International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Review on a Collaborative Way to Detection Outlier Points on Various Data Streams

Isha Sinha[1], *Prof. Rajni Kori*[2]

[1]Dept. of Computer Science and Engineering, LNCTE, Bhopal
ishasinha616@gmail.com
[2]Dept. of Computer Science and Engineering, LNCTE, Bhopal
kori07rajni@gmail.com

ABSTRACT

In the previous time, data continued to evolve, making sound data in statistics the great challenge of obtaining more which became an important issue being researched in various fields within the application domain. The previous discovery was a breakthrough in many scientific research studies because it had a negative impact on results. Detecting outliers is to identify the most reversible contents to enable normal distribution of real data sets. Various algorithms for finding the outlier point have already worked well in such a field. Consequently, machine learning methods develop innovative methods before they become permanent tasks.

Index Terms— Outlier detection, k nearest neighbours (k-NN), local outlier factor (LOF), local projection score (LPS), intrinsic dimension.

Introduction

Determining outlier point output is a key area of machine learning research that focuses first on locating a very different, unique and unexpected local point of view in terms of data entry generated in the input database [1]. In recent years, various studies are needed on the data collected and transmitted to the planning of large data streams to find an outlier point. This creates the latest technical opportunities and challenges for research efforts on external acquisition. To finding the outlier point of data transmission which is a real ongoing and well-thought-out challenge that is entirely in the sequence of arrival or obviously with the appearance of a stamp of objects. The data stream application area contains network traffic, telecommunications data, financial market data, and data from weather and environmental sensors, video surveillance and more recently. Outlier broadcast data acquisition can detect items, i.e., objects or points that are uncharacteristic or unbalanced in relation to the popularity of the whole object or horizontal/window of the data stream.

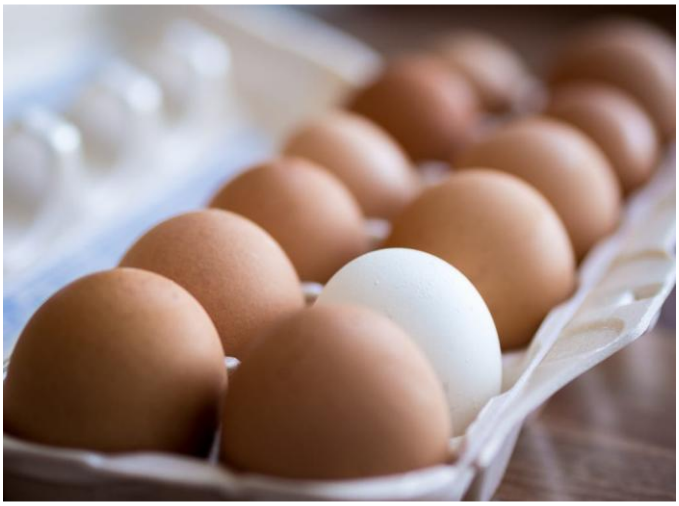

Figure: Outlier Detection points of similar data streams

The process of outlier acquisition, now-a-days, uses modern methods and methods of machine learning with very promising results. There are various categories of machine learning algorithms that can be used for outlier acquisitions in a variety of fields and applications including neural networks, in-depth learning algorithms, remote-based algorithms, straightforward models, and vector-based models. All of the above are two main sets, namely, supervised and unsupervised learning algorithms. For outlier acquisition purposes, it is possible to have supervised and unsupervised learning modes depending on the domain and data being tested.

Often outliers are discarded because of their effect on the complete distribution and statistical analysis of the database. This is a very good way if outliers appear due to some kind of error (measurement error, data corruption, etc.), but usually the source of the outliers is not clear. There are many cases where certain 'extreme' events from time to time are outliers from outside the normal distribution of the database, but it is a valid measure and not due to error. In these cases, the choice of how to deal with outliers is unclear, and the choice has a significant impact on the results of any statistical analysis done on the database. However, their size used to calculate point output is not gradually updated, and many techniques include various outliers data scans that make it impossible to manage the data stream. For example, [2] [3] use the Sparsity Coefficient to calculate the sparsity of data, and this is based on the technique of deep data separation that should always be cut from data distribution. This will be expensive and such updates will require a lot of data scanning. [4-6] use data sparsity metrics that include the concept of the nearest k (k-NN) neighbors to calculate the distance. This is not a valid method for data streams and as a single data test is not sufficient to store k-NN data for data points. On the other hand, the approaches to getting the previous point in the data streams [7] depend on the full-data space to find the vendors and thus the identified vendors cannot be determined by these methods. Therefore, it is interesting to suggest various new strategies that clearly explain the disadvantages of these existing methods.

