


ITC 1/50 Information Technology and Control Vol. 50 / No. 1 / 2021 pp. 138-152 DOI 10.5755/j01.itc.50.1.25588	A Novel Density-based Technique for Outlier Detection of High Dimensional Data Utilizing Full Feature Space	
	Received 2019/03/29	Accepted after revision 2021/02/17
	 http://dx.doi.org/10.5755/j01.itc.50.1.25588	

HOW TO CITE: Rehman, M. U., Khan, D. M., Saher, N., Shahzad, F. (2021). A Novel Density-based Technique for Outlier Detection of High Dimensional Data Utilizing Full Feature Space. *Information Technology and Control*, 50(1), 138-152. <https://doi.org/10.5755/j01.itc.50.1.25588>

A Novel Density-based Technique for Outlier Detection of High Dimensional Data Utilizing Full Feature Space

Mujeeb Ur Rehman

Department of Information Technology (DIT), the Islamia University of Bahawalpur, Pakistan;
Department of Computer Science, Khwaja Fareed University of Eng. and IT (KFUEIT),
Rahim Yar Khan, Pakistan; phone: +92 68 5882400; fax: +92 68 5882400; e-mail: mujeeb.rehman@kfueit.edu.pk

Dost Muhammad Khan

Department of Information Technology (DIT), the Islamia University of Bahawalpur, Pakistan;
e-mail: khan.dostkhan@iub.edu.pk

Najia Saher

Department of Information Technology (DIT), the Islamia University of Bahawalpur, Pakistan;
e-mail: najiasaher@gmail.com

Faisal Shahzad

Department of Information Technology (DIT), the Islamia University of Bahawalpur, Pakistan;
e-mail: faisalsd@gmail.com

Corresponding author: mujeeb.rehman@kfueit.edu.pk

Recently, anomaly detection has acquired a realistic response from data mining scientists as a graph of its reputation has increased smoothly in various practical domains like product marketing, fraud detection and so many other fields. High dimensional data subjected to outlier detection poses exceptional challenges for data mining experts and it is because of natural problems of the curse of dimensionality and resemblance of distant and adjoining points. Customary methodologies concentrate largely on low dimensional data and hence show ineffectiveness while discovering anomalies in a data set comprised of a high number of dimensions. It becomes a very difficult and tiresome job to dig out anomalies present in high dimensional data set when all subsets of projections need to be explored. All data points in high dimensional data behave like similar observations because of its intrinsic feature i.e., the distance between observations approaches to zero as the number of dimensions extends towards infinity. This research work proposes a novel technique that explores deviation among all data points and embeds its findings inside well established density-based techniques. This is a state of art technique as it gives a new breadth of research towards resolving inherent problems of high dimensional data where outliers reside within clusters having different densities. The datasets from UCI Machine Learning Repository are chosen to test the proposed technique and then its results are compared with that of density-based techniques to evaluate its efficiency.

KEYWORDS: Anomaly Detection, Local Neighborhood-based Anomaly Detection, Projected Outlier Local Outlier, High Dimensional Data.

1. Introduction

An outlier could be differentiated from an inlier in such a way that it could be considered a very different observation that might demonstrate very beneficial for some individual or organization. Outlier and noise are two very different entities as the only former one is wanted. Several benefits are enjoyed in practical fields by separating regular data from unexpected data. These irregular forms are also acknowledged as aberration, anomaly, contaminant, discordant observation and exception in many different application fields [9]. A very precise characterization of anomaly could be described as; it is a point that behaves relatively in a different way from other points concerning some characteristics. Density-based anomaly detection generates two kinds of data points, either inlier or outlier as shown in Fig. 1. Inlier is a data point that is surrounded densely by its neighboring points whereas an outlier has relatively fewer neighbor points and hence behaves like an abnormal entity. Several issues need to be considered while detecting anomalies amongst a particular class of data set. These issues require to preprocess certain questions as suggested by Ranga Suri et al. [38], e.g. what method to choose? (either distance-based or density-based), what type of data is? (either numerical or categorical), what is the mode of analysis? (either online or offline). Local neighborhood-based anomaly

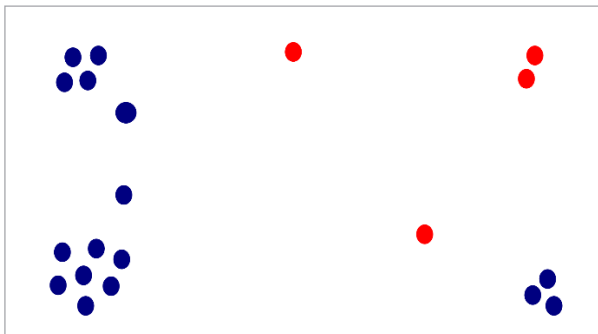
detection reveals that regular data points occupy the condensed neighborhood, from the other perspective, anomalies are far away from their neighbors, that is., these irregular points inhabit the less condensed neighborhood. Anomaly detection for low dimensional data is processed exhausting conventional procedures which turn into vastly hostile in the perspective of high dimensional data [1]. High dimensional data reveals its inherent problem which shows that the average outcome of all dimensions creates anomalies indistinguishable inside data points. LSOF proves very efficient method while detecting outliers from high dimensional data as it reduces variance among neighboring data points [2]. This problem needs to be engaged in so that anomalies could be made distinguishable. It is observed that low-dimensional projections (spaces comprising a subset of attributes) contain tremendously bulk of anomalies hidden inside high-dimensional data streams [28]. High dimensional subspaces recognize these anomalies as projected anomalies, that is, one anomaly present in one projection might behave normally in another projection [21].

High dimensional data has been employed recently in many different practical fields; it includes recommendation systems, stock exchanges, medical data, electronic vendors and unstructured data [32]. Concrete

data and Ionosphere data proves to be a good example of high dimensional data and could be exploited for data digging purposes.

Two major problems are observed regarding anomaly detection for high dimensional data. The specificity of likenesses between data points weakens when the number of dimensions exceeds some limits. A study in [14] demonstrates that, with the propagation of dimensionality, the Euclidean distance between the adjoining neighbor and that to the furthestmost point shrinks and causes a reduction in the gap between these two extreme points.

