A Review On: Finding Outlier Points On Real Dimensional Data Sets

Bhagyashri Karkhanis, Sanjay Sharma

Abstract: With the latest rate of increase in research into finding outlier point has been studied broadly in area of data mining as well as machine learning. However as the appearance of enormous dimensional data sets in real-life applications to finding outlier point from outlier detection faces a series of new challenging problem in now-a-days. Detecting outliers is to identify the objects that extensively turn aside commencing the common distribution of the real data. Such that items may be seen as suspicious data items due to the different mechanism of generation. Various algorithms have already worked well in such an environment for finding outlier point. Consequently, machine learning methods are developing up-to-date outlier detection methods becomes insistent tasks.

Index Terms: Intrinsic dimension, k nearest neighbours (k-NN), local outlier factor (LOF), local projection-based outlier detection (LPOD), local projection score (LPS), outlier detection, resolution based outlier factor (ROF).

1 INTRODUCTION

To finding outlier point is a key problem of research area in data mining as well as machine learning that initial focus to find local outlier point that are significantly different, outstanding and unpredictable regarding the mainstream of dimensional data in an input database [1]. In recent years various research are required to data collected and transferred in the arrangement of main data streams for finding outlier point. This creates latest technological opportunities as well as challenges for research attempts in outlier detection. To finding outlier point of main data stream is a real-time challenging problem continuous and well thought-out i.e. implicitly by arrival sequence or clearly by timestamp evolution of items. Application area of main data streams consist of network traffic, telecommunications data, financial market data, and data from sensors that scrutinize the weather and environment, surveillance video and soon. Outlier detection from stream data can find items i.e. objects or points that are uncharacteristic or unbalanced about the popular of items in the entire or a horizon/window of the stream. Detecting outliers is to categorize the objects that significantly diverge from the common allocation of the data. Finding outlier point in main data streams can be valuable in many research areas such as analysis and scrutinizing of network traffic data e.g., connection-oriented records, web log, wireless sensor networks and financial transactions, etc.

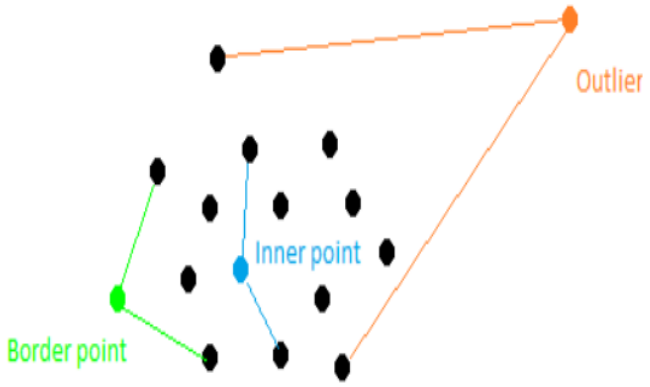

Fig.1 *Outlier Detection points*

To encourage this research is that outliers accessible in real data streams are set in a few lower-dimensional subspaces. At this time, a subspace refers to as the real data space of outlier points. The survival of projected outliers is inspired that as the data dimensionality reduce on outlier data have a tendency to develop into regularly far-away from each other. Thus, the high esteem of data point's outlier-ness will be converted into progressively more fragile and thus undistinguishable. Only in sensible data or small dimensional data subspaces can considerable outlierness of data be scrutinized. This happening is usually referred to as the curse of dimensionality. For the reason that most of modern outlier detection methods achieve outlier detection in the full data space thus the projected outliers cannot be found by any of these techniques. This will show the way to a loss of remarkable and potentially valuable unusual patterns hidden in high-dimensional data streams. However their dimensions exploited for calculating point's outlierness are not gradually update and lots of techniques involve various scans of outlier data making them incompetent of handling data streams. For example, [2][3] use the Sparsity Coefficient to calculate data sparsity and this is based on technique an equi-depth data partition that has to be cut down frequently from the data stream. This will be expensive and such updates will require multiple scans of data. [4][5]6] use data sparsity metrics that are involving the concept of concept of k-nearest neighbours (k-NN) to calculate distance. This is not right way for data streams also as one examination of data is not adequate for maintaining k-NN information of data points.

