[Altaf*, 5(2): February, 2016]

ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.785

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

RELATIVE STUDY OF OUTLIER DETECTION PROCEDURES

Insha Altaf
Dept. of IT, National Institute of Technology, Srinagar, J&K, India

DOI: 10.5281/zenodo.46489

ABSTRACT

Data Mining just alludes to the extraction of exceptionally intriguing patterns of the data from the monstrous data sets. Outlier detection is one of the imperative parts of data mining which Rexall discovers the perceptions that are going amiss from the normal expected conduct. Outlier detection and investigation is once in a while known as Outlier mining. In this paper, we have attempted to give the expansive and a far reaching literature survey of Outliers and Outlier detection procedures under one rooftop, to clarify the lavishness and multifaceted nature connected with each Outlier detection technique. Besides, we have likewise given a wide correlation of the different strategies for the diverse Outlier techniques. Outliers are the focuses which are unique in relation to or conflicting with whatever is left of the information. They can be novel, new, irregular, strange or uproarious data. Outliers are in some cases more fascinating than most of the information. The principle difficulties of Outlier detection with the expanding many-sided quality, size and assortment of datasets, are the manner by which to get comparable Outliers as a gathering, and how to assess the Outliers data set.

KEYWORDS: Outliers, data mining, Clustering, Neural Network Outlier, Univariate Outlier detection, K-means algorithm.

INTRODUCTION

Data Mining is a non-paltry method of recognizing legitimate, novel, conceivably valuable lastly justifiable patterns [1]. Presently, data mining is turning into a critical instrument to change over the data into data. It is fundamentally utilized as a part of misrepresentation detection, promoting and exploratory disclosure. Data mining really alludes to removing the concealed intriguing patterns from the expansive measure of datasets and databases [2]. Mining is essentially used to reveal the patterns of the data, however this can be done on the specimen of data. The mining procedure will be totally fizzled if the specimens are not the great representation of the vast body of the data. Besides, the revelation of a specific example in a specific arrangement of the data does not as a matter of course imply that the example is discovered somewhere else in the bigger data from which that specimen is drawn [3].One of the essential reasons of utilizing the data mining is to adequately and productively break down the accumulation of the different perceptions as per their conduct. So as to do as such, clustering or cluster investigation is a decent option. Cluster analysis or clustering is the classification of the arrangement of perceptions lying under one cluster are diverse in some sense from the other cluster. It is an unsupervised learning method which really goes for discovering the thick and sparse locales in the dataset [4].Outlier detection is superb and an extremely significant idea of data mining which is likewise alluded as Outlier mining. Outlier detection alludes to the issue of discovering the patterns in the gigantic datasets that does not demonstrate the understanding with the summed up expected conduct. More often than not, these odd patterns are termed as Outliers, oddities, harsh perceptions, shortcoming, special cases, shock, and contaminants [5]. Clustering is an exceptionally compelling method to discover different Outliers, however just clustering is not adequate for dissecting and the detection of Outliers in light of the fact that at whatever point we are managing the expansive datasets and databases, the uncommon events are not kept to the Outliers rather they get to be higher dimensional Outliers and to manage these higher dimensional Outliers, versatile model based clustering is required where the clustering is scaled at the larger amounts, in order to add to the methods for taking care of vast databases, inside of the restricted computational assets, for example, memory and calculation time. In addition, in high dimensional space, the data is sparse and the thought of Proximity neglects to hold its weightiness.

Indeed, in high dimensional data, each Point is a practically equidistant from one another. Accordingly, because of the base distinction of the data points Outlierness will turn out to be progressively frail and undistinguishable. Hence, for high dimensional data, the idea of discovering important Outliers turns out to be considerably more perplexing and

awkward. Be that as it may, in the meantime, we can't say totally that the Outliers are the repercussion of clustering, however a portion of the past literature has consented to it or a portion of the specialists have consented to it. By or their methods Outliers are the Points that don't exist in the cluster, accordingly, these procedures verifiably characterize the Outliers as the foundation noise in which the clusters are implanted yet late literature characterizes the Outliers as the Points that are neither a part of the cluster nor a part of the foundation noise; rather they are particularly the Points that are all that much unique in relation to the standard. [5].Sometimes, these Outliers are such a great amount of intriguing for us that they go about as a managing media to identify the different abnormalities in the base datasets, (for instance in system, Intrusion detection). With the assistance of such Outliers, we are exceptionally well fit for finding the noise or inconsistencies that are really making ruin for our data such Outliers are otherwise called solid Outliers [5].

UNXIVARIATE OUTLXIER ANALYSIS

Univariate Outliers are the cases that have a strange worth for a solitary variable. One approach to recognize univariate Outliers is to change over the majority of the scores for a variable to standard scores. In the event that the simplex size is little (80 or less cases), a case is an Outlier if its standard score is ±2.5 or past. On the off chance that the example size is bigger than 80 cases, a case is an Outlier if its standard score is ±3.0 or past. This method applies to interim level variables, and to ordinal level variables that are dealt with as metric.

K-MEANS CLUSXTERING

In data mining cluster analysis should be possible by different methods. One such method is K-implies algorithm which parcel n-perceptions into k-clusters in which every perception has a place with the cluster with the nearest mean. K-implies clustering works on real perceptions and a solitary level of clusters is made. K-implies clustering is frequently more suitable than hierarchical clustering for a lot of data. Every perception in your data is dealt with as an item having an area in space. Allotments are shaped based upon the way that questions inside of each cluster are as near one another as could be allowed, and as a long way from articles in different clusters as could be allowed. We can look over five changed distance measures, contingent upon the sort of data we are clustering. Each cluster in the parcel is characterized by two things, its part protests and centroid (the Point to which the whole of distances from all items in that cluster is minimized), or focus. Cluster centroids are registered contrastingly for each distance measure, to minimize the whole regarding the measure that you determine. An iterative algorithm is utilized, that minimizes the entirety of distances from every item to its cluster centroid, over all clusters. This algorithm continues moving the articles between clusters until the whole can't be diminished further giving an arrangement of clusters that are as reduced and very much isolated as could be expected under the circumstances.