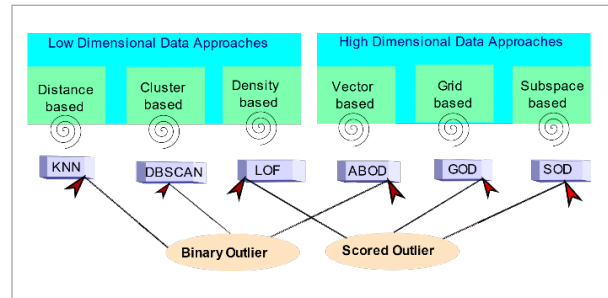
Figure 1
Outlier vs Inlier (Density-based anomaly detection)



The complexity of anomaly detection algorithms suffers from the curse of dimensionality, that is; its complexity rises exponentially as dimensionality grows unbounded. When the number of attributes exceeds some limit then a typical anomaly detection algorithm behaves inflexible and unreliable. Therefore, these algorithms become inappropriate and unsuitable when deployed in practical domains [33, 47]. Fig. 2 demonstrates a summarized picture of outlier detection techniques regarding low and high dimensional data. It further classifies algorithms in two broad categories on the basis of output, either binary or score. Low dimensional data approaches comprise distance, density, and cluster-based techniques. Vector, subspace, and grid-based techniques are devised and experimented on high dimensional datasets. Density-based Spatial Clustering of Applications (DBSCAN) identifies anomalies as noise. It is a local neighborhood-based methodology that makes groups of data points having random shapes. Its mechanism is based on two elementary ideas which are density connect-ability and density reachability. Its opera-

tion is concerned with the minimum number of data points and the size of the epsilon neighborhood [42].

Figure 2
An Overview of Anomaly Detection Techniques



Another local neighborhood-based methodology known as Local Outlier Factor (LOF) has attracted the attention of researchers as it discovers scored outliers. The core idea behind its operation is that the local density of a certain data point is compared and matched with the local density of its neighbor points. A user selects parameter 'k' which determines the number of neighbors to be processed. Many variants have been proposed to improve the efficiency of the LOF algorithm.

Local Correlation Integral (LOCI) has been acknowledged as a comprehensive anomaly detection technique. Its specialty is that it discovers lonely anomalies along with assembly or group of anomalies. Earlier techniques demand users to choose cutoffs so that a data point could be decided either a normal point or anomaly whereas LOCI determines automatic cutoff and hence gives relief to its users. Another special feature revealed by this methodology is that a point to be observed captures an abundance of information in the vicinity of that point. That is, micro-clusters, macro-clusters, their diameters, and inter-cluster distances are determined through this technique. Optimized results are expected when LOCI is studied and analyzed while tackling inherent problems of high dimensional data. Techniques devised for low dimensional data work efficiently when the number of dimensions is a few. Six to fifteen dimensions are very common in low dimensional data (e.g., Breast Cancer Dataset present on UCI Machine Learning Laboratory Repository), hence the distance between points could be easily differentiated through any normal distance measuring method, e.g., Euclid-

ean metric method. Whereas high dimensional data contains a relatively high number of dimensions (e.g., Sales Transactions Weekly Dataset present on UCI Machine Learning Laboratory Repository), that is; from fifty to hundred or from hundred to thousand dimensions. Normal distance metric methods fail to distinguish between outliers and inliers as all observations seem equally distant from one another and it happens because of the inherent feature of high dimensional data. Since traditional techniques utilize normal distance measuring methods, hence these fail altogether to detect anomalies present in high dimensional data. Dutta and Banerjee [10] have proved that traditional outlier detection techniques show failure when the dimension size of data exceeds 250.

This research work is arranged as follows: A local neighborhood-based approach is proposed for exploring anomalies in a dataset having a high number of dimensions. The proposed approach exploits the benefit of some existing techniques, i.e., Distributed LOF [44], INFLO [18], COF [40] and LoOP (which is statistical technique) [25]. The rest of the research work is depicted as follows: In Section 2, we have discussed research motivation, questions and research objectives. Section 3 and 4 elaborate related work and proposed methodology respectively. Experimental work with results is discussed in Section 5. Limitation of the proposed technique is presented briefly in Section 6. Finally, Section 7 concludes the paper.

2. Problem Description

2.1. Research Motivation

During the past few years, low dimensional data is being interchanged with high dimensional data because of speedy advancements in technology. So it is a very essential and demanding situation to invent such systems and algorithms which can challenge and resolve high dimensional data problems. Generation of big data and large data sets have motivated many scientists to redesign algorithms and techniques regarding anomaly detection in high dimensional data. When we deal with real data or real problems, we often deal with high dimensional data that consists of dozens of dimensions. For data miners, finding anomalies within multiples of dimensions becomes not an easy job.

Though it is very common to tackle such situations with dimensionality reduction techniques like PCA (Principal Component Analysis), yet many datasets necessitate considering all dimensions equally relevant. Subspace based techniques like SOD (Subspace based Outlier Detection) are considered suitable but suffer from the curse of dimensionality. Proposed work focusses on full feature spaces of datasets to resolve the issue of least difference in data points when dimensionality grows to remarkable volume. Further it also bears fruits of authentic outlier detection techniques like INFLO (influenced outlierness) technique which detects clusters of different densities residing near to one another.

2.2. Problem Statement

Exploring anomalies from high dimensional data through a subset of feature spaces is costly in terms of time and accuracy as digging out subspaces itself is a time-consuming job. Conventional methodologies cannot detect anomalies from high dimensional data due to the specificity of resemblances between data points but these methodologies could be adapted to tackle the above-described problem. In our case, outcomes are estimated to be more precise and computationally less costly as compared to outcomes obtained through a subset of feature subspaces. The exploitation of subspaces or subsets of features resolves likeness of similar data points in a dataset having a large number of dimensions, hence this approach has been utilized in many subspace based outlier detection techniques. Only the brute force technique guarantees cent percent accuracy while trying all combinations of different subspaces but it is not feasible in reality. Evolutionary techniques like the Genetic Algorithm handles time complexity efficiently but generates optimized results with each next iteration. Conventional density-based outlier detection technique LOF and its variants INFLO and COF are considered state of art techniques while projecting on low dimensional data only. Since these techniques do not utilize a subspace-based approach, so its adaptation for high dimensional data assures better results in terms of time complexity, optimization and memory required to process data.

2.3. Research Questions

An analysis needs to be conducted on local neighborhood-based anomaly detection algorithms by revising the variance of attributes for high dimensional data

set. The following are major research questions that will be explored and answered inevitably.

- 1 The likeness of data points regarding high dimensional data needs to respond more intelligently [14]. All data points resemble each other concerning the distance between them. We are to discover whether it is possible to maximize the difference in the distance among data points? Another research area in this regard is to find or improve distance measuring methods to maximize the distance between data elements. For example, the Manhattan Distance metric determines more distance-variation in data points as compared to Euclidean distance, and hence it is suggested to utilize it in high dimensional data sets [11].
- 2 Curse of dimensionality makes projected subspaces based outlier detection infeasible for high dimensional data sets. A valid question enquires to check possibility of replacing Projected Subspaces based techniques with full space-based techniques?
- 3 Traditional techniques work on full feature spaces of low dimensional data. These methods fail regarding high dimensional data as outliers are supposed to lie in projected features [30]. A question arises whether traditional techniques devised for full feature subspaces should be adapted (improved) or new techniques (in terms of approaches like vector-based, subspace-based) should be discovered.

