



Study of Outlier Detection Techniques for Low and High Dimensional Data

ASHWINI G.SAGADE¹, RITESH THAKUR²

¹Dept of Computer Engineering, IOKCOE, Pune University, Pune, India, Email: ashwinisagade25@yahoo.com.

²Dept of Computer Engineering, IOKCOE, Pune University, Pune, India, Email: hod_comp_iok@yahoo.com.

Abstract: Outlier detection has been used for centuries to detect and, where appropriate, remove anomalous observations from data. Outliers arise due to mechanical faults, changes in system behavior, fraudulent behavior, human error, instrument error or simply through natural deviations in populations. Their detection can identify system faults and fraud before they escalate with potentially catastrophic consequences. It can identify errors and remove their contaminating effect on the data set and as such to purify the data for processing. The original outlier detection methods were arbitrary but now, principled and systematic techniques are used, drawn from the full gamut of Computer Science and Statistics. In this paper, we introduce a study of contemporary techniques for outlier detection. We identify their respective motivations and distinguish their advantages and disadvantages in a comparative review.

Keywords: Data Mining, High Dimensional Dataset, Information Theory, Outlier Detection.

I. INTRODUCTION

Outlier detection refers to the problem of finding patterns in data that do not conform to expected behavior. These non-conforming patterns are often referred to as outliers, outliers, discordant observations, exceptions, aberrations, surprises, peculiarities or contaminants in different application domains. Of these, outliers and outliers are two terms used most commonly in the context of outlier detection; sometimes interchangeably. Outlier detection finds extensive use in a wide variety of applications such as fraud detection for credit cards, insurance or health care, intrusion detection for cyber-security, fault detection in safety critical systems, and military surveillance for enemy activities.

What are outliers? Outliers are patterns in data that do not conform to a well defined notion of normal behavior. Figure1 illustrates outliers in a simple 2-dimensional data set. The data has two normal regions, N1 and N2, since most observations lie in these two regions. Points that are sufficiently far away from regions, e.g., point's o1 and o2, and points in region O3, are outliers.

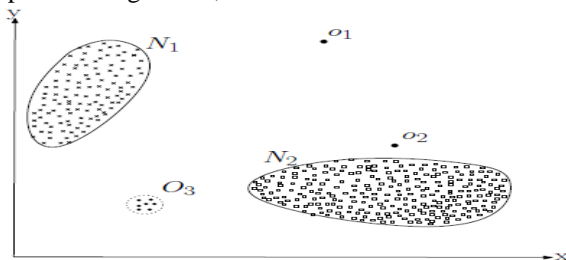


Fig1. A Simple example of outliers in a 2-dimensional data set.

At an abstract level, an outlier is defined as a pattern that does not conform to expected normal behavior. A straightforward outlier detection approach, therefore, is to define a region representing normal behavior and declare any observation in the data which does not belong to this normal region as an outlier. But several factors make this apparently simple approach very challenging: Defining a normal region which encompasses every possible normal behavior is very difficult. In addition, the boundary between normal and anomalous behavior is often not precise. Thus an anomalous observation which lies close to the boundary can actually be normal, and vice-versa. When outliers are the result of malicious actions, the malicious adversaries often adapt themselves to make the anomalous observations appear like normal, thereby making the task of defining normal behavior more difficult.

In many domains normal behavior keeps evolving and a current notion of normal behavior might not be sufficiently representative in the future. The exact notion of an outlier is different for different application domains. For example, in the medical domain a small deviation from normal might be an outlier, while similar deviation in the stock market might be considered as normal. Thus applying a technique developed in one domain to another is not straightforward. Availability of labeled data for training/validation of models used by outlier and hence is difficult to distinguish and remove. Due to the above challenges, the outlier detection problem, in its most general form, is not easy to solve. In fact, most of the existing outlier detection techniques solve a specific formulation of the problem. The formulation is

induced by various factors such as nature of the data, availability of labeled data, type of outliers to be detected, etc.