LITERATXURE SURXVEY

Outliers are the patterns of the data that don't consent to the general expected conduct. Consider the accompanying fig1 beneath [6] the data has two typical districts N1and N2, so the vast majority of the perceptions lie in these two locales, Points that are adequately far from the areas; e.g., pts.O1, O2, and the Points in O3, are the Outliers.

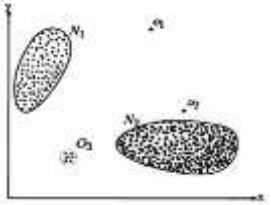

Fig 1. A simple example of Outliers in a 2-dimenxsional data Set

In spite of the fact that, the reality of the matter is that the Outliers are available in the data due to a few malignant exercises like credit card misrepresentation, digital intrusion or breakdown of the framework yet in the meantime, they are extremely intriguing to the expert and the interestingness or Rexall life pertinence of the Outliers is a key element of Outlier detection. As it is Rexall fascinating to the expert, so we can't contrast it and noise evacuation [6] and noise convenience both of which manage the undesirable noise. Noise can be termed as a deterrent in the data investigation so it is not required in the data. Noise evacuation is driven by the need to uproot the undesirable articles before any data examination is performed on the data. Noise Accommodation alludes to vaccinating measurable model

estimations against atypical perceptions. Oddity detection is additionally a developing idea that is infrequently blended with Outlier detection, however it is to some degree discovering the novel data and it ought not be mistaken for Outlier detection [7].It Rexall appears to be extremely basic that Outliers are the perceptions that are veering off from their typical or expected conduct, it's anything but difficult to imagine them, yet in Reality it is not a straightforward errand. There are numerous troubles to characterize the ordinary conduct or a typical district in order to make the Outliers more conspicuous. These challenges are as [6]:

1. To incorporate and to cover the each conceivable ordinary conduct in the locale it is a dull undertaking.
2. As there is a flimsy line between the peripheral area and ordinary district, in this way, some of the time the perceptions lying in the typical locale are considered as Outliers and the other way around.
3. Novel data and Noise are available in the data which has a tendency to be like the real data and subsequently hard to recognize them from Outliers, and further to evacuate them.
4. Moreover, which procedure would be connected to uproot a specific sort of Outlier is again an extremely troublesome undertaking. In this way, these are a percentage of the viewpoints not Rexall makes the Outlier detection.

A. Various Factors Determining the Outlier Detection Problem

As distant perceptions appears to be exceptionally intriguing to us, so the detection of Outliers likewise turn into a vital Point, there are a few components that decide how to detail an Outlier detection issue:

1. Nature of Input Data

The imperative variable of any Outlier detection is to think about the input data. Input is by and large an accumulation of data cases and each data example can be depicted as far as traits which can be further alluded as variable, trademark highlight and so forth. Each data case, if comprises of standout variable, then it is known as univariate and on the off chance that it comprises of more than one variable then it is multivariate. If there should arise an occurrence of multivariate, data cases, all qualities may be of same sort or may be the blend of various data sorts. We can have distinctive sorts of input data, for example, continuous data, categorical data, spatial data, so relying on the way of the data, we can apply the specific Outlier detection procedure [8].

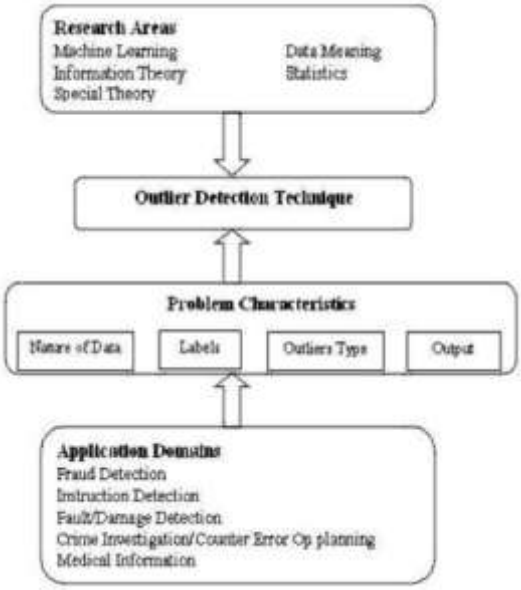

Fig2. Key Component Associated with Outlier Detection Techniques

2. Types of Outlier

A critical part of an Outlier detection strategy is to consider the nature and kind of the Outliers. Outliers can be sorted into the accompanying five classes – Type – A (Point Outliers), Type–B (Contextual Outliers), Type – C (Collective Outliers), Type – D (Rexall Outliers), Type – E (Erroneous Outliers).

Type-A (Point Outliers):- On the off chance that an individual data occurrence can be considered as abnormal concerning whatever remains of the data, then the example is termed as a Point Outlier. It is one of the least complex types of Outliers and is the key center of dominant part of exploration in Outlier Detection [9]. For instance : If we consider the credit card fraud detection with the dataset, relating to an individual's credit card , then the exchange for which the sum spent is extremely higher contrasted with the typical scope of use for that individual is termed as an Outlier [10].

Type – B (Contextual Outliers):- On the off chance that a data case is an uncommon event as for some particular connection and it is a typical event regarding some another setting, then such sorts of data occasions are known as Contextual data sets, the Contextual Outliers incorporate the time arrangement data excessively [10], here the time is a Contextual attribute which decides the position of an example on the whole grouping, then it is generally called as the type B–Outliers. The decision of applying a Contextual Outlier detection system is controlled by the weightiness of the Contextual Outliers in the biggest application space. Applying a Contextual Outlier detection procedure bodes well, if Contextual characteristics are promptly accessible and consequently characterizing a setting is straightforward.