3. Related Work

Hawkins defined “Outlier Detection” which is accepted globally, that is, “An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism [13]. There are many well-known applications of outlier detection like credit card fraud detection, intrusion detection, fault detection, etc. [4]. In a broad sense, there are two classes of outlier detection methods, that is; either supervised or unsupervised, and its choice is dependent upon the nature of data being processed. Major categories of outlier detection are distance-based, density-based and subspace-based. While comparing the above three categories, distance-based approaches pro-

duce binary outliers, density-based methods generate scored outliers and subspace-based techniques create both kinds of outliers, binary and scored. In paper [46], a local density estimator (variable sample technique) is implemented by using the T-Forrest algorithm. It splits the data into subspaces and finally density of each instance determines score (outlierness).

K-means algorithm is modified to cope with high dimensional problems by introducing multiple centroids and local search strategy during the iterative process [17, 23, 41].

Jindi and Huaji [19] used a method of density clustering to examine the outliers of data by observing the cow’s behavioral characteristics. Density-based clustering can detect local outliers whereas distance-based methods cannot find out. DBSCAN, a well-known clustering method, explores clusters of arbitrary shapes and detects outliers as noise points.

ABOD (Angle Based Outlier Detection) proves to be very efficient for outlier detection of high dimensional data as it is not sensitive to the curse of dimensionality [26]. Broad angles reveal that normal data points exist among clusters whereas a small angle indicates that these observations could be suspected as outliers.

Yan et al. [44] presents a new strategy to obtain optimization for the LOF algorithm. It not only improves costs within each stage but also decreases communication costs for each stage. LOF requires finding k-distance, local reachable density (LRD) and local outlier factor.

Regression analysis in high dimensional data requires careful investigation to avoid some statistical issues like misspecification of the model and inappropriate predictions [7]. Wang et al. [43] proposes a multiple outliers detection approach through multiple testing procedures. To enhance effectiveness, a relatively reliable normal subset of points is obtained by refining outlier detection rule.

Yuan et al. [45] proposed the neighbor-density-deviation-based outlier factor (NDDOF) algorithm which can detect outliers amongst different density clusters. Further it can detect outliers within objects having relatively smaller clusters.

Liu et al. [29] introduce a trajectory outlier detection

algorithm (TRAOD) which proves bad for local trajectory, so Lee et al. [27] compensated it by proposing another density-based trajectory outlier detection algorithm (DBTOD).

Jindi and Hua [19] suggest not ignoring global neighborhood while focusing on local neighbors and detect density-based outliers. Its basic purpose is to determine the degree of an outlier compared with other outliers globally and hence improves its rank or score.

Kriegel et al. [25] proposed a very effective search strategy for finding outliers in relevant subspaces, a set of attributes. This strategy is applied to spatial data containing spatial attributes. LoOP: local outlier probabilities, is an algorithm that is very similar to LOF (Local Outlier Factor) except that it does not provide an outlier factor. Rather it utilizes probabilistic set distance to measure the probability of a data point being an outlier.

HiCS: High contrast subspaces resemble subspace outlier detection but its main distinction is that it explores high contrast subspaces which have more probability of holding outliers hidden in subsets of attributes [22]. It detects outliers from the dataset by using LOF but other similar methods can also be utilized.

OutRank, outlier ranking is an algorithm that focuses on finding rank or score of data points. It measures outlierness of data points by analyzing subspaces. Regarding this, it exploits the similarity of subspace measurements and subspace clustering methods [31].

Projected Clustering based on K-Means (PCKA) is a partitioned distance-based clustering algorithm. PCKA is suitable for relevancy analysis of a set of dimensions called subspaces but it lacks redundancy analysis. Proposed PCKA is in improved form as it not only performs relevancy analysis but also redundancy analysis [8].

In paper [6], it is discovered that LOCI is a versatile technique that explores wealth of information by detecting clusters within clusters, but it proves computationally expensive. Feature extraction methods help to reduce redundant and irrelevant data and ultimately help in enhancing the speed of selected techniques [3]. Dimensionality reduction (DM) has been recognized as a good technique to diminish time

complexity but it's not valid when all dimensions are significant [2, 16].

Anomaly detection is often considered an mandatory tool in exploratory data analysis (EDA). The scientists have found the principal component analysis (PCA) as one of the most popular method for EDA with high-dimensional data. In particular, two-dimensional projections with a few leading PC directions have been found beneficial for detecting hidden patterns in HD data [5, 12]. Nevertheless, it is pre-determined that the estimation of PCA for high-dimensional data is often erratic, so the sample version of PCA may not discover anomalies residing in some population PC directions that are not realistically projected [20, 34]. Also, since the PC projection plot can only show two directions at a time, it may fail to reveal anomalies that are well concealed in a subspace generated by several PC directions.

Outlier detection regarding distance-based approaches [24] accepts two parameters radius ϵ and a percentage π , where π percent of all other points must have a distance from point p less than ϵ . kNN distance models are used to determine labeled outliers where k and ϵ parameters determine whether a data point is normal or outlier [35]. A variant of LOF known as COF (Connectivity based outlier) was proposed by Tang et al. [40] which solves the problem of low density and isolation. Previously LOF was unable to differentiate between low density and isolation of data points. There is another variation of LOF, i.e., INFLO (Influenced Outlier) [18] which solves those problems in which clusters of different densities could not be separated clearly. It solves this problem by taking the ratio of the average density of things in the vicinity of a point.

Grid-based subspace outlier detection (GOD) [30] partitions data space into an equal depth grid (number of cells in each cell). After calculating the sparsity coefficient of k dimensional grid cells, a negative sparsity coefficient of data points residing in lower-dimensional cells marks these as outliers.

Rehman and Khan [36] have done extensive effort on evaluating different proximity functions when applied to density-based techniques. Results are analyzed and compared in terms of outlier score, inlier score, time complexity, dimensionality variation and for different values of k (minimum points).

A novel method, LSC (Local Subspace Classifier) is used in [15] that is based on the feature vector extraction method. LSC determines outlier measure based on time increment for distance applied on the model. This method was improved in terms of computation in [37] by proposing method Fast LSC. In this approach, clustering is used to reduce the amount of data and hence proves ten times faster as compared to the LSC method.

Tang and He [39] detect outliers based on distance function utilizing a density-based approach. He uses three types of measures to determine density estimation which are classic k nearest neighbors, reverse nearest neighbors and shared nearest neighbors.

A comprehensive and precise comparison shown in Table 1 reveals approaches to be adopted, the type of outliers to be detected and the pros and cons of methodology to be utilized.

