both, their spatial and their non-spatial attributes. A performance evaluation, analytical as well as experimental, showed the effectiveness and efficiency of GDBSCAN on large spatial databases.

III. OUTLIER DETECTION USING CLUSTERING METHODS

Clustering algorithms are classified into the following types: partitioned clustering, hierarchical clustering, density-based clustering and grid-based clustering [4]. Clustering is a process of partitioning data sets into sub classes known as clusters. It gives us the natural grouping or cluster from the data set. It is unsupervised classification that means it has no predefined classes. This paper presents the various partitioning techniques in clustering algorithms and the advantages individually. The outlier detection is one of the challenging areas in data stream. In CluStream [1], the algorithm continuously maintains a fixed number of micro-clusters. Such an approach is especially risky when the data stream contains noise. Because a lot of new micro-clusters will be created for the outliers, many existing micro-clusters will be deleted or merged. Ideally, the streaming algorithm should provide some mechanism to distinguish the seeds of new clusters from the outliers. Discovery of the patterns hidden in streaming data imposes a great challenge for cluster analysis in that paper, Cao et al. have proposed a new algorithm named as DenStream, for clustering an evolving data stream. In this method, clusters of arbitrary shape in data streams, and it is insensitive to noise were discussed. The structures of p-micro-clusters and o-microclusters maintain sufficient information for clustering, and a novel pruning strategy is designed to reduce the memory utilization. The results of the research work carried out by a number of synthetic and real data sets gives the efficiency of DenStream in discovering clusters of arbitrary shape in data streams. HPStream [1] introduces the concept of projected cluster to data streams. However, it cannot be used to discover clusters of arbitrary orientations in data streams. Discovery of the patterns hidden in streaming data are a good effort for cluster analysis. The aim of clustering is to group the streaming data into meaningful classes. [16]

Irene Ntoutsi, Arthur Zimek et al have developed a density-based projected clustering algorithm, HDDStream, for high dimensional data streams. This work summarizes both the data points and the dimensions where these points are grouped together and maintains these summaries online, as new points arrive over time and old points expire due to ageing. The results illustrated the effectiveness and the efficiency of HDDStream. The Forest Cover Type dataset from UCI KDD Archive contains data on different forest cover types, containing 581,012 records. The challenge in this dataset is to predict the correct cover type from cartographic variables. The problem is defined by 54 variables of different types: 10

quantitative variables, 4 binary wilderness area attributes and 40 binary soil type variables. The class attribute contains 7 different forest cover types. In, this research, clustering quality of HDDStream was superior to the clustering quality of the canonical competitor. [28]

S. D. Pachgade and S. S. Dhande et al constructed a Hybrid approach for outlier detection method. Due to reduction in size of dataset, the computation time reduced. Then threshold value from user have taken and calculated outliers according to given threshold value for each cluster to get outliers within a cluster. Hybrid approach taken less computation time. The approach needed to be implemented on more complex datasets. Experiments were conducted in Matlab 7.8.0 (R2009a) on various data sets. Data are collected from UCI machine learning repository that provided various types of datasets. This dataset can be used for clustering, classification and regression. Dataset has multiple attribute and instances. Data File Format is in .data and .xls excel file or .txt or .csv file format. This data are useful to find cluster based the outliers. [19] "Density-based clustering for real-time stream data" by Chen, Yixin, and Li Tu constructs an algorithm D-Stream for clustering stream data using a density-based approach. The algorithm uses an online component which defines every input data into a grid and an offline component which computes the grid density and clusters the grids depends on the density of the cluster. The algorithm can find clusters of arbitrary shape. The researchers compare the qualities of the clustering results by D-Stream and those by CluStream.[13] Due to the on-convexity of the synthetic data sets, CluStream cannot get a correct result. Its quality cannot be compared to that of D-Stream. Therefore, they have compared only the sum of squared distance (SSQ) of the two algorithms on the network intrusion data from KDD CUP-99the computations made to detect and remove the sporadic grids in order to dramatically improve the space and time efficiency without affecting the clustering results with high speed data stream clustering. Both algorithms are tested on the KDD CUP-99. D-Stream is 3.5 to 11 times faster than CluStream and scales better results.