Type – C (Collective Outliers):- On the off chance that an individual data instance is not odd but rather its gathering with the whole dataset is abnormal, then it is termed as a Collective Outlier, the individual data examples which are termed as Collective Outlier may not be the Outlier themselves but rather their event together as an accumulation is irregular, consequently it is a Collective Outlier [11]. Collective Outliers have been investigated for grouping data [12] graph data [13] and spatial data [14]. It ought to be noticed that while Point Outliers can happen in any data set, in which the data cases are connected. In control, event of the Contextual Outliers relies on upon the accessibility of the connection properties in the data. A Point Outlier or a Collective Outlier can likewise be Contextual Outlier, if broke down regarding a setting. In this way, the Point Outlier detection issue can be changed to a Contextual Outlier detection issue by fusing the context advice activity [12].

Type – D (Rexall Outliers): - These are the Rexall distant perceptions which are of enthusiasm to the framework expert. These perceptions do have some interestingness that finds the investigator something new and imaginative and in the event that they are evacuated at any rate, we are totally left with the ordinary area however this does not mean at all that we are contrasting them and noise or novel data. They can't be viewed as noise rather they are the Rexall Outliers [12].

Type – E (Erroneous Outliers): - In the event that some perception is noted inaccurately as an Outlier, because of some characteristic issue, or some catastrophic failure, then these are by error Outliers or we can say illusive Outliers. They Rexall take the result of the data in some other way. Xx

3. DAXTA LABELS (Fundamentals Approaches of the Outlier Detection)

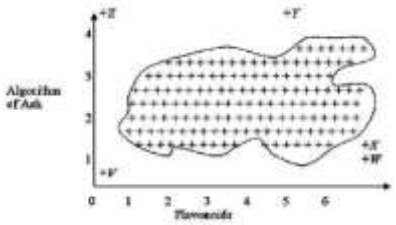

Fig 3. Shows the normal region and indicating V, W, X, Y, Z as Outliers

The marks as a rule connected with the data occurrences shows the ordinary conduct of the data or the atypical conduct of master and consequently requires considerable push to get the named training data set.

(I) Supervised Outlier Detection

Techniques prepared in a supervised mode expect the accessibility of the training data set which has named cases for the ordinary and the Outlier class. Any inconspicuous data occurrence is contrasted against the model with figure out

which class it fits in with. Yet, there are two noteworthy issues that might emerge in supervised Outlier detection; to begin with, the odd cases are few, when contrasted with the typical occasions, in the training data. Also, acquiring precise and agent marks, particularly, for the Outlier class is generally testing [15] [16]. In addition there are numerous methods that infuse simulated Outliers in a typical dataset to acquire a marked preparing data set.

(ii) Semi Supervised Outlier Detection:

Strategies which are working in a semi supervised mode, accept that the training data has the marked examples, just for the normal class. Since they don't require names for the Outlier class, they are all the more broadly utilized when contrasted with the supervised systems. For instance, in space craft fault detection [17], an Outlier situation will be a mishap, which is difficult to demonstrate .The run of the mill approach utilized as a part of such systems is to assemble a model for the class relating to the normal conduct, and after that utilization the model to distinguish Outliers in the test data.

(iii) Unsupervised Techniques:

Unsupervised Outlier detection procedures don't require training data, and therefore are most broadly relevant. The strategies in this class make the understood presumption, that the ordinary occurrences are significantly more incessant when contrasted with the Outliers, if this is not the case then, these systems are not that much effective, and they experience the ill effects of the high false alert rate [17][18]. In the rest of this paper, we order the broadness, wealth and complexities of the different Outlier detection methodologies. We have attempted to give the literature audit of critical Outlier detection techniques, for example, Statistical Methods, Parametric and Non Parametric methods, Proximity Based Methods, Distance Based Methods, and Density Based Methods, Clustering Methods lastly Neural Network techniques to recognize and dissect the Outliers, besides, we have given their examination as well.

B. STATISTICAL METHODS

Statistical methods are one of the most punctual algorithms or the models that can be utilized by the different Outlier detection methodologies [19]. They are broadly utilized not just to distinguish the remote perceptions, rather to break down them in order to ponder the complete dataset based on them. One of the single dimensional univariate method is Grubb's method which is Extreme Studentized Deviate [20] which ascertains a Z Value as the contrast between the mean estimation of the property and the question esteem isolated by the standard deviation for the characteristic where the mean and standard deviations are computed from all quality qualities including the inquiry worth and Z esteem for the inquiry is contrasted and the 1% or 5% essentialness level [21]. Additionally, this system does not require any parameters from the client as all the parameters are straightforwardly gotten from the data itself. Be that as it may, in the event that we need to pinpoint the Outliers for both the univariate and multivariate then as per Laurikkala et al, [22] use casual box plots. Fig3. A Data Distribution with 5 Outliers (V, W, X, Y and Z) [23]. These box plots produces graphical representation and permits the human evaluator to for the most part pinpoint the peripheral Points, incorporates the typical investigation, they can deal with Rexall esteemed, ordinal and all out traits box plots a considerable measure the lower amazing, lower quartile, median, upper quartile and upper extreme Points. So as to comprehend this, consider the figure 3. Here, in figure 3 above [23], we are having a normal region, and the Points V, Z, Y, X, W, are the Outliers which are demonstrated past the normal region, with a specific end goal to demonstrate this thing in casual box plot, we will demonstrate to it as in fig 4 underneath