Table 1

Comparison of Approaches for Outlier Detection Techniques

Work	Method	Approach	Outlier Type	Data Dimension	Special Feature	Shortcomings
Jin et al. [18]	INFLO	Density-based	Scored	Low	Clusters of different densities	High dimensions are not considered
Chandola et al. [6]	LOCI	Density-Based	Scored	Low	Clusters having wealth of information	Computationally expensive
Kriegel et al. [26]	ABOD	Vector-based	Binary	High	Angles are more stable than the distance	Strength of outliers is not measured
Aggarwal and Yu [1]	Evolutionary Technique	Grid based	Binary	High	Sparsity Coefficient is used	Quality depends on grid resolution
Kriegel et al. [25]	Statistical Technique	Subspace based	Scored	High	KNN and Subspace are computed	expensive
Zhang and Yin [46]	Sample Technique	Density-based	Scored	Low	Tree height is minimized.	Results get influenced due to randomness
Keller et al. [22]	HiCS	Subspace based	Scored	High	High contrast subspaces	Computationally expensive
Agrawal [2]	LSOF	Subspace based	Scored	High	Variance among neighbors is reduced	Parameter β is not determined automatically
Yan et al. [44]	Distributed LOF	Density-based	Scored	Low	Improves cost of each stage	Not suitable for high dimensions
Dighe and Gawde [8]	PCKA	Distance-based	Binary	Low	Performs relevancy analysis	Lacks redundancy analysis
Liu et al. [29]	DBTOD	Subspace based	Binary	High	Trajectory of anomalies is detected.	Not suitable for spatial and temporal data

4. Proposed Methodology

4.1. Problem Statement

Let a DB contains a “ d ” number of dimensions and ‘ N ’ number of data points. Let ‘ D ’ denotes set of dimension and represented by $D = \{m_p, m_q, \dots, m_d\}$ and ‘ P ’ represents set of data points where $P = \{p_p, p_q, \dots, p_n\}$. The given task is to find the distance between any two points (D_m, P_n) of DB, i.e., n^{th} data point having m^{th} dimension. Standard deviation is determined by calculating the variance of each attribute set for the same dimension (attribute ‘ p_i ’ of dimension “ d_k ”). Attributes having larger variance are normalized. Local outlier factor is determined by fixing parameter ‘ k ’ (number of neighbors to be processed), which finally gives scores of each point that could be evaluated and compared with that of traditional techniques (e.g. INFLO) and subspace-based techniques (e.g. SOD).

4.2. Variance Calculation

In this step, each attribute is examined and gets classified according to variance present in data. Then all attributes having low standard deviation are normalized as per classified. Standard Deviation is determined after calculating the variance of each attribute. Standard Deviation for Total Population is determined by using Equation 1.

$$\sqrt{\frac{1}{N} \sum_{i=1}^n (x_i - \mu)^2}. \quad (1)$$

Standard Deviation for Sample Population is determined by using equation 2.

$$\sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \mu)^2}, \quad (2)$$

where N, x_i, μ stand for the number of points, i^{th} data point and mean-value of all data points, respectively.

4.3. Finding Outlier Degree of each Data Point

k -distance of a data point ‘ o ’, “ $dis_k(o)$ ” is determined where parameter ‘ k ’ (minimum points) is chosen by the user. In the next step, the k -distance neighborhood of a tuple “ $Neigh_k(o)$ ” is calculated. Reachability distance of a neighboring point ‘ p ’ with respect to ‘ o ’ is measured using equation 3.

$$Rdis_k(p,o)=\text{maximum}(\text{disk}(o),\text{disk}(p,o)). \quad (3)$$

After finding $Rdis_k(p, o)$, local reachability density of point ‘ o ’ “ $Lrd_k(o)$ ” is calculated via equation 4 where ‘ p ’ represents all neighboring points in the neighborhood of ‘ o ’. $Lrd_k(o)$ determines the density of each point in the neighborhood of the point of o after some value of k is chosen.

$$Lrdk(o) = \frac{Neighk(o)}{\sum_{o' \in Neighk(o)} Rdisk(o',o)}. \quad (4)$$

In the last step, we measure Local outlier factor “ $LOF_k(o)$ ” of point ‘ o ’ which gives the degree of outlierness determined by equation 5.

$$LOFk(o) = \frac{\sum_{o' \in Neighk(o)} \frac{Lrdk(o')}{Lrdk(o)}}{Neighk(o)}. \quad (5)$$

A higher value of LOF reveals a higher degree for outlierness of a data point whereas lower degree depicts that point as an inlier.

Proposed Methodology

Input: numerical data having “ d ” dimensions and ‘ N ’ records

Output: data points with a higher degree of outlierness (high LOF)

Step 1: Apply dimension reduction or search relevant attributes (if applicable)

Step 2: Examine the standard deviation of each attribute and identify those having lower values

Step 3: Normalize all attributes having lower values.

Step 4: Find k -distance of a tuple, $disk(o)$.

Step 5: Determine the k -distance neighborhood of a tuple, $Neighk(o)$.

Step 6: Find reachability distance of a tuple ‘ p ’ with respect to ‘ o ’.

Step 7: Determine the local reachability density of tuple, $Lrdk(o)$.

Step 8: Local outlier factor $LOFk(o)$ of instance ‘ o ’ is calculated.

4.4. Comparison of Outlierness with Different Perspectives

Top ten outliers are compared in terms of its strength (score) with traditional density-based and subspace-based techniques. All data points are assigned score to differentiate between outliers and inliers using RapidMiner and ELKI tools. Finally results are analysed and discussed to conclude the pros and cons of the proposed technique.

5. Experimental Work

As described earlier, similar distances are exhibited by data points when the number of dimensions grows large enough. Hence Local outlier factor of all data points exhibits a similar score for high dimensional data that shows the similarity of all points with respect to distance. As the value of Euclidean distance is smaller than that of Manhattan distance (also known as taxicab metric), so we get different results for local outlier factor outlier applied on the same dataset as shown in Table 2. We can see that the difference of LOF for Manhattan distance is higher than that of Euclidean distance. A Manhattan distance also known as Taxicab distance replaces Euclidean geometry with Taxicab geometry in which the distance between two data points is the sum of the absolute differences of their cartesian coordinates.

So it is obvious that Manhattan distance should be preferred for high dimensional data whenever some outlier detection technique is experimented.

In another experiment, different proximity functions are used to calculate the outlier score when the dimension size of the dataset is gradually increased. As mentioned before, theoretically distance between two data points approaches to zero as dimension size reaches infinity. Practically this distance is so small that it cannot differentiate between outlier (abnormal point) and inlier (normal point). Figure 3 clearly shows that the average outlier score of all data points declines fair enough as dimension size grows. Three different proximity functions, i.e. Euclidean, Manhattan and Squared Euclidean are compared to reveal the effect on outlier scores when the number of dimensions is changed in ascending order. Squared Euclidean distance proves effective