Frank Rehm and Frank Klawonn et al discussed an algorithm to calculate the noise distance in noise clustering based on the preservation of the hyper volume of the feature space. They have applied NC on FCM, other clustering algorithms, such as GK, GG and other prototype based clustering algorithms can be adapted. The aim of the study is not only to reduce the influence of outliers, but also to clearly identify them. [21]

S. Vijayarani et al discussed the research work, two partitioning clustering algorithms CLARANS and -CLARANS (Enhanced Clarans) are used for detecting the outliers in data streams. Two performance factors such as clustering accuracy and outlier detection accuracy are used for observation. By examining the computational results, it is observed that the proposed ECLARANS clustering algorithm performance is more accurate than the existing algorithm CLARANS. In this paper, they have analysed the performance of CLARANS and ECLARANS clustering algorithm s. the result of the computation showed that the proposed ECLARANS is more efficient than CLARANS clustering. Vijayarani, S., and P. Jothi have analysed the clustering and outlier performance of BIRCH with CLARANS and BIRCH with k-Means clustering algorithm for detecting outliers. They have used two biological data sets that are Pima Indian diabetes and breast cancer (Wiscosin). From that research, the clustering and outlier detection accuracy is more efficient in BIRCH with CLARANS clustering than BIRCH with k-Means with clustering [24]. S. Vijayarani and P. Jothi have compared two clustering algorithms namely CURE with k-Means and CURE with CLARANS is used to find the outliers in data streams. Various types of data sets and two performance factors such as clustering accuracy and outlier detection accuracy are used for analysis. From that research work, the proposed CURE with CLARANS clustering algorithm performance is more accurate than the existing algorithm CURE with k-Means.

IV. OUTLIER DETECTION BY k-MEANS AND k-MEDOIDS

In this paper, well known partitioning based methods k-Means and k-Medoids are compared. k-Mean algorithm is one of the centroid based technique. It takes input parameter k and partition a set of n object from k clusters. The similarity between clusters is measured in regards to the mean value of the object. The random selection of k object is first step of algorithm which represents cluster mean or center. By comparing most similarity other objects are assigning to the cluster .The survey given here explores the behaviour of these two methods. Elahi, et al. has proposed Efficient Clustering-Based Outlier Detection Algorithm for Dynamic Data Stream. This research work 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 [22].

Rajendra Pamula and Jatindra Kumar Deka et al carried out the research work titled "An Outlier Detection Method based on Clustering" based on clustering method to capture outliers. They have applied k-Means clustering algorithm is used to divide the data set into clusters. The points which are lying near the centroid of the cluster are not probable candidate for outlier and to prune out such points from each cluster. A distance based outlier score for remaining points were calculated. The computations needed to calculate the outlier score reduces considerably due to the pruning of some points. The results demonstrate that even though the number of computations is less, the proposed method performs better than the existing methods [23].

Christopher. T and T. Divya have analysed the performance of CURE with k-Means and CURE with CLARANS clustering algorithm. From the experimental results it is observed that the outlier detection accuracy is more efficient in CURE with CLARANS clustering while compared to CURE with k-Means with clustering. Neeraj Bansal and Amit Chugh have compared to the result of different Clustering techniques in terms of time complexity and proposed a new solution by adding fuzziness to

International Journal of Data Mining Techniques and Applications
Volume 5, Issue 1, June 2016, Page No.96-101
ISSN: 2278-2419

already existing Clustering [25]. Aruna Bhat explored a novel technique for face recognition by performing classification of the face images using unsupervised learning approach through k-Medoids clustering. Partitioning around Medoids algorithm (PAM) has been used for performing k-Medoids clustering of the data. k-Medoid clustering using PAM was observed to be more than that of k-Means clustering for all the data sets where outliers and noise was present. The PAM algorithm recognised the faces with different expressions more accurately as summarised [26]. Rani, Deevi Radha, et al proposed weighted k-Means assigned weights to the variables by using weighted k-Means for the dynamic data streams. The k-Means process cannot select the variables automatically for clustering and it is not efficient for large data sets, weights are assigned to the variables. In this process they considered three variables and identify the cluster initial centroid and assign the initial weights to the variable as 0.42, 0.64, and 0.13 and identify the value of the function before assigning weights. So the clustering process is again recomputed with the newly arriving data that becomes as inliers so that useful information may not be loosed and it is carried out until the user specified threshold values is reached. The experiment results shows that weighted k-Means is more efficient for detecting outliers. The Table 1 shows that the summary of various methods proposed by different researchers to find an outlier detection in DM clusters and data streams.