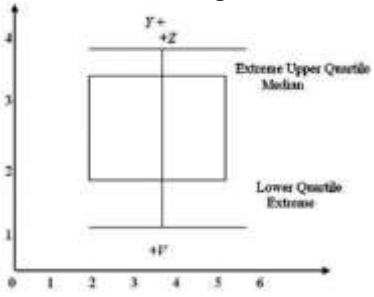

Fig 4. Informal box plot showing the Outliers V, Y, Z

Figure 4 demonstrates a box plot of the Y-hub values for the data in Fig-3 with the lower great, the lower quartile,

median, upper quartile and upper extreme from the normal data and the three Outliers V, Y and Z plotted the Points X and W from Figure-2 are not Outliers as for the Y-hub values. Box plots make no presumptions about the data appropriation display yet are dependent on a human to take note of the great Points plotted on the box plot. There emerges an issue when the dimensionality builds, the data Points are spread through the bigger volumes and turn out to be less thick, this issue is known as the scourge of dimensionality which facilitate expands the spreading so as to prepare time and misshapes the data dissemination the raised bodies. Be that as it may, there are numerous strategies like k-N.N, Neural Networks, MVE i.e., Minimum volume Ellipsoid and so on that are all that much defenseless to the scourge of dimensionality. These are fundamentally the feature selection systems that are utilized basically to expel the noise from the data dispersion and spotlight on the principle cluster of the typical data Points while separating the Outliers [23].

C. PROXIMITY BASED TECHNIQUES

Proximity Based Techniques don't require any earlier presumption about the data – dissemination and they are nearly easy to actualize in spite of the fact that they are computationally extremely mind boggling as they more often than not experience the ill effects of the exponential computational development. The computational many-sided quality is specifically relative to both the dimensionally of the data m and the quantity of the records n. Essential thought in the Proximity Based Techniques is to display the Outliers as Point which are detached from the remaining data. This procedure essentially incorporates three vital methods – nearest neighbor analysis. K-NN i.e., K Nearest Neighbor [23][24] method utilized which is otherwise called occasion – based learning or lazy learning is utilized for classifying the items based on nearest preparing samples in the component space. It is just approximated locally and all the calculation as a rule ascertains the nearest neighbors of a record utilizing a suitable distance metric known as Euclidean Distance and can be given by mathematical statement 1 and is basic the vector distance though the Mahalanobis distance is given by comparison 2

Ascertain the distance from a pt. to the centroid (M) characterized by corresponded properties given by the covariance matrix (C). For vast and higher dimensional datasets, Mahalonobis distance is computationally extremely costly as it requires a go through the whole data set to recognize the property co-connection. If there should arise an occurrence of the clustering methods, which is another Proximity based method clustering algorithms are made with a specific end goal to decide – the thick districts of the data set. In the following step, some metric to quantify the attack of the data Points to the distinctive clusters is utilized as a part of request to register an Outlier score for the data Point, for instance, while utilizing k-implies algorithm, the distance of the data Point to the closest centroid might be utilized to gauge its irregular conduct one of the challenges with the clustering algorithms is that they verifiably accept the particular sorts of cluster shakes relying on the particular algorithm or distance function utilized inside of the clustering algorithm [24].Density based methods give a more elevated amount of interpretability, when the inadequate areas in the data can be displayed as far as mixes of the first characteristic. We will examine the clustering and density based methods further in this paper.

D. PARAMETRXIC TECHNIQUES

Parametric methods scale themselves extremely well in the event of the different modifications and optimizations done in the standard algorithms. In addition, the models utilizing the parametric techniques become just with the model multifaceted nature not the data size. Nonetheless, they confine their pertinence by upholding the preselected distribution model to fit the data. In the event that the client is as of now mindful about their data fits such a distribution model, then these methodologies are exceptionally exact yet numerous data sets don't fit a specific model [25]. Regression Method is one of the major parametric methods. The regression analysis means to discover a reliance of one or more irregular variables y on another or more variables x. This includes examining the conditional probability distribution y/x. The regression model can either be a linear or nonlinear model that fits the data, contingent on the decision from the clients [26]. Any data Point is set apart as an Outlier if a momentous deviation happens between the actual value and its normal quality created by the regression model. Extensively talking there are fundamentally two approaches to utilize the data in the dataset for building the regression model for Outlier detection, to be specific the opposite hunt and direct search methods [26]. The converse search method builds the relapse model by utilizing all data accessible and after that the data with the best blunders are considered as Outliers and rejected from the model. In the immediate pursuit approach, one forms a model based on the part of the data and afterward includes new data Points incrementally when the preparatory model development has been done. At that point, the model is stretched out by including most fitting data, which are those articles in whatever remains of the populace that have the least

deviations from the model developed in this way. The data added to the model in the last round, thought to be the slightest fitting data are viewed as Outliers [18] [26].

E. NON-PARAMETRXIC METHXODS:

The Outlier detection procedures in this class don't make any suspicions about the statistical distribution of the data. The most essential methodologies for Outlier detection in this class are histograms and Kernel Density Function or Kernel Feature Space [27] Methods.

Histograms:

It is a standout amongst the most well-known systems of Non-Parametric method which is essentially used to keep up a profile of the data. Histogram techniques by nature are based on the frequency or tallying of the data. The histogram based Outlier detection methodology is regularly connected when the data has a solitary component. Scientifically, a histogram for a component of data comprises of various disjoint canisters (or cans) and the data are mapped into one receptacle. The stature of the container compares to the quantity of perceptions that fall into the receptacles [23] [28]. Subsequently, if n be the aggregate no. of occasions, k be the aggregate number of the containers and mi be the quantity of the data Point in the ithbin ($1 \leq i \leq k$) the histogram fulfills the accompanying condition:

1. The histogram strategies commonly characterize a measure between another test occurrence and the histogram based profiles figure out whether it is an Outlier or not. In particular, there are three conceivable routes for building a histogram.
2. Histogram development is just based on the typical data. They are utilized just to represent the profile of the typical data. The test arrange fundamentally assesses whether the component esteem in the test example falls in any of the populated canisters of the built histograms. If not the test occasion is marked as an Outlier [28].
3. Secondly, the histograms can be developed based upon the Outliers. In that capacity, the histograms catch the profile for the Outliers. A test occasion that can be categorized as one of the populated receptacles is marked as an Outlier [28]. Such systems are especially prevalent in intrusion detection group and fraud detection.
4. The histogram can be developed based on a blend of the typical data and the Outliers. Subsequent to normal data commonly command the entire data set, therefore the histogram speaks to an approximated profile on the ordinary data. The sparsity of the receptacle in the histogram can be characterized as the proportion of the frequency of the canister against the normal frequency of the considerable number of containers in the histogram. A canister is considered as scanty if such proportion is lower than the client determined limit [28].