Figure 3

Effect of dimensionality on outlier score for different proximity functions

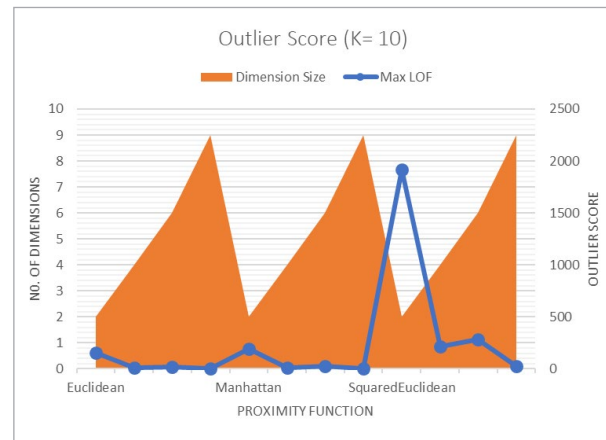


Table 2

Comparison of LOF for Euclidean and Manhattan Distance

Dataset	Euclidean Distance, k=2	Manhattan Distance, k=2
ID=1: 10.0 9.0 1.0 2.0 2.0 11.0	lof=1.019	lof=1.025
ID=2: 8.0 0.0 1.0 2.0 2.0 8.0	lof=0.981	lof=0.987
ID=3: 4.0 8.0 1.0 2.0 2.0 7.0	lof=1.019	lof=1.012
ID=4: 0.0 4.0 1.0 2.0 2.0 14.0	lof=1.098	lof=1.306
ID=5: 3.0 1.0 1.0 2.0 2.0 2.0	lof=0.981	lof=0.987
Five data points were chosen (randomly) from dataset to determine the LOF of each data point	Min-LOF=0.981 Max-LOF=1.098	Min-LOF=0.987 Max-LOF=1.306
	Difference-LOF=0.117	Difference-LOF=0.319

for density-based outlier detection techniques when dimension size is large enough.

In this experimental work, two unsupervised datasets named “Concrete Data” and “Appliances Energy Prediction Dataset” are collected from the UCI Machine Learning Laboratory. As a matter of the proposed technique, we determine the mean and standard deviation of each attribute for these datasets. Standard deviation is used to determine variance or spread out present in all attributes. Attributes having lower standard deviation are selected for the normalization process. Attributes showing large variance contribute more for any proximity function and hence require no normalization.

Algorithms of the same class are those algorithms which work on the principle of local density. These algorithms are also known as variants of LOF, which are COF, INFLO and LOOP. Each algorithm calculates the outlier score of each point with respect to the local density of its neighboring data points. In Figure 4 (a) and 4 (b), we evaluate the strength of outliers by comparing the proposed technique with others of the same class. We compare these algorithms in terms of maximum score (max score), minimum score (min score), number of outliers and number of inliers. When experimented on Concrete Dataset shown in Figure 4 (a), for the proposed methodology, outlier scores (max score is 4.0 and min score is 0.93) is relatively higher than that of others, whereas the number of outliers (889) gets better strength in COF more than that of proposed. It is because, COF

implementation is based on the connectivity of all data points, hence it finds several outliers in a more precise way. But when we compare run time of COF and proposed methodology, then the proposed one proves better regarding time complexity. Figure 4 (b) exhibits experimentation on Energy Dataset and shows almost similar results as shown for the first dataset. The maximum outlier score is higher than that of other techniques of the same class. There is one exception that the number of outliers for INFLO differ in both experiments, for the reason that all techniques treat its neighboring points in a slightly different way.

ABOD and SOD are considered reliable outlier detection techniques regarding high dimensional data. These two techniques utilize different approaches as the former one is vector based and calculates outlier scores based on the deviation of angle of a certain point with respect to other data points. The second technique is subspace-based, i.e. different subset of attributes are used to find appropriate subspace that holds outliers embedded in it. Figure 5 (a) and 5 (b) show a comparison of the proposed technique with angle based and subspace-based techniques. Figure 5 (a) reveals that the outlier score of the proposed technique is better than that of ABOD, whereas SOD behaves better in terms of outlier score and number of outliers as well, it is because of finding suitable subspaces that contain distinct outliers. As far as the time complexity of SOD is concerned, it does not defeat the proposed technique. In the second experiment

Figure 4 (a), (b)

Comparing Proposed technique with others of the same and different classes: a) Concrete Data b) Energy Data

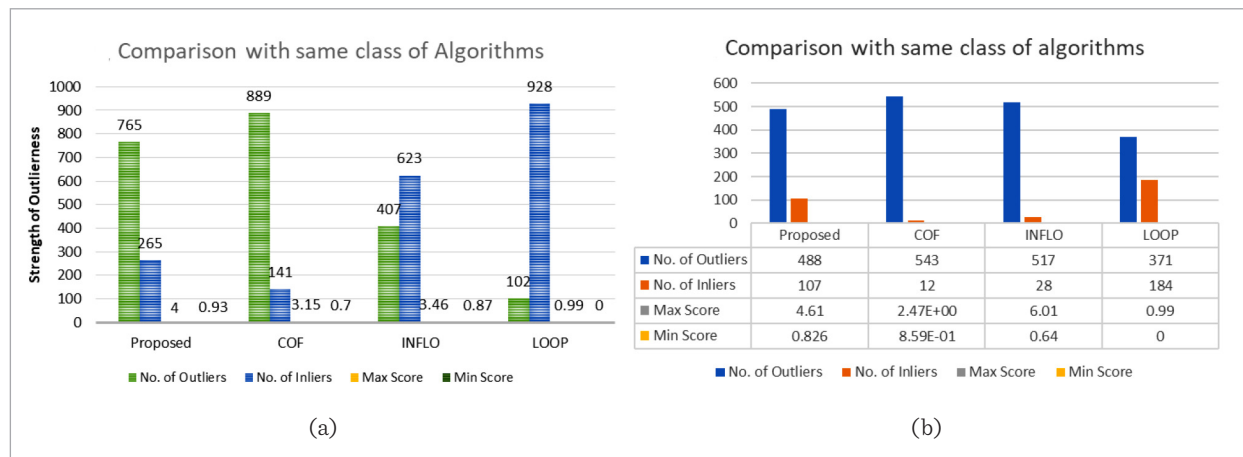
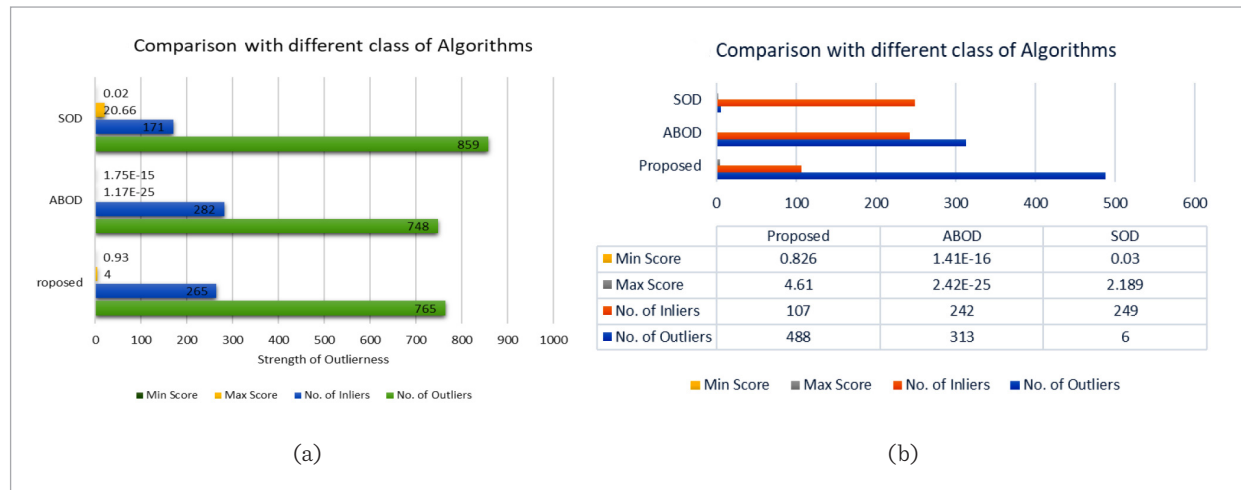