- *Bhagyashri Karkhanis is currently pursuing masters degree program in computer science and engineering, India.Email: karkhanis.bhagyashri26@gmail.com*
- *Sanjay Sharma is currently working as assistant professor at Oriental Institute of Science and Technology, India. Email: sanjaysharma@oriental.ac.in*

One the other hand, the techniques for approaching to finding outlier point in data streams [7] rely on full data space to detect outliers and thus projected outliers cannot be determined by these techniques. As such, it is attractive to propose a various new techniques that well explain the disadvantages of these existing methods.

2 Problem Statement

Outlier detection from stream data can find items i.e. objects or points that are abnormal or irregular regarding the common of items in the entire or a horizon/window of the data stream. They make available a collection of solutions to embark upon these disadvantages. But they focus on several of these open research issues, such as the relationship between numerous characteristic reduction methods and the resulting classification accuracy. The main objective is to recognize a set of features that best estimated the novel data without a reduction in the classification result. Other problems are based on computational cost of feature reduction algorithms and large amount of data upcoming requires developing computationally efficient feature reduction techniques which can be achieved concurrently. To finding outlier point various algorithms are originated upon statistical modeling techniques it can be any of them whether predictive or direct. Predictive techniques use tagged data using training sets to produce a finding outlier point data model i.e. contained by which outliers reduce for a domain which is subsequently utilized to categorize original data objects. Bit direct techniques are consist of deviation, proximity, statistical clustering, and density based techniques, refer to those in which labelled training sets are occupied and for that explanation the organization of objects as finding outlier point is implemented through the measurement of statistical heuristics. Although characteristically more composite than predictive techniques, direct methods are not as much of constrained as detection is not dependent upon pre-defined models.

3 Various Outlier Detection Techniques

Today's various systems are capable to generate and capture real-time data continuously. Various application works on this real-time data acquisition systems, condition monitoring systems, and financial activity systems. It is one of the challenging tasks on real data to efficiently detect outliers occurring in definite data streams. Conventional outlier detection approaches are no longer realistic as they only deal with statics data sets and entail multiple scans of data to generate effective results. In data streams surroundings outlier detection algorithms (e.g., [21]) need to process each data item within an exacting time constraint and can only meet the expense of to analyze the whole real data set with a distinct scrutinize of information. In finding outlier point has extensively research problem in fault-tolerance, anomaly detection, credit card fraud detection, medical diagnosis in various types of streaming data. To finding outlier point are uncharacteristic patterns in data; explicitly they are outlines that do not demonstrate normal behaviour. The extensive diversity of finding outlier point methods are summarized and categorized in the remainder of this research paper. In the next five sections, here they categorized to finding outlier point techniques are as follows:

3.1 Nearest neighbour
3.2 Density
3.3 Cluster
3.4 Statistical approach
3.5 Robust distance and
3.6 Depth based outlier detection techniques.

Each outlier detection techniques are described as follows:

3.1 Nearest Neighbour Based Outlier Detection Techniques:

Nearest neighbor based anomaly detection techniques require a distance or similarity measure between two data points. Nearest-neighbor based algorithms allocate the anomaly score of data instances comparative to their neighbourhood. They take for granted that finding outlier points are distant from their neighbours points or that their neighborhood is sparse.

3.2 Density Based Outlier Detection Techniques:

Density based outlier detection techniques calculate approximately the thickness of the neighborhood of each data instance. An illustration that deception in a neighborhood with low density is asserted to be outliers while an illustration that lies in a dense neighborhood is announced to be common. Density based techniques perform inadequately if the data has regions of unstable densities.