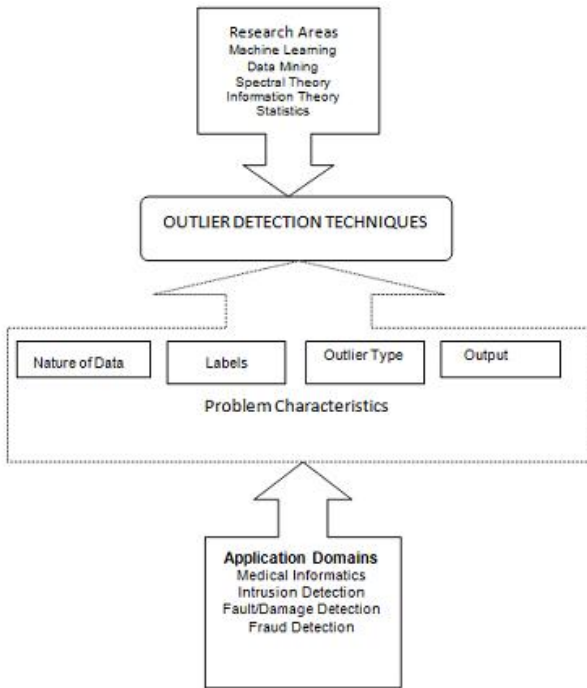


Fig2. Different aspects of an outlier detection problem

Often, these factors are determined by the application domain in which the outliers need to be detected. Researchers have adopted concepts from diverse disciplines such as statistics, machine learning, data mining, information theory, spectral theory, and have applied them to specific problem formulations. Figure 2 shows the above mentioned key components associated with any outlier detection technique. As mentioned earlier, a specific formulation of the problem is determined by several different factors such as the nature of the input data, the availability (or unavailability) of labels as well as the constraints and requirements induced by the application domain.

A. Nature of Input data

Input is generally a collection of data instances (also referred as object, record, point, vector, pattern, event, case, sample, observation, entity). Each data instance can be described using a set of attributes (also referred to as variable, characteristic, feature, Field, dimension). The attributes can be of different types such as binary, categorical or continuous. Each data instance might consist of only one attribute (univariate) or multiple attributes (multivariate).

B. Type of outliers

1. **Point Outliers:** If an individual data instance can be considered as anomalous with respect to the rest of data, then the instance is termed as a point outlier.
2. **Contextual Outliers:** If a data instance is anomalous in a specific context (but not otherwise), then it is termed as

a contextual outlier (also referred to as conditional outlier).

3. **Collective Outliers:** If a collection of related data instances is anomalous with respect to the entire data set, it is termed as a collective outlier.

C. Data Labels

The labels associated with a data instance denote if that instance is normal or anomalous. It should be noted that obtaining labeled data which is accurate as well as representative of all types of behaviors, is often prohibitively expensive. Labeling is often done manually by a human expert and hence requires substantial effort to obtain the labeled training data set [1],[2]. Supervised data set uses labeled objects belonging to the normal and outlier classes to learn the classifier and assign appropriate labels to test objects. Semi-supervised data set firstly learns a model denoting normal behavior from given training data set of normal objects and further calculates the likelihood of test objects. Unsupervised data set detects outliers in unlabeled data set. Considering that the most of the objects in data set are normal. This approach is applied to various kinds of outlier detection methods and data sets.

D. Output of outlier detection

An important aspect for any outlier detection technique is the manner in which the outliers are reported. Typically, the outputs produced by outlier detection techniques are one of the following two types [3]:

1. **Scores:** Scoring techniques assign an outlier score to each instance in the test data depending on the degree to which that instance is considered an outlier. Thus the output of such techniques is a ranked list of outliers.
2. **Labels:** Techniques in this category assign a label (normal or anomalous) to each test instance. Scoring based outlier detection techniques allow the analyst to use a domain specific threshold to select the most relevant outliers. Techniques that provide binary labels to the test instances do not directly allow the analysts to make such a choice, though this can be controlled indirectly through parameter choices within each technique.

II. APPLICATIONS OF OUTLIER DETECTION

A. Intrusion Detection

Intrusion detection refers to detection of malicious activity (break-ins, penetrations, and other forms of computer abuse) in a computer related system. These malicious activities or intrusions are interesting from a computer security perspective. An intrusion is different from the normal behavior of the system, and hence outlier detection techniques are applicable in intrusion detection domain.

B. Fraud Detection

Fraud detection refers to detection of criminal activities occurring in commercial organizations such as banks, credit card companies, insurance agencies, cell phone companies, stock market, etc. The malicious users might be the actual customers of the organization or might be posing as a

Study of Outlier Detection Techniques for Low and High Dimensional Data

customer (also known as identity theft). The fraud occurs when these users consume the resources provided by the organization in an unauthorized way.