Figure 5 (a), (b)

Comparing Proposed technique with others of different class: a) Concrete Data b) Energy Data



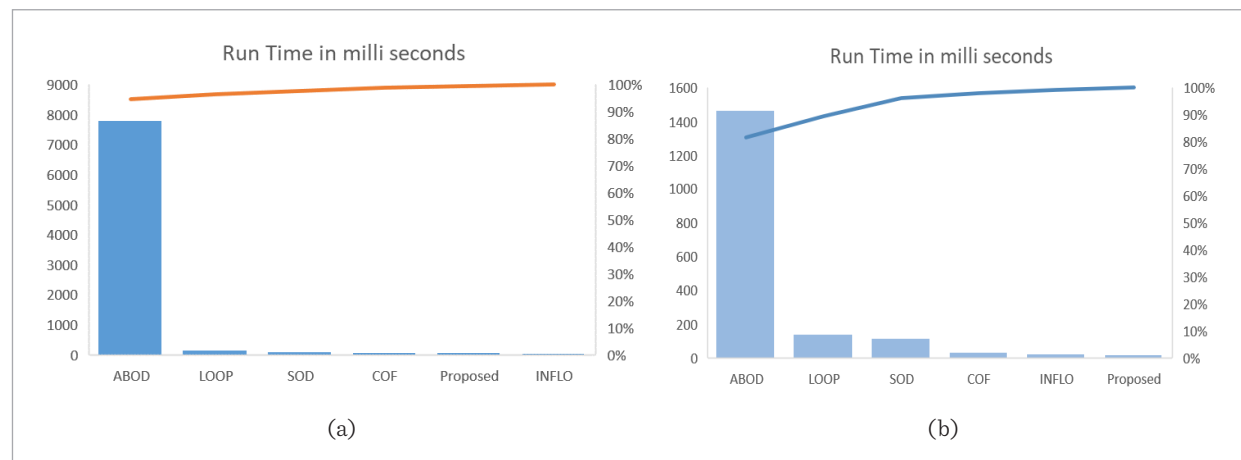
shown in Figure 5 (b), we observe results as expected in terms of outlier score and number of outliers when compared with other class of algorithms i.e. SOD and ABOD.

Time complexity has more concerns for any algorithm/technique when the dimension size of data is large enough. We have already discussed that curse of dimensionality causes an exponential rise in run time as the number of dimensions grows. In this research work, we have compared the time complexity of the

proposed technique with techniques of the same class/approach and of different as well. Figures 6 (a) reveals that the runtime of the proposed methodology is less than other techniques for the same class and different classes as well. Only INFO shows better results but its outlier strength is less than that of proposed technique. In fact there exists a tradeoff between accuracy and runtime while comparing with techniques of the same and different classes. Figure 6 (b) also verifies the above claim that the runtime of the proposed tech-

Figure 6 (a), (b)

Runtime (milliseconds) comparison with other techniques: : a) Concrete Data b) Energy Data



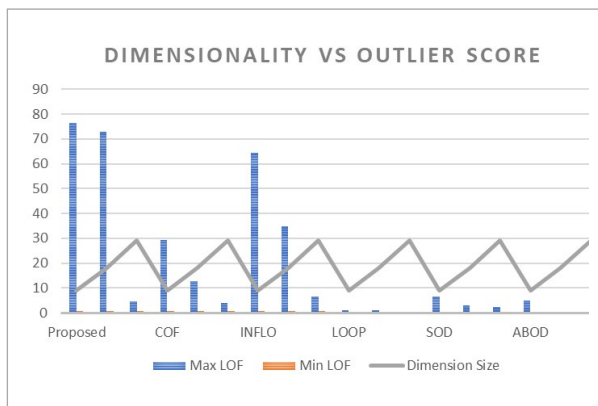
nique is very efficient as compared with techniques of the same class and different classes as well

It is a well-established fact that a true relationship exists between the number of dimensions and outlier scores. We have described before that when dimension size is smaller then there is no need to worry as traditional techniques work effectively and efficiently. A high number of dimensions, i.e., high dimensional data requires proper selection of proximity function as the distance between any two data points should be visible in terms of its difference. The outlier score of each data point is directly proportional to the dis-

tance between that point and its neighboring points. As a matter of proposed technique, we have determined variance amongst each attribute. All attributes having the least spread-out are normalized so that these attributes should not compromise the effect of attributes having large standard deviation values. Figure 7 demonstrates how the outlier score behaves when the size of the dimension is increased when applied on different outlier detection techniques. It is obvious for all methods that the outlier score is inversely proportional to dimension size. When compared proposed novel density-based techniques with others, we see that the outlier score has higher values for all dimensions relatively.

Figure 7

Effect of Dimensionality on Outlier Factor: Energy Data

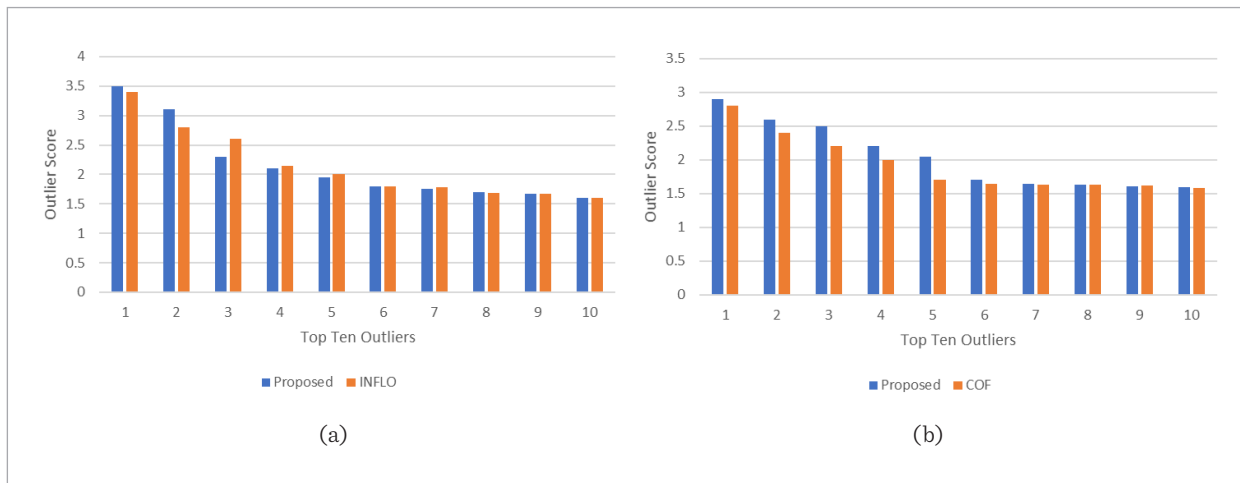


Data miners show more interest in data points having the highest scores as these points are likely to contain information that might prove treasure for any organization or company.