3.3 Cluster Based Outlier Detection Techniques:

In this technique here they use various data cluster is a collection of data objects related to one another within the equivalent data object cluster and remaining dissimilar data objects to the other clusters.

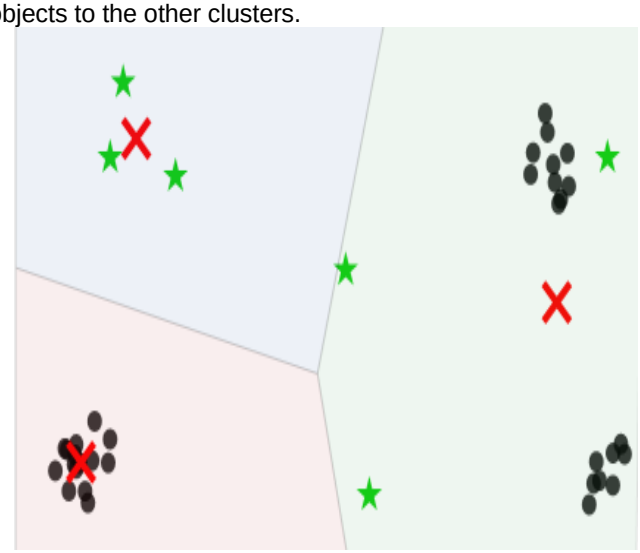

Fig.2 *Clustering with outliers*

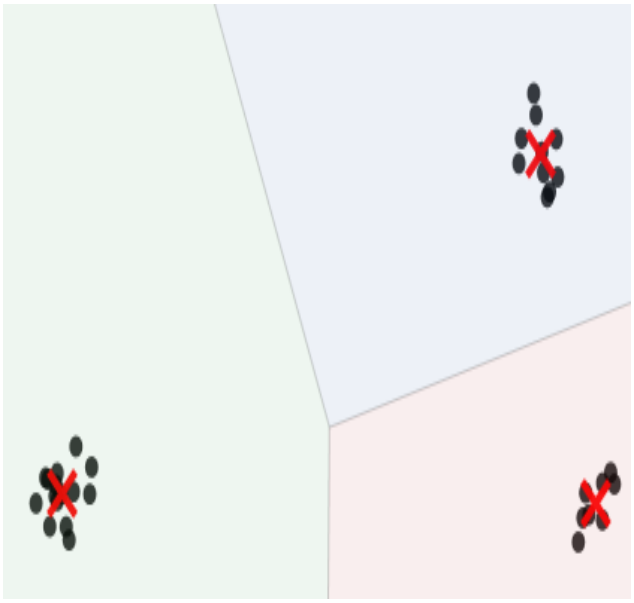

Fig.3 *Clustering without outliers*

Normal data fit in to a data cluster in the given entire data while outliers either they do not be in the right place to any another cluster region. Clustering is not a novel idea but data clustering together with finding outlier point is a contemporary scientific restraint under fast development of data points. Anomaly detection procedures can deal with is any one of the three modes: Supervised, Semi supervised and unsupervised anomaly detection.

3.4 Statistical Approach Based Outlier Detection Techniques:

Statistical approaches were the most ancient algorithm used for outlier detection. Some of the preliminary are suitable only for single dimensional data sets [22]. Statistical models are usually corrected to quantitative real data sets or at the very smallest amount quantitative ordinal data distributions where the ordinal data can be altered to appropriate statistical values for statistical processing. It [23] used the straightforward statistical to finding outlier point methods via casual box plots to locate particular outliers point in both univariate and multivariate case.

3.5 Robust Distance Based Outlier Detection Techniques:

Outlier observations can be finding out by using various distance based methods. One can get outliers by distance for each inspection using Location and Scale Estimator. The following are the process endured for detection of anomaly observation using robust distance.