Problem Statement

Outlier detection is a notoriously hard task: detecting anomalies can be difficult when it comes to clusters, and these clusters should be large enough to build a reliable model. The problem of contamination, i.e., using input databases contaminated by outliers, makes this task even more difficult as it incorrectly undermines the final model if the training algorithm is not robust.

These problems, which are present in many real-world sets, are not always addressed in activities that describe new unsupervised methods because these algorithms may point to a different application case. These factors trigger the need for a complete benchmark that combines different strategies for complex data sets. The main objective is to see a set of features that greatly balance novel data without reducing the effect of classification. Other problems stem from the computational costs of feature reduction algorithms and the large amount of data upcoming that requires the development of effective mitigation strategies that can be achieved simultaneously. Finding an outlier point, various algorithms based on mathematical modeling processes can be any of them, be it predictive or direct. Predictive techniques using tagged data using training sets to produce a data acquisition model, i.e., which contains deductions by domain vendors that will be used to classify real data items. Bit direct techniques consist of deviation, proximity, statistical clustering, and density based techniques, refer to those in which labeled 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 there is a more integrated feature than predictive strategies, the direct methods are not limited as the detection does not depend on the predefined models.

Outlier Detection Approach

Today's various modern systems are able to generate and capture real-time data continuously. Various applications apply to these real-time data acquisition systems, status monitoring systems, and financial activity systems. It is one of the most challenging tasks in real data to properly find outliers objects from certain data streams. Standard outlier detection methods are no longer usable as they only work with statistical data sets and include multiple data scans to produce effective results. In data streams, outliers acquisition algorithms (e.g., [9]) need to process each data object within a fixed barrier and can meet the cost of analyzing all real data with separate data analysis. In finding the outlier point, there is a major research problem in fault-tolerance, anonymous detection, credit card fraud detection, medical diagnosis on various types of streaming data. Outliers point finding has uncharacteristic patterns in the data; they are obviously structures that do not reflect normal behavior. The wide range of findings for outliers points is summarized and the remaining segments of this research paper. In the next five sections, here they categorized to finding outlier point approach are as follows:

- ✓ Nearest neighbour outlier detection techniques
- ✓ Density outlier detection techniques
- ✓ Cluster outlier detection techniques
- ✓ Statistical approach outlier detection techniques
- ✓ Robust distance outlier detection techniques and
- ✓ Depth based outlier detection techniques.

Each outlier detection techniques are described as follows:

- **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 neighbour's points or that their neighborhood is sparse.

- **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.

- **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.