C. Medical and Public Domain

Outlier detection in the medical and public health domains typically work with patient records. The data can have outliers due to several reasons such as abnormal patient condition or instrumentation errors or recording errors. Several techniques have also focused on detecting disease outbreaks in a specific area. Thus the outlier detection is a very critical problem in this domain and requires high degree of accuracy.

D. Industrial Damage Detection

Industrial units suffer damage due to continuous usage and the normal wear and tear. Such damages need to be detected early to prevent further escalation and losses. The data in this domain is usually referred to as sensor data because it is recorded using different sensors and collected for analysis. Outlier detection techniques have been extensively applied in this domain to detect such damages.

E. Image Processing

Outlier detection techniques dealing with images are either interested in any changes in an image over time (motion detection) or in regions which appear abnormal on the static image.

F. Other Domains

Outlier detection has also been applied to several other domains such as speech recognition, novelty detection in robot behavior, traffic monitoring, click through protection, detecting faults in web applications, detecting outliers in biological data, detecting outliers in census data, detecting associations among criminal activities, detecting outliers in Customer Relationship Management (CRM) data, detecting outliers in astronomical data and detecting ecosystem disturbances.

III. OUTLIER DETECTION TECHNIQUES FOR LOW DIMENSIONAL DATA

A. Classification Based Techniques

Classification is used to learn a model (classifier) from a set of labeled data instances (training) and then, classify a test instance into one of the classes using the learnt model (testing). The training phase learns a classifier using the available labeled training data. The testing phase classifies a test instance as normal or anomalous using the classifier.

1. Multi-class classification based outlier detection techniques assume that the training data contains labeled instances belonging to multiple normal classes. Such outlier detection techniques learn a classifier to distinguish between each normal class against the rest of the classes. See Figure 3(a) for illustration. A test instance is considered anomalous if it's not classified as normal by any of the classifiers.
2. One-class classification based outlier detection techniques assume that all training instances have only

one class label. Such techniques learn a discriminative boundary around the normal instances using a one-class classification algorithm, e.g., one-class SVMs, one-class Kernel Fisher Discriminants, as shown in Figure 3(b). Any test instance that does not fall within the learnt boundary is declared as anomalous.

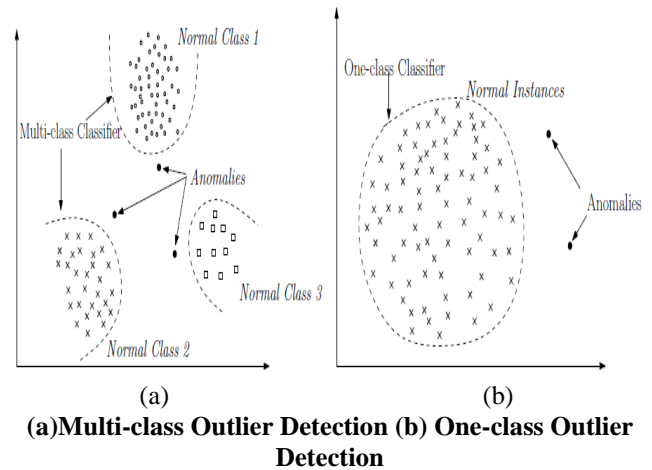


Fig3. Outlier detection using classification.

Advantages and Disadvantages of Classification Based Techniques:

1. Classification based techniques, especially the multi-class techniques, can make use of powerful algorithms that can distinguish between instances belonging to different classes.
2. The testing phase of classification based techniques is fast since each test instance needs to be compared against the pre-computed model.
3. Multi-class classification based techniques rely on availability of accurate labels for various normal classes, which is often not possible.
4. Classification based techniques assign a label to each test instance, which can also become a disadvantage when a meaningful outlier score is desired for the test instances. Some classification techniques that obtain a probabilistic prediction score from the output of a classifier can be used to address this issue.

B. Statistical Techniques

Statistical outlier detection methods [4][5] rely on the statistical approaches that assume a distribution or probability model to fit the given dataset. Under the distribution assumed to fit the dataset, the outliers are those points that do not agree with or conform to the underlying model of the data. The statistical outlier detection methods can be broadly classified into two categories, i.e., the parametric methods and the non-parametric methods.

C. Parametric Methods

Parametric statistical outlier detection methods explicitly assume the probabilistic or distribution model(s) for the given data set. Model parameters can be estimated using the training data based upon the distribution assumption. The

major parametric outlier detection methods include Gaussian model based and regression model based methods.