3.6 Depth Based Outlier Detection Techniques:

Data depth is an important concept to Multivariate data analysis. Using the different conception of data depth, one can compute depth values for all sample points in the data cloud. Order the depth values based on center noticeable ranking. It means that the data points with the highest depth called the deepest or central point or it basically called center. The data points with lowest depth values are called outliers. Based on this arranging of intensity one can calculate Multivariate location, scale, skewness and kurtosis and Graphical methods such as Bag plot, Sunburst plot, Perspective plot, DD plot, Contour plot, Blotched bag plots for analyzing the distributional characteristics of the Multivariate data cloud and detect outliers.

4 Literature Survey

The common problem of identifying outliers has been addressed by various approaches that can be approximately confidential as global versus local outlier models. An overall model shows by outlier model show the ways to a binary decision of whether or not a given data object is to finding outlier point. A local outlier idea was to a certain extent allocates a degree of outlierness to each object. Such an “outlier factor” is a value distinguishing each object in “how much” this object is an outlier. Many applications where it is attractive to rank the outliers in the database and to get back the top-n outliers a local outlier approach is apparently desirable. A different arrangement of outlier approaches discriminates between supervised and unsupervised approaches. A supervised approach is supported by set of various data observations where the circumstance of being an outlier point or not is acknowledged and the differences between those unusual categories of surveillances is become skilled at various data points are required [8]. Generally, supervised methods are also comprehensive and can be measured as very unbalanced classification difficulties i.e. class of finding outlier point has intrinsically reasonably. In another cases finding outlier point is come across as an unsupervised difficulty from the time when one does not have an adequate amount of earlier information for supervised learning methods. But statistical methods to the identification of finding outlier point are based on acknowledged distributions of objects. Various data objects are tests and they are optimized for each data distribution reliant on the definite constraints of the equivalent allocation the number of anticipated outliers and the space where to suppose an outlier. Problems of these conventional approaches are observably the essential theory of a precise distribution in order to concern a precise test. Additionally, all analysis is univariate and examines a single attribute to find out an outlier point. Related approaches join given models and supervised learning methods but still imagine the distribution of objects to be known in progress [9][10]. Occasionally, the data are imagined to consist of k Gaussian distributions and to signifies and standard deviations are calculated data driven. On the other hand, these schemes are not really robust since mean and standard deviation are to a certain extent sensitive to outliers and the prospective outliers are still measured for the calculation step. In [12], author has to finding distances from the k-nearest neighbors and they are used that data objects are categorized by their distances to their k^{th} nearest neighbor. On the other hand, as alteration to high-dimensional data only the time-complexity concern is tackled. The intrinsic difficulties of high-dimensional data are not concentrating on this approach. Different problems are even intensified since the estimate is based on space filling curves. Another approximation based on suggestion points is proposed in [11]. This estimation is only on low-dimensional data shown to be precious. The initiative of using the k nearest neighbours already resembles density based approaches that think about ratios between the local