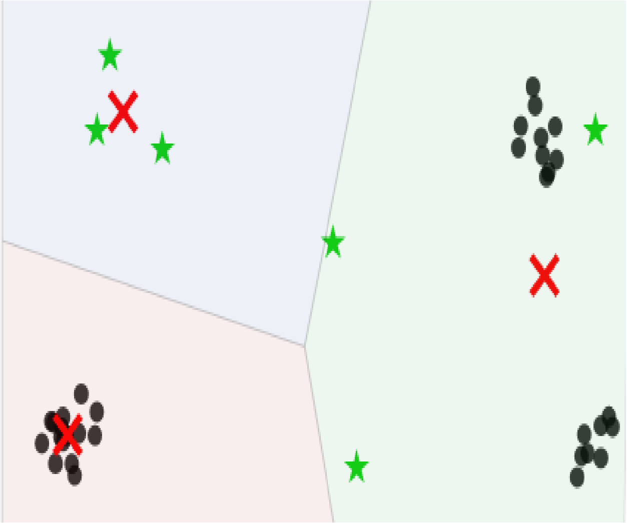

Figure: Clustering with outliers

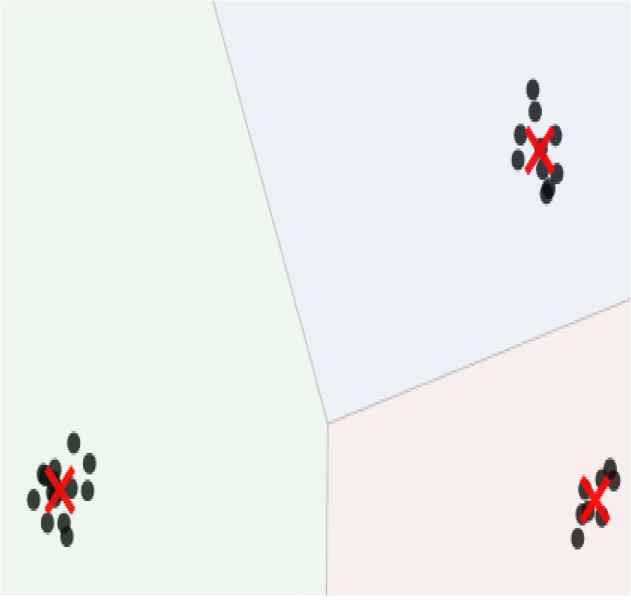

Figure: Clustering without outliers

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.

- **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 [10]. 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 [11] used the straightforward statistical to finding outlier point methods via casual box plots to locate particular outliers point in both univariate and multivariate case.

- **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.

- **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.

Literature Survey

A common problem identifying outliers is considered to be a variety of confidential as global versus local outlier models. The overall model shows the outlier model showing the binary decision methods that the data object provided to find the outlier point. The concept of outlier space gave a certain amount of relevance to each object. Such an “outlier” factor is the number that separates each item from the “how much” of the outlier factor. Many applications where it is attractive to rank the outliers in the database and get back the top-n outliers, a local outlier approach is apparently desirable. The supervised approach is supported by the setting of various data studies when determining the status of an outlier point, or not and differences between those abnormal categories of observation to be effective in different data areas are necessary [8].

Barbara et. al. [12] suggests a rigorous process for outliers in training data. The authors begin by distinguishing common items from outliers in training data, using frequent item-set mining, and then using a clustering-based process to obtain outliers. Second-thinking strategies can also be used in semi-supervised mode, where training data is combined with test data items compared to clusters to obtain points outside the test data object. If training data contains multiple class items, semi-supervised clustering can be used to improve clusters.

Zengyou He et. al. [13], called Discover CBLOF, provides outlier score known as the Cluster-based Local Outlier Factor (CBLOF) for each data point. The CBLOF measurement captures the size of the cluster in which the data object belongs, and the distance of the object to its cluster centriod. Initially they divided the data set to be set into clusters with a squeezer algorithm, and these clusters were divided into two clusters, namely large clusters and small clusters, based on the number of points in each collection. For each data point they count CBLOF and announce clusters.

Bay and Schwabacher et. al. [14] have shown that with enough detailed information, a simple pruning step can lead to the usual confusion of nearby neighborhood searches that will descend almost to the line. After calculating the nearest neighbors by a data point, the algorithm sets the outer limit of any data point to the weak outlier points obtained so far. Using this pruning process, the process discards nearby items, which is why it is unpopular. The performance of this algorithm is largely based on three assumptions, the violation of which could lead to malicious operation.

In [15], a feature bagging approach for finding outliers is suggested. It combines results from many outlier detection algorithms that are implemented using various features. The outlier detection algorithm uses a small set of randomly selected elements from the original feature set. As a result, each outlier detector identifies different outliers, and provides all data objects with outlier scores associated with the potential of becoming outliers. The outlier scores calculated by the individual outlier detection algorithms are then combined with the purpose of finding the better-quality outliers.

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.