Top ten outliers for proposed and traditional techniques are compared as shown in Figure 8 (a) and 8 (b). The proposed technique reveals outstanding results for the COF algorithm, whereas the INFLO algorithm behaves slightly weak for the third and fourth outliers. Above all, there is an average improvement of outlier scores for the proposed technique when compared with that of traditional local density-based techniques.

Figure 8 (a), (b)

A Comparison of Top ten Outliers (Outlier Factor) for Proposed and INFLO/COF



6. Limitation

The above-proposed technique works on numerical or continuous data only but it could be adapted for other data types if the distance between data points is quantifiable. For example, the edit distance metric calculates the distance between words containing alphabetical letters.

7. Conclusion

During last decade, scientists have recognized anomaly detection as a hot research topic in the domain of data mining. Advancement in computer technology has motivated researchers to shift their focus from low dimensional data to high dimensional data. Techniques to investigate high dimensional data could be categorized into two aspects, either to explore through full feature space or just subspaces. Local neighborhood-based techniques like LOF, LOCI, COF and IN-FLO have proved excellent because of its ability to separate clusters of arbitrary shapes. Unfortunately,

the above-mentioned techniques work efficiently only for low dimensional data. High dimensional data wishes its explorers to take care of its embedded issues which are the similarity of data points and curse of dimensionality. Full feature spaces are concerned with the likeness of data points so traditional techniques fail altogether. On the other hand, the accuracy of results is compromised when subspace-based anomaly detection is exploited. This study involves the differentiation of normal and abnormal points through normalized distance metric methods. Each attribute of the data set is examined to find variance so that each attribute is classified and normalized accordingly. In this regard, Local neighborhood-based methodology is adapted for full feature space to detect anomalies present in high dimensional data.

Acknowledgement

Authors acknowledge the Department of Computer Science and IT, The Islamia University of Bahawalpur Pakistan and Department of Computer Science, Khawaja Fareed University Rahim Yar Khan Pakistan, for facilitating a suitable environment for the successful completion of this research work.

References

1. Aggarwal, C. C., Yu, P. S. Outlier Detection for High Dimensional Data. *ACM Sigmod Record*, 2001, 37-46. <https://doi.org/10.1145/375663.375668>
2. Agrawal, A. Local Subspace Based Outlier Detection. *International Conference on Contemporary Computing*, 2009, 149-157. https://doi.org/10.1007/978-3-642-03547-0_15
3. Ayesha, S., Hanif, M. K., Talib, R. Overview and Comparative Study of Dimensionality Reduction Techniques for High Dimensional Data. *Information Fusion*, 2020. <https://doi.org/10.1016/j.inffus.2020.01.005>
4. Bai, M., Wang, X., Xin, J., Wang, G. An Efficient Algorithm for Distributed Density-Based Outlier Detection on Big Data. *Neurocomputing*, 2016, 19-28. <https://doi.org/10.1016/j.neucom.2015.05.135>
5. Benito, M., Parker, J., Du, Q., Wu, J., Xiang D., Perou, C. M. Adjustment of Systematic Microarray Data Biases. *Bioinformatics*, 2004, 20, 105-114. <https://doi.org/10.1093/bioinformatics/btg385>
6. Chandola, V., Banerjee, A., Kumar, V. Survey of Anomaly Detection. *ACM Computing Surveys*, 2009, 41, 1-72. <https://doi.org/10.1145/1541880.1541882>
7. Chatterjee, S., Hadi, A. S. Sensitivity Analysis in Linear Regression. *John Wiley & Sons*, 2009, 327.
8. Dighe, M., Gawde, G. Improving Projected Clustering Algorithm for a High Dimensional Dataset. *IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, 2016, 1411-1415. <https://doi.org/10.1109/RTEICT.2016.7808064>
9. Domingues, R., Filippone, M., Michiardi, P., Zouaoui, J. A Comparative Evaluation of Outlier Detection Algorithms: Experiments and Analyses. *Pattern Recognition*, 2018, 74, 406-421. <https://doi.org/10.1016/j.patcog.2017.09.037>
10. Dutta, J. K., Banerjee, B. Improved outlier Detection Using Sparse Coding-Based Methods. *Pattern Recognition Letters*, 2019, 122, 99-105. <https://doi.org/10.1016/j.patrec.2019.02.022>