As the histogram based detection methods are easy to actualize and subsequently are very well known in space, for example, intrusion detection. Yet, one essential inadequacy of such methods for multivariate data is that they are not ready to catch the communications between qualities. An Outlier may have the attribute values that are exclusively visit, however their blend is exceptionally uncommon. Another issue with the histogram is that clients need to decide an ideal size of the containers to build the histogram

Kernel Function

This is another extremely well known non-Parametric approach for the Outlier detection [29] this includes utilizing kernel functions. Another case which lies in the low probability density range is pronounced as an Outlier. Formally, if x1, x2… ….xn are IID (incrementally and indistinguishably conveyed) tests of irregular variable x2then the kernel density guess of its probability density function [29] is

Where k is the kernel function and h is the data transfer capacity (smoothing parameter). Entirely logged off k is taken to be the standard Gaussian function with mean h=0 and difference .$\sigma 2=1$

Kernel density estimation of Probability Density Function (PDF) is relevant to both univariate and multivariate data. Notwithstanding, its estimation of a multivariate data is significantly more computationally costly than the univariate data. This renders the kernel density estimation methods somewhat wasteful in Outlier detection for high dimensional data [28] [29].

The significant contrasts between the Parametric and non-Parametric methods is that previous accept the hidden distribution of the given data and appraisal the parameters of the distribution model from the given data while the later don't expect any learning of distribution characteristics [29]

C. DISTANCE BASED METHODS

Distance based Outlier systems are a standout amongst the most generally acknowledged and every now and again utilized strategies that are utilized as a part of machine learning and data mining and it totally relies on upon the idea of neighborhood (KNN) of the data Points[30]. This idea can likewise be termed as Nearest Neighbor Analysis and it can likewise be connected for various purposes, for example, classification, clustering and in particular Outlier detection. The most noteworthy component of the nearest neighbor based Outlier detection system is that they have an unequivocal thought of Proximity that is characterized as a distance and comparability measure for any of two individual data occasions, or a set or a succession of occurrences. Firstly, in distance based methods, distance between the data Points must be figured with the assistance of Lp measurements, as Manhattan distance and Euclidean distance metrics for measuring the distance, some non-metric distance functions are likewise utilized for making the distance based meanings of Outliers exceptionally broad. For groupings, a distance metric between two successions should be characterized. For Spatial data, Kou et al [2006] consolidate spatial correlation between data Points while determining the distance [31] [32]. A distance measure can be utilized for the data containing the blend of all out and ceaseless properties for Outlier detection. The connections between the two occurrences can be characterized by including distance for unmitigated and nonstop qualities independently. For straight out properties, the no. of the traits for which the two occurrences have the same qualities characterizes the distance between them. For nonstop attributes, a Covariance framework is kept up to catch the conditions between the consistent qualities [33]. Besides, there is an issue that how one can pronounce an example as an Outlier. This should be possible by sorting the nearest neighbor based Outlier detection system into two classes based on how the Outliers are measured as for the nearest neighbors. The primary classification comprises of those systems which measures the distance of a case from its nearest neighbor set and apply the diverse tests to recognize the Outliers, for example, the occasion is more than the distance d from its nearest Point or the distance is more than the distance d from its nearest neighborhood. [34]. in this manner, the idea of the distance based Outliers does not accept any basic data distributions and sums up numerous ideas from distribution based methods. Additionally, the distance based methods scale better to higher dimensional space and can be figured all the more successfully and proficiently when contrasted with the measurable methods.

D. DENSITY BASED OUTLIER DETECTION METHODS

Density based methods utilize more perplexing system or computationally more unpredictable to demonstrate the Outlierness of the data Points when contrasted with the distance based methods. It not just discovers the neighborhood densities of the Point being concentrated additionally the nearby densities of its nearest neighbors. In spite of the fact that the density based methods connotes a more grounded displaying capacities of the Outlier, yet they require the costly calculations in the meantime. [28] [35] the essential thought of LOF method is that, the circulation of distances between a data Points and all different Points will appear to be like the cumulative distance distribution for all pair-wise distances if there are numerous other close-by Points. This is a backhanded method for recognizing Outliers. There may be Points that are genuine Outliers and for which the tops in their distance distribution may coordinate the crests in the cumulative distribution. An Outlier score can be doled out to any given data Point, known as Local Outlier Factor (LOF), contingent upon its distance from its local neighborhood [34]. In this manner this plan finds the local Outlier score of any data Point and is not influenced by the varieties in the density of dissemination of the data. The benefit of LOF over the simple nearest neighbor approach proposed by Ramaswamy et al. [2000], [36] is LOF of an item mirrors the density contrast between its density and those of its neighborhood [36]. The lower the density of p or higher the density of p's neighbor, the bigger the estimation of LOF which shows that p has a higher level of being an Outlier. More often than not LOF misses the potential Outliers whose nearby neighborhood density is near that of its neighbors To conquer this issue, another connectivity based Outlier element (COF) plan that enhances the viability and productivity of LOF plan when the pattern itself has comparative neighborhood density as an Outlier[37][38] So, In a net might, we can say that despite the fact that the density based methods are computationally more unpredictable and costly when contrasted with the distance based methods yet at the same time they are far superior as it were that the density based methods researches not just the neighborhood density of the Point being concentrated additionally the neighborhood densities of the nearest neighbors. Also, it can be further scaled to the higher dimensional data all the more effortlessly, adequately and effectively. Dimensional data is the Clustering. In the literature survey we have