1. Gaussian Models

Detecting outliers based on Gaussian distribution models have been intensively studied. The training stage typically performs estimation of the mean and variance (or standard deviation) of the Gaussian distribution using Maximum Likelihood Estimates (MLE). To ensure that the distribution assumed by human users is the optimal or close-to-optima underlying distribution the data fit, statistical discordant tests are normally conducted in the test stage [4][6][7]. So far, over one hundred discordance/outlier tests have been developed for different circumstances, depending on the parameter of dataset (such as the assumed data distribution) and parameter of distribution (such as mean and variance), and the expected number of outliers [8][9].

2. Regression Models

If the probabilistic model is unknown regression can be employed for model construction. The regression analysis aims to find a dependence of one/more random variable(s) Y on another one/more variable(s) X . This involves examining the conditional probability distribution $Y|X$. Outlier detection using regression techniques are intensively applied to time-series data [10][11][12][13][14]. The training stage involves constructing a regression model that fits the data. The regression model can either be a linear or non-linear model, depending on the choice from users. The test stage tests the regression model by evaluating each data instance against the model. More specifically, such test involves comparing the actual instance value and its projected value produced by the regression model. A data point is labeled as an outlier if a remarkable deviation occurs between the actual value and its expected value produced by the regression model.

D. Non-parametric Techniques

The outlier detection techniques in this category use non-parametric statistical models, such that the model structure is not defined a priori, but is instead determined from given data. Such techniques typically make fewer assumptions regarding the data, such as smoothness of density, when compared to parametric techniques.

1. Histogram Based:

The simplest non-parametric statistical technique is to use histograms to maintain a profile of the normal data. Such techniques are also referred to as frequency based or counting based. A basic histogram based outlier detection technique for univariate data consists of two steps. The first step involves building a histogram based on the different values taken by that feature in the training data. In the second step, the technique checks if a test instance falls in any one of the bins of the histogram. If it does, the test instance is normal, otherwise it is anomalous. A variant of the basic histogram based technique is to assign an outlier score to each test instance based on the height (frequency) of the bin in which it falls. The size of the bin used when

building the histogram is key for outlier detection. If the bins are small, many normal test instances will fall in empty or rare bins, resulting in a high false alarm rate. If the bins are large, many anomalous test instances will fall in frequent bins, resulting in a high false negative rate. Thus a key challenge for histogram based techniques is to determine an optimal size of the bins to construct the histogram which maintains low false alarm rate and low false negative rate.

Advantages and Disadvantages of Statistical Techniques:

1. If the assumptions regarding the underlying data distribution hold true, statistical techniques provide a statistically just able solution for outlier detection.
2. The outlier score provided by a statistical technique is associated with a confidence interval, which can be used as additional information while making a decision regarding any test instance.
3. If the distribution estimation step is robust to outliers in data, statistical techniques can operate in a unsupervised setting without any need for labeled training data.
4. The key disadvantage of statistical techniques is that they rely on the assumption that the data is generated from a particular distribution. This assumption often does not hold true, especially for high dimensional real data sets.
5. Even when the statistical assumption can be reasonably justified, there are several hypothesis test statistics that can be applied to detect outliers; choosing the best statistic is often not a straightforward task.
6. Histogram based techniques are relatively simple to implement, but a key shortcoming of such techniques for multivariate data is that they are not able to capture the interactions between different attributes.

E. Distance Based Techniques

There have already been a number of different ways for defining outliers from the perspective of distance related metrics. Most existing metrics used for distance based outlier detection techniques are defined based upon the concepts of local neighborhood or k nearest neighbors (k NN) of the data points. The notion of distance-based outliers does not assume any underlying data distributions and generalizes many concepts from distribution-based methods. Moreover, distance-based methods scale better to multi-dimensional space and can be computed much more efficiently than the

F. Statistical based methods

1. Local Neighborhood Methods: The first notion of distance-based outliers, called DB-Outlier, is due to Knorr and Ng [15]. It is defined as follows. A point p in a data set is a $DB(k,\lambda)$ -Outlier, with respect to the parameters k and λ , if no more than k points in the data set are at a distance λ or less (i.e., λ neighborhood) from p . This definition of outliers is intuitively simple and straightforward. The major disadvantage of this method, however, is its sensitivity to the parameter λ that is difficult to specify a priori. As we know, when the data dimensionality increases, it becomes increasingly difficult to specify an appropriate circular local neighborhood (delimited by λ) for outlier-ness evaluation of

Study of Outlier Detection Techniques for Low and High Dimensional Data

each point since most of the points are likely to lie in a thin shell about any point [16]. Thus, a too small λ will cause the algorithm to detect all points as outliers, whereas no point will be detected as outliers if a too large λ is picked up. In other words, one needs to choose an appropriate λ with a very high degree of accuracy in order to find a modest number of points that can then be defined as outliers.