density around an object and the local density approximately its neighbouring objects. These approaches consequently initiate the idea of local outliers. The basic concept of this algorithm is to find data object of the database denoting a degree of outlierness based on density-based local outlier factor (LOF) [13]. In this paper [14], the author has to propose one more local outlier detection schema named local outlier integral (LOCI) based on the various observation of data object in a multi-granularity deviation factor (MDEF). Here author shows that the difference between the LOF and the LOCI outlier model here they uses ε-neighbourhoods rather than k^{th} nearest neighbours for using MDEF of LOCI. The authors propose an estimated method for calculating the LOCI values of each database object for any ε value. The consequences are demonstrated as a relatively intuitive outlier plot. This way, the approach becomes much less responsive to input parameters. In addition author has to initiate for finding outlier point based method on the LOCI model method. Another method resolution-based outlier factor (ROF) [15] is a combination of the local and the global outlier standard and their schema is based on the new scheme of a transform of resolution. Approximately, the "resolution" identifies the number of objects considered to be neighbours of a given real data object and these method is data driven based on distances rather than on formations like the k nearest neighbors or a ε-neighborhood that rely on user-specified parameterization. The thought remind you of a grid-based subspace clustering move toward where not dense but sparse grid cells are required to description objects within sparse grid cells as outliers. Since this is exponential data dimensionality discovers the most considerable arrangement of attributes where the point is an outlier. This is an attractive characteristic because an explicit and concise explanation why a definite point is measured to be an outlier has not been presented by any other outlier detection model so extreme. In this method author has aiming to give explanation to find outlier point here author deriving subsets of characteristics where an object is an outlier most considerably supported on a global outlier model method. Using this outlier models the new proposed technique is unsupervised and can be think about as a local concept. As using this local outlier detection models they have shown enhanced exactness than global outlier detection models. Consequently, as one of the most important local methods, LOF will be used as participant in evaluation to our novel approach. In this paper [16], author has tried to develop a better learning method to identify outliers out from normal observations. The concept of this learning method is to make use of local neighbourhood information of an observation to determine whether it is an outlier or not. To confine the neighborhood information precisely an idea local neighbourhood information concept called LPS is initiated to compute the anomalous degree of an apprehensive observation. Formally, the LPS are dependable with the perception of nuclear norm and can be acquired by the procedure of low-rank matrix approximation. Furthermore, distinct offered distance-based and density-based detection methods the proposed method is robust to the parameter k of k-NN embedded within LPOD. Using this method they are effectiveness algorithms on applying various outlier data sets. Experimental results show that the LPS are good at ranking the most excellent candidates for individual outliers and the show of LPOD is capable at many characteristics. While LPOD make use of k-NN to get neighbourhood information its competence relies on k-NN and its concert will be influenced by the distance formulation of k-NN to some area. Here author has introduce a new outlier scoring method for finding local outliers to distinguish between inliers and outliers in the surrounding area of a test position the continuous intrinsic dimension (ID) which has been shown to be equivalent to determine of the discriminative power of similarity functions. Continuous ID allows for inliers members of a subspace cluster with other members of the same cluster and distinguishes ability from non-members of local outliers. Local outliers have a tendency to increase in the estimated value of their continuous intrinsic dimension can be regarded as an extension of Karger and Ruhl's expansion dimension to a statistical setting in which the sharing of distances to a uncertainty point is modeled with continuous random variable. The proposed [17] author has local outlier score, IDOS well-known LOF outlier score can be summarized by explanation of the model of continuous intrinsic dimensionality initiated. Here author has to compare with IDOS; LOF is make public to have the potential for assessment of local density within local clusters i.e. groups of inliers than IDOS has in its measurement of local intrinsic dimensionality which would make it harder to discriminate outliers in the neighborhood of such clusters values of data objects. An experimental analysis shows that the precision of IDOS considerably increase to finding outlier point based on scoring methods mainly when the real data sets are huge and high-dimensional datasets show their superiority in terms of both effectiveness and efficiency. In this paper, they propose [18] novel parameter-free approach angle-based outlier detection (ABOD) and several alternatives evaluating the variance in the angles between the differences vectors of a point to the other points to finding outlier point based on the variance of angles between pairs of data points. This approach the consequences of the "curse of dimensionality" are alleviated on mining high-dimensional data where distance-based approaches often fail to offer high quality results. The basic concept of this method ABOD, they proposed two alternatives: Fast ABOD as acceleration suitable for low-dimensional but big data sets and LB-ABOD, a filter-refinement approach as acceleration suitable also for high-dimensional data. The main advantage of this proposed method does not rely on any constraint collection manipulating the feature of the accomplished ranking and here author has try to find rank the best candidates for being an finding outlier point with high precision and recall value. Here experimental assessment has to compare angle-based outlier detection to the well-started distance-based technique LOF for a variety of artificial data set and a real life data set and give you an idea about angle-based outlier detection to achieve mainly well on high-dimensional data. As increasing dimensions of data objects it is difficult to find out data points which are not fitting in group i.e. cluster called outlier. This method is using to finding outlier point has significant in real life applications area of fraud detection, intrusion detection and various areas in which increasing data dimensions. Here author has to propose another method to divides original high dimensional data set in subspace clusters using subspace clustering method