watched that such a large number of data mining algorithms discover the Outliers as a repercussion of clustering themselves and they characterize the Outliers as the Points that don't lie in or situated far from the cluster. Parceling Clustering Method is the principal classification of partitioning so as to clustering methods that performs clustering the data set into the particular no. Of clusters .The no. of the clusters to be gotten is signified by k, is determined by human clients. It as a rule begins with the introductory Position and afterward the objective function is improved until the data set amplifies the ideal estimation of the data set. In the Partitioning methods, different centroid based methods, mediods based methods, PAM, CLARA, k-means, and CLARANS and so on methods are utilized [39]. Another vital classification of clustering method is hierarchical clustering. In hierarchical clustering, the entire data set is further disintegrated into different subsets or little datasets. Hierarchical clustering is further isolated into two classes i.e., Agglomerative methods and divisive methods. An Agglomerative method as a rule begins with each Point as a distinct cluster and it joins two nearest clusters in each back to back stride until the limit condition is met [39]. A divisive method, in opposition to an agglomerative method, starts with all the Points as a solitary cluster and parts it in the following back to back stride until the threshold condition is met. Agglomerative methods are more prevalent being used. [37][39]. Different Hierarchical Methods are MST clustering, CURE and CHAMALEON. In addition, in extensive databases, BIRCH [40], is utilized and it can be improved for higher dimensional data.

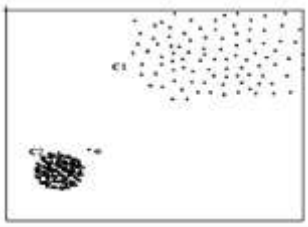

Fig 5. : A sample dataset showing the advantage of LOF over Distance based Methods for Outlier detection [39]

E. Clustering Based Methods

The critical and regularly utilized classification of Outlier detection methodology that is generally utilized for moderately low dimensional data is the Clustering. In the literature survey we have watched that such a large number of data mining algorithms discover the Outliers as a side effect of clustering themselves and they characterize the Outliers as the Points that don't lie in or situated far from the cluster. Partitioning Clustering Method is the main class of clustering methods that performs clustering by partitioning the data set into the particular no. of clusters .The no. of the clusters to be gotten is indicated by k, is determined by human clients. It as a rule begins with the starting position and after that the objective function is optimized until the data set expands the ideal estimation of the data set. In the Partitioning methods, different centroid based methods, mediods based methods, PAM, CLARA, k-means, and CLARANS and so on methods are utilized [39]. Another critical classification of clustering method is hierarchical clustering. In hierarchical clustering, the entire data set is further deteriorated into different subsets or little datasets. Hierarchical clustering is further separated into two classes i.e., Agglomerative methods and divisive methods. An Agglomerative method more often than not begins with each Point as a particular cluster and it consolidates two nearest clusters in each successive stride until the threshold condition is met [39]. A divisive method, as opposed to an agglomerative method, starts with all the Points as a solitary cluster and parts it in the following sequential stride until the threshold condition is met. Agglomerative methods are more famous being used. [37][39]. Different Hierarchical Methods are MST clustering, CURE and CHAMALEON. In addition, in extensive databases, BIRCH [40], is utilized and it can be upgraded for higher dimensional data

F. NEUXRAL NETWORK METHODS:

Neural Network methodologies are normally non Parametric and model based and suits well to the hidden pattern and are fit for learning extensive complex class limits .The whole data set must be crossed different times to permit the system to settle and model the data accurately .Neural Networks are nearly less powerless to the scourge of Dimensionality when contrasted with the statistical methods; the neural systems are further of two sorts – Supervised Neural Methods and Unsupervised Neural Methods[23].Supervised Neural Networks utilize the classification of the data to drive the learning process. In the event that this classification of the data is inaccessible, then it is known as unsupervised neural system .Unsupervised neural systems contain hubs which contend to speak to partitions of the

data set. Likewise with Perceptron-based neural systems, choice trees or k-implies, they require a training dataset to permit the system to learn. They independently cluster the input vectors through hub position to permit the fundamental data distribution to be displayed and the ordinary/anomalous classes separated [41]. They expect that related vectors have normal element values and depend on recognizing these elements and their qualities to topologically display the data distribution. The neural system utilizes the class to alter the weights and thresholds to guarantee the system that can effectively classify the entire system. These methods are additionally used to distinguish the noise and novel data [42]. Neural Network is an exceptionally vital methodology that assumes an imperative part in the Outlier detection. The essential thought is to prepare the neural network on the normal training data and afterward distinguish oddities and uncommon analyzing so as to fascinate events the reaction of the trained neural system to the test input. On the off chance that the system acknowledges a test input, it is ordinary and if the system rejects a test is an Outlier [43].This is the straight forward neural network application method that is essentially connected in the Neural Network Inclusion Detector System (NNID), for distinguishing the Outliers in therapeutic demonstrative data [43] for identifying credit card fraud and for image sequence data [44]. The NNID framework prepares a back propagation network on the normal training data (an arrangement of client orders) to recognize the clients who execute those charges. Amid testing a test example is classified to have a place with one of the learnt client profiles, if the output client does not coordinate the genuine client who produced the command, it implies an intrusion. This is a sort of supervised methodology where every preparation case has a related class name. Generally, the examination of the exhibitions is done on these fine diverse neural networks. – Perceptron, Back Propagation, Perceptron Back Propagation Hybrid Radial Based function (RBF) and ARTMAP. Also, there is a database mining framework known as Cardwatch [45] that is utilized for credit card misrepresentation detection and is based on neural network learning modules, and it gives an interface to an assortment. In this way, Neural Networks are Rexall useful in distinguishing and evacuating the Outliers. All these methods are dissected and looked at in the table1 and fig6 demonstrates the variety of Efficiency with dimensions for all the methods. Also, fig7 and fig 8 demonstrates the variety of versatility and computational multifaceted nature with the measurements separately for all the methods.