Table 1: Comparison of Various Articles

Paper Ref.No	Methods Used	Data Sets/Applications Used	Results	
			Accuracy	Sensitivity
13,31	D-STREAM	Network Intrusion Detection Stream Data(Mit Lincoln Laboratory)	96.5	--
16	DENstream	Network Intrusion Detection data set Charitable Donation data set (KDD CUP'99)	94%	95%
17	DBSCAN	SEQUOIA 2000 benchmark data	Suitable for large spatial databases.	
19	Modified and hybrid clustering approach	Medical Data Set, WDBC (Diagnosis)	Results are visualized, less computation time,	
21	FCM	Benchmark Data Set And Weather Data Set	Outliers are clearly identified	
23	k-Mean	Medical Data Set	the number of computations is less and better performance	
24	BIRCH with k-Means and BIRCH with CLARANS	Pima Indian Diabetes Dataset	accuracy is more Efficient in BIRCH with CLARANS.	
25	CURE with k-Mean and CLARANS	Breast Cancer Wisconsin and Pima Indian data set (768 instances and 8 attributes).	More accuracy in CURE with CLARANS	
26	k-Means and (PAM)k-mediods	Eigen Faces	(PAM) k-mediod produces better results.	
28	HDDstream	Forest cover type data from UCI	detecting drastic changes in the underlying stream population	

V. CONCLUSION

Outlier detection in Data streams has become a subject of dynamic research in computer science such as, distributed systems, database systems, and data mining. Lot of research work has been carried in this field to develop an efficient clustering algorithm for data streams. In this paper, popular outlier detection algorithms are surveyed and discussed based on clustering. In statistical method, the data distribution may be unknown. Streaming data are not in fixed length like static data .The distance based outlier detection may be useful for working with time series and multi dimensional streaming data. In density based approach accuracy is guaranteed in text and image data. These methods are structured into many criteria depending upon whether they work directly with data streams. Most clustering algorithms are not capable to find outliers in data stream. In addition, this paper discusses about partition based clustering methods k-Mean and (PAM) k-Medoid with data streams. k-Mean is computationally expensive but it is most useful for dynamic data streams. CLARAN is the best algorithm for high dimensional data. But (PAM) k- Mediods are only used for small set of data items, k-Mediods has less accuracy while compared to k-Means. The future work determines to develop an effective clustering algorithm for detecting outliers in data stream, considering the merits and demerits of the surveyed methodology.

References

[1] Aggarwal, Charu C., Jiawei Han, Jianyong Wang, and Philip S. Yu, "A framework for projected clustering of high dimensional data streams", Proc. of the Thirtieth international conference on Very large data bases, Vol. 30, 2004, pp. 852-863.