and here they try to improved k-means algorithms outlier cluster is establish which is additional amalgamated with other clusters depending upon compromise task. Various outlier clusters which are not going to combine with any other subspace cluster to find final outlier cluster. Here author [19] investigates various researches over many concepts of high dimensional data mining, information retrieval to finding outlier point in multi dimensional data ensemble subspace clustering, spam detection, improved k-means algorithm based on association rules. As these type of data is require to information systems so all these concepts can be used for improvement in data mining as well as machine learning methods. All these approaches are helpful for designing many strong applications for information retrieval. One application can be Spam Outlier Detection using Ensemble subspace clustering. In which spam outliers in analysis dataset of e-commerce can be detected. In this subspace clustering can be done trailed by outlier detection and again ensemble with other subspaces for enormous accuracy. In progress if we append spam detection logic then there will not be any concern for fraud reviews by someone. Whatever clusters are recognized as an outlier cluster from high dimensional data sets these can be highlighted or in some cases make some authorized essential accomplishments against all these entities. Second is they can put into practice elimination logic in datasets so that while performing data analysis when outliers are detected primarily if coming data is belonging to same dimension set will be rejected form adding it to the database. In this paper, here they propose [20] a hybrid semi-supervised anomaly detection model for high-dimensional data. Here author has using proposed detection model that consists of two parts: a deep auto encoder (DAE) and an ensemble *k*-nearest neighbor graph-(*K*-NNG) based anomaly detector. The deep auto encoder (DAE) is promoting from the ability of nonlinear mapping method and to begin with only trained the essential features of data objects in unsupervised mode and to transform into high-dimensional data. In this method they are sharing of the training dataset is more dense in the compact feature dimensional data space to various nonparametric KNN-based detect anomaly detectors method with a part of a real life dataset rather than using the whole specific training set and this process greatly condenses the computational charge. Experimental results and statistical significance analysis shows that proposed method is evaluated on several real-life datasets and their performance confirms that the proposed hybrid model improves the anomaly detection accuracy and also they reduces the computational complexity than standalone algorithms.

5 Conclusion

To finding outlier point it is a significant part of machine learning task with many critical applications, such as medical diagnosis, fraud detection, and intrusion detection. Due to large amount of data objects in real-life applications outlier detection faces various challenges to find these outlier points, so as we reduce the dimensionality efficiently they enlarge data object value or they combine traditional algorithms to build strong approximation for both high dimensional data and low dimensional data. High-dimensional data can be seen as part of the variety challenge of big data. To understand, the new challenge in dimensionality and understandability. Volume of data increases, but also the dimensionality; a large set of low-dimensional sensors can be seen as a high-dimensional multivariate time series. In particular, we try to develop new outlier detection methods from two perspectives i.e. detecting the out-of-the-way characteristics of a data object and detecting out-of-the-way data objects of a data set. We try to improve high dimensional data by using subscale algorithm to detect rank based outlier by its neighboring behavior.

6 References

[1]. J.Han and M.Kamber. Data Mining: Concepts and Techniques. Morgan Kaufman Publishers, 2000.

[2]. C. C.Aggarwal and P.S.Yu. Outlier Detection in High Dimensional Data. In Proc. of 2001 ACM SIGMOD International Conference on Management of Data (SIGMOD'01), Santa Barbara, California, USA, 2001.

[3]. C.Zhu, H.Kitagawa and C.Faloutsos. Example-Based Robust Outlier Detection in High Dimensional Datasets. In Proc. of 2005 IEEE International Conference on Data Management (ICDM'05), pp 829-832, 2005.