Table1 Comparison of Outlier Detection Methodologies

Methodology used	Computational Complexity	Efficiency	H.D.D	Feature Space	Practical Applicability	Ordinary data Streams
Statistical Methods	Very Complex	Less	Not Applicable	Univariate	Statistical Data	Applicable
Parametric Method	Less Complex	More	Not Applicable	Univariate	Data sets with Prior Knowledge	Applicable
Non parametric Method	Less Complex	Efficient	Not Applicable	Univariate/ Multivariate	Profile of the data Is maintained	Applicable
Distance based Method	Easy	Efficient	Scalable	Multivariate	Based on closeness Of individual points	Not Applicable
Density based Method	Very Complex	More Efficient	Effectively Scalable	Multivariate	Based on the Closeness of points And nearest Neighbor too	Not Applicable
Clustering based Methods	Less Complex	Very Efficient	Effectively Scalable	Univariate/ Multivariate	Based on the Clustering of similar Data	Depends on The cluster
Neural Networks	Very Less	Very Efficient	Effectively Scalable	Multivariate	Applied on the Normal training data	Depends On training data

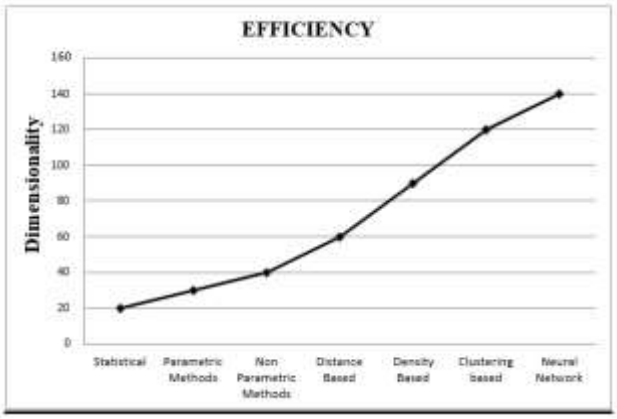

Fig 6 Shows the Efficiency of various Outlier methodologies with the increase in dimensionality

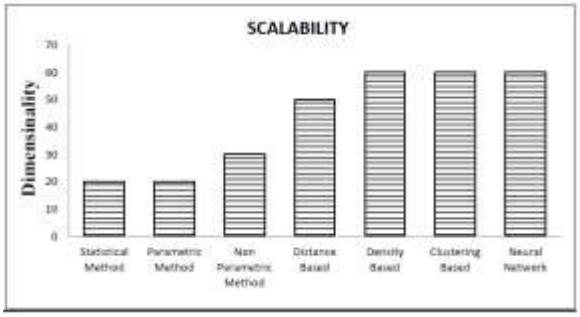

Fig 7 shows the scalability of the different Outlier detection methodologies with the increase in dimensionality

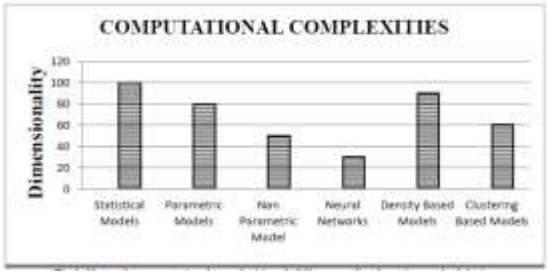

Fig 8. Shows the computational complexities of different Outlier detection methodologies

CONCLUSION

In this paper, it has emphasized that there is no universally accepted gamut of any methodology to detect and analyses the Outliers. However, the diversity, multiplicity and the comparative overview of all the Outlierdetection methodologies are tried to accommodate in this paper .Moreover, different ideologies of different researchers about various Outlier detection methodologies are Collectively organized under one roof so as to provide the comprehensive overview of these techniques which are very beneficial if further algorithms are made on these techniques.

REFERENCES

1. Aggarwal, C. C., Yu, S. P., "An effective and efficient algorithm for high-dimensional outlier detection, The VLDB Journal, 2005, vol. 14, pp. 211-221.
2. Abe, N, Zadrozny, B, and Langford, J. 2006.Outlier detection by active learning. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM Press, New York, NY, USA, 504 - 509.