[2] Hawkins, Douglas M., Identification of outliers, Vol. 11. London: Chapman and Hall, 1980.
[3] Chandola, Varun, Arindam Banerjee, and Vipin Kumar, "Anomaly detection: A survey", ACM computing surveys, Vol. 41, No. 3, 2009.
[4] Han, Jiawei, and Micheline Kamber. "Data Mining: Concepts and Techniques, 2nd editionMorgan Kaufmann Publishers." San Francisco, CA, USA (2006).
[5] Chawla, Sanjay, and Pei Sun. "Outlier detection: Principles, techniques and applications." In Proceedings of the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Singapore. 2006.
[6] Yamanishi, Kenji, Jun-Ichi Takeuchi, Graham Williams, and Peter Milne. "On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms." Data Mining and Knowledge Discovery 8, no. 3 (2004): 275-300.
[7] Sivaram, Saveetha, "An Efficient Algorithm For Outlier Detections" , Global Journal Of Advance Engineering And Technologies, Vol 2,PP,35-40,January 2013.
[8] Yamanishi, J. T. K., and Y. Maruyama. "Data mining for security." NEC journal of advanced technology 2, no. 1 (2005): 63.
[9] Leroy, Annick M., and Peter J. Rousseeuw. "Robust regression and outlier detection." Wiley Series in Probability and Mathematical Statistics, New York: Wiley, 1987 1 (1987).
[10] Knor, Edwin M., and Raymond T. Ng. "Algorithms for mining distance based outliers in large datasets." In Proceedings of the International Conference on Very Large Data Bases, pp. 392-403. 1998.
[11] He, Zengyou, Xiaofei Xu, and Shengchun Deng. "Discovering cluster-based local outliers." Pattern Recognition Letters 24, no. 9 (2003): 1641-1650.
[12] Bu, Yingyi, Lei Chen, Ada Wai-Chee Fu, and Dawei Liu. "Efficient anomaly monitoring over moving object trajectory streams." In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 159-168. ACM, 2009.
[13] Chen, Yixin, Xin Dang, Hanxiang Peng, and Henry L. Bart. "Outlier detection with the kernelized spatial depth function." Pattern Analysis and Machine Intelligence, IEEE Transactions on 31, no. 2 (2009): 288-305.
[14] Dang, Xin, and Robert Serfling. "Nonparametric depth-based multivariate outlier identifiers and robustness properties." submitted for journal publication (2006).
[15] Bakar, Zuriana Abu, Rosmayati Mohemad, Akbar Ahmad, and Mustafa Mat Deris. "A comparative study for outlier detection techniques in data mining." In Cybernetics and Intelligent Systems, 2006 IEEE Conference on, pp. 1-6. IEEE, 2006.
[16] Cao, Feng, Martin Ester, Weining Qian, and Aoying Zhou. "Density-Based Clustering over an Evolving Data Stream with Noise." In SDM, vol. 6, pp. 328-339. 2006.
[17] Sander, Jörg, Martin Ester, Hans-Peter Kriegel, and Xiaowei Xu. "Density-based clustering in spatial databases: The algorithm gdbscan and its applications." Data mining and knowledge discovery 2, no. 2 (1998): 169-194.
[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] Pachgade, Ms SD, and Ms SS Dhande. "Outlier detection over data set using cluster-based and distance-based approach." International Journal of Advanced Research in Computer Science and Software Engineering 2, no. 6 (2012): 12-16.
[20] Chen, Yixin, and Li Tu. "Density-based clustering for real-time stream data." In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 133-142. ACM, 2007.
[21] Rehm, Frank, Frank Klawonn, and Rudolf Kruse. "A novel approach to noise clustering for outlier detection." Soft Computing 11, no. 5 (2007): 489-494.
[22] Elahi, Manzoor, Kun Li, Wasif Nisar, Xinjie Lv, and Hongan Wang. "Efficient clustering-based outlier detection algorithm for dynamic data stream." In Fuzzy Systems and Knowledge Discovery, 2008. FSKD'08. Fifth International Conference on, vol. 5, pp. 298-304. IEEE, 2008.
[23] Pamula, Rajendra, Jatindra Kumar Deka, and Sukumar Nandi. "An outlier detection method based on clustering." In Emerging Applications of Information Technology (EAIT), 2011 Second International Conference on, pp. 253-256. IEEE, 2011.
[24] Vijayarani, S., and P. Jothi. "An efficient clustering algorithm for outlier detection in data streams." International Journal of Advanced Research in Computer and Communication Engineering 2, no. 9 (2013): 3657-3665.
[25] Christopher, T., and T. Divya. "A Study of Clustering Based Algorithm for Outlier Detection in Data streams." In Proceedings of the UGC Sponsored National Conference on Advanced Networking and Applications. 2015.
[26] Bhat, Aruna. "K-MEDOIDS CLUSTERING USING PARTITIONING AROUND MEDOIDS FOR PERFORMING FACE RECOGNITION." Int. J. Soft Comp. Mat. Cont 3, no. 3 (2014): 1-12.
[27] Singh, Shalini S., and N. C. Chauhan. "K-Means v/s K-medoids: A Comparative Study." In National Conference on Recent Trends in Engineering & Technology, vol. 13. 2011.
[28] Ntoutsi, Irene, Arthur Zimek, Themis Palpanas, Peer Kröger, and Hans-Peter Kriegel. "Density-based Projected Clustering over High Dimensional Data Streams." In SDM, pp. 987-998. 2012.
[29] Rani, Deevi Radha, Navya Dhulipala, Tejaswi Pinniboyina, and Padmini Chattu. "OUTLIER DETECTION FOR DYNAMIC DATA STREAMS USING WEIGHTED K-MEANS." International Journal of Engineering Science and Technology 1, no. 3 (2011): 7484-7490.
[30] Mahalanobis, Prasanta Chandra. "On the generalized distance in statistics."Proceedings of the National Institute of Sciences (Calcutta) 2 (1936): 49-55.
[31] Dang, Xin, and Robert Serfling. "Nonparametric depth-based multivariate outlier identifiers, and masking robustness properties." Journal of Statistical Planning and Inference 140, no. 1 (2010): 198-213.