[4]. J.Zhang, M.Lou, T.W.Ling and H.Wang. HOS-Miner: A System for Detecting Outlying Subspaces of High-dimensional Data. In Proc. of 30th International Conference on Very Large Data Bases (VLDB'04), demo, pages 1265-1268,Toronto, Canada, 2004.

[5]. J.Zhang, Q.Gao and H.Wang. A Novel Method for Detecting Outlying Sub-spaces in High-dimensional Databases Using Genetic Algorithm. 2006 IEEE International Conference on Data Mining (ICDM'06), pages 731-740, Hong Kong, China, 2006.

[6]. J.Zhang and H.Wang. 2006. Detecting Outlying Subspaces for High-dimensional Data: the New Task, Algorithms and Performance. Knowledge and Information Systems (KAIS), 333-355, 2006.

[7]. C.C.Aggarwal. On Abnormality Detection in Spuriously Populated Data Streams. SIAM International Conference on Data Mining (SDM'05), Newport Beach, CA, 2005.

[8]. C.Zhu, H.Kitagawa, and C.Faloutsos. Example-based robust outlier detection in high dimensional datasets. In Proc. ICDM, 2005.

[9]. G.Williams, K.Yamanishi, and J.Takeuchi. Online unsupervised outlier detection using finite mixtures with discounting learning algorithms. In Proc. KDD, 2000.

[10]. K.Yamanishi and J.Takeuchi. Discovering outlier filtering rules from unlabeled data: combining a supervised learner with an unsupervised learner. In Proc. KDD, 2001.

[11]. Y.Pei, O.Zaıane, and Y.Gao. An efficient reference-based approach to outlier detection in large datasets. InProc. ICDM, 2006.

[12]. S.Ramaswamy, R.Rastogi, and K.Shim. Efficient algorithms for mining outliers from large data sets. InProc. SIGMOD, 2000.

[13]. M.M.Breunig, H.P.Kriegel, R.Ng, and J.Sander. LOF: Identifying density-based local outliers. InProc. SIGMOD, 2000.
[14]. S.Papadimitriou, H.Kitagawa, P.Gibbons, and C. Faloutsos. LOCI: Fast outlier detection using the local correlation integral. In Proc. ICDE, 2003.
[15]. H.Fan, O.R.Zaıane, A.Foss, and J.Wu. A nonparametric outlier detection for efficiently discovering top-N outliers from engineering data. In Proc. PAKDD, 2006.
[16]. Huawen Liu, Member, IEEE, Xuelong Li, Fellow, IEEE, Jiuyong Li, Member, IEEE, and Shichao Zhang, Senior Member, IEEE "Efficient Outlier Detection for High-Dimensional Data" IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS, 2017.
[17]. Jonathan von Brunken, Michael E.Houle, and Arthur Zimek, "Intrinsic Dimensional Outlier Detection in High-Dimensional Data" NII-2015 - 003E, Mar. 2015.
[18]. Hans-Peter Kriegel, Matthias Schubert, Arthur Zimek" Angle-Based Outlier Detection in High-dimensional Data" ACM 978-1-60558-193, 2008.
[19]. Suresh S.Kapare, Bharat A.Tidke, "Spam Outlier Detection in High Dimensional Data: Ensemble Subspace Clustering Approach" IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (3), 2015, 2326-2329.
[20]. Hongchao Song, Zhuqing Jiang, Aidong Men, and Bo Yang, "A Hybrid Semi-Supervised Anomaly Detection Model for High Dimensional Data" Comput Intell Neurosci. 2017.
[21]. K.Das and J.Schneider, "Detecting anomalous records in categorical datasets," in Proceedings of the ACM KDD, pp. 220–229, 2007.
[22]. Grubbs, F.E., 1969. Procedures for detecting outlying observations in samples. Technometrics, 11: 1-21.
[23]. Laurikkala, J.,M.Juhola1 and E.Kentala, 2000.Informal identification of outliers in medical data. In: Fifth International Workshop on Intelligent Data Analysis in Medicine and Pharmacology, pp: 20-24.