3. Arning, A., Agrawal, R., and Raghavan, P.: 1996, 'A Linear Method for Deviation Detection in Large Databases'. In: Proceedings of the ACM SIGKDD
4. S. Vijayarani : [An Efficient clustering Algorithm, for outlier Detection IJCA,vol 32 oct,2011].
5. Charu C. Aggarwal, Phillip S. Y, An effective and efficient algorithm for higher dimensional outlier detection.
6. Karanjeet Singh and Dr. SuchitraUpadhyay. Outlier Detection: Applications and Techniques IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 3, January 2012 ISSN (Online): 1694-0814
7. Brother ton, T., Johnson, T., and Chadderdon, G.: 1998, 'Classification and Novelty Detection using Linear Models and a Class Dependent - Elliptical Bassi Function Neural Network '. In: Proceedings of the International conference on neural networks. Anchorage, Alaska.
8. Barnett, V. and Lewis, T.: 1994, Outliers in Statistical Data. John Wiley & Sons.3rd edition.
9. V. Chandola, A. Banerjee, and V. Kumar. Outlier Detection-A Survey, Technical Report, TR 07-017, Department of Computer Science and Engineering, University of Minnesota, 2007.
10. Dorronsoro, J. R., Ginel, F., Sanchez, C., and Cruz, C.S. 1997. Neural fraud detection in credit card operations.IEEE Transactions On Neural Networks 8, 4 (July), 827 -834.
11. Keogh, E., Lin, J., and Fu, A. 2005. Hot sax: Effciently finding the most unusual time series subsequence. In ICDM '05: Proceedings of the Fifth IEEE International Conference on Data Mining. IEEE Computer Society, Washington, DC, USA.
12. Teng, H., Chen, K., and Lu, S. 1990. Adaptive real-time outlier detection using inductively generated sequential patterns. In Proceedings of IEEE Computer Society Symposium on Research in Security and Privacy.IEEE Computer Society.
13. Sun, P,Chawla, S., and Arunasalam, B. 2006. Mining for outliers in sequential databases.In SIAM International Conference on Data Mining.
14. Noble, C. C. and Cook, D. J. 2003. Graph- outlier detection. In Proceedings of the 9th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM Press, 631 - 636.
15. Ester, M., Kriegel, H-P., and Xu, X.: 1996, 'A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise'. In: Proceedings ofthe Second International Conference on Knowledge Discovery and Data Mining,Portland, Oregon, pp. 226–231. AAAI Press.
16. Theiler, J. and Cai, D. M. 2003. Resampling approach for outlier detection in multispectral images.In Proceedings of SPIE 5093, 230-240, Ed.
17. Steinwart, I., Hush, D., and Scovel, C. 2005. A classification framework for outlier detection. Journal of Machine Learning Research 6, 211 – 232
18. Fujimaki, R.,Yairi, T., and Machida, K. 2005. An approach to spacecraft outlier detection problem using kernel feature space.In Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. ACM Press, New York, NY,
19. Bolton, R. J. and Hand, D. J.: 2001, 'Unsupervised Profiling Methods for Fraud Detection'. In: Credit Scoring and Credit Control VII, Edinburgh, UK, 5-7 Sept.
20. Barnett, V. and Lewis, T.: 1994, Outliers in Statistical Data. John Wiley & Sons.3rd edition.
21. Huber, P. 1974. Robust Statistics.Wiley, New York.
22. Grubbs, F. E.: 1969, 'Procedures for detecting outlying observations in samples'Technometrics**11**, 1–21.Hickinbotham, S. and Austin, J.: 2000, 'Novelty detection in Airframe Strain Data'. In: Proceedings of 15th International Conference on Pattern Recognition. Barcelona, pp. 536–539
23. Laurikkala, J., Juhola, M., and Kentala, E.: 2000, 'Informal Identification of Outliers in Medical Data'. In: Fifth International Workshop on Intelligent Data Analysis in Medicine and Pharmacology IDAMAP-2000 Berlin, 22 August. Organized as a workshop of the 14th European Conference on Artificial Intelligence ECAI-2000.
24. Hodge, V. and Austin, J. 2004. A survey of outlier detection methodologies.Artificial Intelli-gence Review 22, 2, 85.
25. Byers, S. and Raftery, A. E.: 1998, 'Nearest Neighbor Clutter Removal for Estimating Features in Spatial Point Processes'. Journal of the American Statistical Association **93**(442), 577–584.
26. V. Chandola, A. Banerjee, and V. Kumar. Outlier Detection-A Survey, Technical Report, TR 07-017, Department of Computer Science and Engineering, University of Minnesota, 2007.

27. Rousseeuw, P. J. and Leroy, A. M. 1987. Robust regression and outlier detection. John Wiley & Sons, Inc. New York, NY, USA.
28. Fujimaki, R, Yairi, T., and Machida, K. 2005. An approach to spacecraft outlier detection problem using kernel feature space. In Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. ACM Press, New York, NY,
29. Ji Zhang: A doctoral thesis titled as" Towards Outlier Detection For High-Dimensionaldata Streams Using Projected Outlier Analysis".
30. E. Parzen. On the estimation of a probability density function and mode. Annals of Mathematical Statistics 33, 1065-1076, 1962.
31. W. Jin, A. K. H. Tung, J. Han andW.Wang: Ranking Outliers Using Symmetric Neighborhood Relationship. PAKDD'06, 577-593, 2006
32. Knorr, E. M. and Ng, R. T.: 1998, 'Algorithms for Mining Distance-Based Outliers in Large Datasets '. In: Proceedings of the VLDB Conference. New York, USA, pp. 392–403.
33. Y kou,CT Lu,RF Dos Santos-Spatial outlier Detection- a graph based approach published in Tools with Artificial Intelligence 2007, ICTAI 2007,19th IEEE International Conference on Volume 1
34. A descriptive framework for the field of data Mining and Knowledge discovery by Yi Peng, Gang Kou, Yong SHI, and ZHENGXIN CHEN.
35. M. Breuning, H-P. Kriegel, R. Ng, and J. Sander. LOF: Identifying Density-Based Local Outliers. In Proc. of 2000 ACM SIGMOD International Conference on Management of Data (SIGMOD'00), Dallas, Texas, pp 93-104, 2000.
36. Zhang doctoral thesis titled as "Towards outlier detection for high-dimensional data streams using projected outlier analysis strategy".
37. Ramaswamy, S., Rastogi, R., and Shim, K. 2000. Efficient algorithms for mining outliers from large data sets.In Proceedings of the 2000 ACM SIGMOD international conference on Management of data.ACM Press, 427.
38. Charu C. Aggarwal, Phillip S. Y, An effective and efficient algorithm for higher dimensional outlier detection.
39. J. Zhang, M. Lou, T. W. Ling and H.Wang. HOS-Miner: A System for Detecting Outlying Subspaces of High-dimensional Data. VLDB Conference, 2004.