In contrast, the cluster-based local outlier factor (CBLOF) [17] uses clustering in order to determine dense areas in the data and performs density estimation for each cluster afterwards. In theory, the clustering algorithm can be used to combine data in the first step. However, in practice, k-methods are often used to take advantage of low computational complexity, which is comparable to the quadratic difficulty of nearest-neighbor. After clustering, CBLOF uses the heuristic to divide the emerging clusters into larger and smaller clusters. Finally, the anomaly points are calculated by the distance each time it travels to its center multiplied by the conditions of its collection. For small groups, the closest large distance range is used. The process of using the number of group members as a measurement factor should measure the size of the collection area as stated by the authors.

We showed in previous work that this assumption is not true [18] and might even result in a incorrect density estimation. Therefore, we additionally evaluate a modified version of CBLOF which simply neglects the weighting, referred to as unweighted-CBLOF (uCBLOF). The results of uCBLOF using a simple two-dimensional dataset where the color corresponds to the clustering result of the preceding k-means clustering algorithm. Similar to the nearest-neighbor based algorithms, the number of initial clusters k is also a critical parameter. Here, we follow the same strategy as for the nearest-neighbor based algorithms and evaluate many different k values. Furthermore, k-means clustering is a non-deterministic algorithm and thus the resulting anomaly scores can be different on multiple runs. To this end we follow a common strategy, which is to apply k-means many times on the data and pick the most stable result. However, clustering-based anomaly detection algorithms are very sensitive to the parameter k, since adding just a single additional centroid might lead to a very different outcome.

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 lifr 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.

Conclusion

The purpose of this review is to introduce a structured and complete state of the art on outlier detection techniques in high dimensional data, and we try to extract the essence of the advanced concept of outlier, focusing on acquisition algorithms provided by various authors. It is generally high-dimensional, which creates a curse of size problem in getting an outlier point and is an important part of machine learning with multiple domains of critical applications, namely medical diagnostics, fraudulent detection, and intrusion detection. Due to the large number of data objects in real-life performance realization, it faces various challenges in obtaining these outlier points, so as we reduce the size appropriately, they increase the data size of the object or combine traditional algorithms to create robust estimates of high dimensional data and low dimensional data. High-dimensional data can be seen as part of a variety of big data challenges.

To understanding a new challenges in high-dimensional data and comprehension data. The volume of data increases but also the high-dimensional data; a large collection of low-dimensional sensors can be seen as a multivariate time series. Due to the fast calculation time, especially for large data sets on high-dimensional multivariate data, we highly recommend trying it on larger databases when looking for global outlier detection.

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] K. Das and J. Schneider, "Detecting anomalous records in categorical datasets," in *Proceedings of the ACM KDD*, pp. 220–229, 2007.

[10] Grubbs, F.E., 1969. Procedures for detecting outlying observations in samples. Technometrics, 11: 1-21.

[11] 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.

[12] Daniel Barbar´a, Yi Li, Julia Couto, Jia-Ling Lin, and Sushil Jajodia. Bootstrapping a data mining intrusion detection system. In SAC '03: Proceedings of the 2003 ACM.symposium on Applied computing, pages 421–425, New York, NY, USA, 2003. ACM.

[13] Zengyou He, Xiaofei Xu, and Shengchun Deng. Discovering cluster-based local outliers. Pattern Recognition Letters, 24(9-10):1641–1650, 2003.

[14] Stephen D. Bay and Mark Schwabacher. Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD 03), pages 29–38, New York, NY, USA, 2003. ACM.

[15] Aleksandar Lazarevic and Vipin Kumar. Feature bagging for outlier detection. In Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, KDD ’05, pages 157–166, New York, NY, USA, 2005. ACM.

[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] He Z, Xu X, Deng S. Discovering Cluster-based Local Outliers. Pattern Recognition Letters. 2003;24(9–10):1641–1650.

[18] Amer M, Goldstein M. Nearest-Neighbor and Clustering based Anomaly Detection Algorithms for RapidMiner. In: Simon Fischer IM, editor. Proceedings of the 3rd RapidMiner Community Meeting and Conferernce (RCOMM 2012). Shaker Verlag GmbH; 2012. p. 1–12.

[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.