2. kNN-distance Methods:

There have also been a few distance-based outlier detection methods utilizing the k nearest neighbors (kNN) in measuring the outlier-ness of data points in the dataset. The first proposal uses the distance to the k th nearest neighbors of every point, denoted as D_k , to rank points so that outliers can be more efficiently discovered and ranked [17]. Based on the notion of D_k , the following definition for D_k n-Outlier is given: Given k and n , a point is an outlier if the distance to its k th nearest neighbor of the point is smaller than the corresponding value for no more than $n-1$ other points. Essentially, this definition of outliers considers the top n objects having the highest D_k values in the dataset as outliers.

Advantages and Disadvantages of Distance-based Techniques:

1. Unlike distribution-based methods, distance based methods are non-parametric and do not rely on any assumed distribution to fit the data.
2. The distance based definitions of outliers are fairly straightforward and easy to understand and implement.
3. Their major drawback is that most of them are not effective in high-dimensional space due to the curse of dimensionality, the high-dimensional data in real applications are very noisy, and the abnormal deviations may be embedded in some lower-dimensional subspaces that cannot be observed in the full data space.

G. Density-based Techniques

Density-based methods use more complex mechanisms to model the outlier-ness of data points than distance based methods. It usually involves investigating not only the local density of the point being studied but also the local densities of its nearest neighbors. Thus, the outlier-ness metric of a data point is relative in the sense that it is normally a ratio of density of this point against the averaged densities of its nearest neighbors. Density-based methods feature a stronger modeling capability of outliers but require more expensive computation at the same time.

1. LOF Method

The first major density-based formulation scheme of outlier has been proposed in [18], which is more robust than the distance-based outlier detection methods. An example is given in [18] (refer to figure 3), showing the advantage of a density-based method over the distance-based methods such as DB (k, λ)-Outlier. The dataset contains an outlier o , and $C1$ and $C2$ are two clusters with very different densities.

The $B(k, \lambda)$ -Outlier method cannot distinguish o from the rest of the data set no matter what values the parameters k and λ take. This is because the density of O 's neighborhood is very much closer to that of the points in cluster $C1$. However, the density-based method, proposed in [18], can handle it successfully.

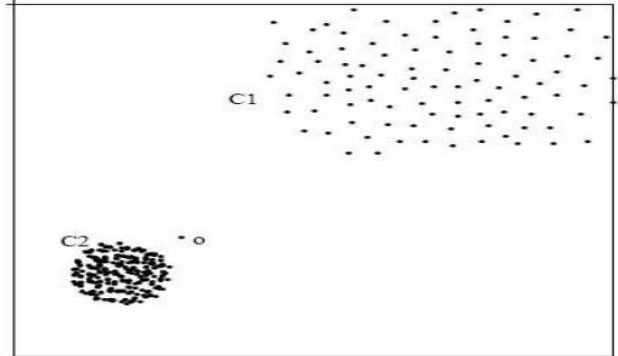


Fig4. A sample dataset showing the advantage of LOF over DB(k, λ)-Outlier.

2. COF Method:

As LOF method suffers the drawback that it may miss those potential outliers whose local neighborhood density is very close to that of its neighbors. To address this problem, Tang et al. proposed a new Connectivity based Outlier Factor (COF) scheme that improves the effectiveness of LOF scheme when a pattern itself has similar neighborhood density as an outlier [19]. In order to model the connectivity of a data point with respect to a group of its neighbors, a set-based nearest path (SBN-path) and further a set-based nearest trail (SBNtrail), originated from this data point, are defined. This SNB trail starting from a point is considered to be the pattern presented by the neighbors of this point. Based on SNB trail, the cost of this trail, a weighted sum of the cost of all its constituting edges, is computed.