11. Feldbauer, R., Flexer, A. A Comprehensive Empirical Comparison of Hubness Reduction in High-Dimensional Spaces. *Knowledge and Information Systems* 59, 2019, 1, 137-166. <https://doi.org/10.1007/s10115-018-1205-y>
12. Filzmoser, P., Maronna, R., Werner, M. Outlier Identification in High Dimensions. *Computational Statistics & Data Analysis*, 2008, 52, 1694-1711. <https://doi.org/10.1016/j.csda.2007.05.018>
13. Hawkins, D. M. *Identification of Outliers*. Springer, vol. 11, 1980. <https://doi.org/10.1007/978-94-015-3994-4>
14. Hinneburg, A., Aggarwal, C. C., Keim, D. A. What is the Nearest Neighbor in High Dimensional Spaces? 26th International Conference on Very Large Databases, 2000, 506-515
15. Hotta, S. Local Subspace Classifier with Transform-Invariance for Image Classification. *IEICE TRANSACTIONS on Information and Systems*, 2008, 91(6), 1756-1763. <https://doi.org/10.1093/ietisy/e91-d.6.1756>
16. Hubert, M., Reynkens, T., Schmitt, E., Verdonck, T. Sparse PCA for High-Dimensional Data with Outliers. *Technometrics*, 2016, 58, 424-434. <https://doi.org/10.1080/00401706.2015.1093962>
17. Hussain, S. F., Haris, M. A K-Means Based Co-clustering (KCC) Algorithm for Sparse, High Dimensional Data. *Expert Systems with Applications*, 2019, 118, 20-34. <https://doi.org/10.1016/j.eswa.2018.09.006>
18. Jin, W., Tung, A. K., Han, J., Wang, W. Ranking Outliers Using Symmetric Neighborhood Relationship. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2006, 577-593. https://doi.org/10.1007/11731139_68
19. Jindi, L., Huaji, Z. Outlier Detection in Dairy Cows Estrus Based on Density Clustering. *IEEE International Conference on Computer and Communications (ICCC)*, 2017, 2291-2294. <https://doi.org/10.1109/CompComm.2017.8322943>
20. Jung, S., Marron, J. S. PCA Consistency in High Dimension, Low Sample Size Context. *The Annals of Statistics*, 2009, 37, 4104-4130. <https://doi.org/10.1214/09-AOS709>
21. Kaur, A., Datta, A. Detecting and Ranking Outliers in High-Dimensional Data. *International Journal of Advances in Engineering Sciences and Applied Mathematics*, 2019, 11, 75-87. <https://doi.org/10.1109/SWOD.2007.353201>
22. Keller, F., Muller, E., Bohm, K. HiCS: High Contrast Subspaces for Density-Based Outlier Ranking. *IEEE 28th International Conference on Data Engineering*, 2012, 1037-1048. <https://doi.org/10.1109/ICDE.2012.88>
23. Khan, D. M., Shahzad, F., Saher, N., Mohamudally, N. *Optimization and Analysis of Clusters*. Science International Lahore, ISSN 1013-5316, 2014, Vol. 5(26), 1951-1971.
24. Knorr, E. M., Ng, R. T. A Unified Approach for Mining Outliers. *Proceedings of the 1997 conference of the Centre for Advanced Studies on Collaborative research*, 1997, 11.
25. Kriegel, H.P., Kröger, P., Schubert, E., Zimek, A. LoOP: Local Outlier Probabilities. *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, 2009, 1649-1652. <https://doi.org/10.1145/1645953.1646195>
26. Kriegel, H.P., Schubert, M., Zimek, A. Angle-Based Outlier Detection in High-Dimensional Data. *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008, 444-452. <https://doi.org/10.1145/1401890.1401946>
27. Lee, J. G., Han, J., Li, X. Trajectory Outlier Detection: A Partition-and-Detect Framework. *IEEE 24th International Conference on Data Engineering*, 2008, 140-149. <https://doi.org/10.1109/ICDE.2008.4497422>
28. Li, Y., Kitagawa, H. Db-outlier Detection by Example in High Dimensional Datasets. *IEEE International Workshop on Databases for Next Generation Researchers*, 2007, 73-78. <https://doi.org/10.1109/SWOD.2007.353201>
29. Liu, Z., Pi, D., Jiang, J. Density-based Trajectory Outlier Detection Algorithm. *Journal of Systems Engineering and Electronics*, 2013, 24, 335-340. <https://doi.org/10.1109/JSEE.2013.00042>
30. Mishra, S., Chawla, M. A Comparative Study of Local Outlier Factor Algorithms for Outliers Detection in Data Streams. *Emerging Technologies in Data Mining and Information Security*, 2019, 347-356, Springer Singapore. https://doi.org/10.1007/978-981-13-1498-8_31
31. Müller, E., Assent, I., Iglesias, P., Mülle, Y., Böhm, K. Outlier Ranking via Subspace Analysis in Multiple Views of the Data. *IEEE 12th International Conference on Data Mining*, 2012, 529-538. <https://doi.org/10.1109/ICDM.2012.112>
32. Naik, P., Wedel, M., Bacon, L., Bodapati, A., Bradlow, E., Kamakura, W. Challenges and Opportunities in High-Dimensional Choice Data Analyses. *Marketing Letters*, 2008, 19(3), 201-213. <https://doi.org/10.1007/s11002-008-9036-3>

33. Nawaz, A. K. K., Khan, D. M., Saher, N., Shahzad, F. The Application of the Subdivision Algorithm for Surface Modeling. *Science International Lahore*, 2016, 28(2).
34. Paul, D. Asymptotics of Sample Eigenstructure for a Large Dimensional Spiked Covariance Model. *Statistica Sinica*, 2007, 1617-1642. www.jstor.org/stable/24307692.
35. Ramaswamy, S., Rastogi, R., Shim, K. Efficient Algorithms for Mining Outliers from Large Data Sets. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 2000, 427-438. <https://doi.org/10.1145/342009.335437>
36. Rehman, M. U., Khan, D. M. Local Neighborhood-Based Outlier Detection of High Dimensional Data Using Different Proximity Functions. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 2020, 11(4), 133-137. <https://doi.org/10.14569/IJACSA.2020.0110418>
37. Shibuya, H., Maeda, S. Anomaly Detection Method Based on Fast Local Subspace Classifier. *Electronics and Communications in Japan*, 2016, 99(1), 32-41. <https://doi.org/10.1002/ecj.11770>
38. Suri, N. M. R., Murty, M. N., Athithan, G. *Outlier Detection: Techniques and Applications*. Springer Nature, 2019
39. Tang, B., He., H. A Local Density-Based Approach for Outlier Detection. *Neurocomputing*, 2017, 241, 171-180. <https://doi.org/10.1016/j.neucom.2017.02.039>
40. Tang, J., Chen, Z., Fu, A.W.C., Cheung, D.W. Enhancing the Effectiveness of Outlier Detections for Low-Density Patterns. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2002, 535-548. https://doi.org/10.1007/3-540-47887-6_53
41. Tang, T., Chen, S., Zhao, M., Huang, W., Luo, J. Very-Large-Scale Data Classification Based on K-Means Clustering and Multi-Kernel SVM. *Soft Computing*, 2019, 23, 3793-3801. <https://doi.org/10.1007/s00500-018-3041-0>
42. Tripathy, S., Sahoo, L. Improved Method for Noise Detection by DBSCAN and Angle Based Outlier Factor in High Dimensional Datasets. *ICCCE*, Springer Singapore, 2020, 213-221. https://doi.org/10.1007/978-981-13-8715-9_27
43. Wang, T., Li, Q., Chen, B., Li, Z. Multiple Outliers Detection in Sparse High-Dimensional Regression. *Journal of Statistical Computation and Simulation*, 2018, 88, 89-107. <https://doi.org/10.1080/00949655.2017.1379521>
44. Yan, Y., Cao, L., Kulhman, C., Rundensteiner, E. Distributed Local Outlier Detection in Big Data. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, 1225-1234. <https://doi.org/10.1145/3097983.3098179>
45. Yuan, Y., Cao, H., Zhang, Y., Xie, Q., Yao, R. Outlier Mining Based on Neighbor-Density-Deviation with Minimum Hyper-Sphere. *Information Technology and Control*, 2016, 45(3), 267-277. <https://doi.org/10.5755/j01.itc.45.3.13164>
46. Zhang, C., Yin, A. Anomaly Detection Algorithm Based on Subspace Local Density Estimation. *International Journal of Web Services Research (IJWSR)*, 2019, 16, 44-58. <https://doi.org/10.4018/IJWSR.2019070103>
47. Zhang, J., Lou, M., Ling, T.W., Wang, H. Hos-miner: A System for Detecting Outlying Subspaces of High-Dimensional Data. *Proceedings of the Thirtieth International Conference on Very Large Data Bases*, 2004, 30, 1265-1268. <https://doi.org/10.1016/B978-012088469-8/50123-6>