3. INFLO Method

Even though LOF is able to accurately estimate outlier-ness of data points in most cases, it fails to do so in some complicated situations. For instance, when outliers are in the location where the density distributions in the neighborhood are significantly different, this may result in a wrong estimation. An example where LOF fails to have accurate outlier-ness estimation for data points has been given in [20]. The example is presented in Figure 4. In this example, data p is in fact part of a sparse cluster $C2$ which is near the dense cluster $C1$. Compared to objects q and r , p obviously displays less outlier-ness. However, if LOF is used in this case, p could be mistakenly regarded to having stronger outlier-ness than q and r . Authors in [20] pointed out that this problem of LOF is due to the inaccurate specification of the space where LOF is applied. To solve this problem of LOF, an improved method, called INFLO, is proposed [20].

4. MDEF Method

In [21], a new density-based outlier definition, called Multi-granularity Deviation Factor (MEDF), is proposed.

Intuitively, the MDEF at radius r for a point p_i is the relative deviation of its local neighborhood density from the average local neighborhood density in its r -neighborhood.

Advantages and Disadvantages of Density-based techniques

1. The density-based outlier detection methods are generally more effective than the distance-based methods. However, in order to achieve the improved effectiveness, the density based methods are more complicated and computationally expensive.
2. For a data object, they have to not only explore its local density but also that of its neighbors. Expensive kNN search is expected for all the existing methods in this category.
3. Due to the inherent complexity and non-updatability of their outlier-ness measurements used, LOF, COF, INFLO and MDEF cannot handle data streams efficiently.

H. Clustering Based Techniques

Clustering is used to group similar data instances into clusters. Clustering is primarily an unsupervised technique though semi-supervised clustering has also been explored lately. Even though clustering and outlier detection appear to be fundamentally different from each other, several clustering based outlier detection techniques have been developed. Clustering based outlier detection techniques can be grouped into three categories.

1. Normal data instances belong to a cluster in the data, while outliers either do not belong to any cluster. Techniques based on the above assumption apply a known clustering based algorithm to the data set and declare any data instance that does not belong to any cluster as anomalous. Several clustering algorithms that do not force every data instance to belong to a cluster, such as DBSCAN, ROCK, and SNN clustering can be used. The FindOut algorithm is an extension of the WaveCluster algorithm in which the detected clusters are removed from the data and the residual instances are declared as outliers.
2. Normal data instances lie close to their closest cluster centroid, while outliers are far away from their closest cluster centroid. Techniques based on the above assumption consist of two steps. In the First step, the data is clustered using a clustering algorithm. In the second step, for each data instance, its distance to its closest cluster centroid is calculated as its outlier score.
3. Normal data instances belong to large and dense clusters, while outliers either belong to small or sparse clusters.

Techniques based on the above assumption declare instances belonging to clusters whose size and/or density is below a threshold as anomalous. Several variations of the third category of techniques have been proposed. The technique proposed by called Find CBLOF, assigns an outlier score known as Cluster-Based Local Outlier Factor (CBLOF) for each data instance. The CBLOF score captures the size of the cluster to which the data instance belongs, as

well as the distance of the data instance to its cluster centroid.

Advantages and Disadvantages of Clustering Based Techniques:

1. Clustering based techniques can operate in an unsupervised mode.
2. Such techniques can often be adapted to other complex data types by simply plugging in a clustering algorithm that can handle the particular data type.
3. The testing phase for clustering based techniques is fast since the number of clusters against which every test instance needs to be compared is a small constant.
4. Performance of clustering based techniques is highly dependent on the effectiveness of clustering algorithm in capturing the cluster structure of normal instances.
5. Many techniques detect outliers as a by-product of clustering, and hence are not optimized for outlier detection.

I. Information theoretic outlier detection techniques

Information theoretic techniques[2] analyze the information content of a data set using different information theoretic measures such as Kolomogorov Complexity, entropy, relative entropy, etc. Such techniques are based on the following key assumption: Outliers in data induce irregularities in the information content of the data set. Let $C(D)$ denote the complexity of a given data set, D . A basic information theoretic technique can be described as follows. Given a data set D , Find the minimal subset of instances, I , such that $C(D)-C(D-I)$ is maximum. All instances in the subset thus obtained, are deemed as anomalous. The problem addressed by this basic technique is to find a pareto-optimal solution, which does not have single optima, since there are two different objectives that need to be optimized. Several techniques have been proposed that perform approximate search for the most anomalous subset.

This uses an approximate algorithm called Local Search Algorithm (LSA) to approximately determine such a subset in a linear fashion, using entropy as the complexity measure. A similar technique that uses an information bottleneck measure was proposed by [Ando 2007]. Information theoretic techniques have also been used in data sets in which data instances are naturally ordered, e.g., sequential data, spatial data. In such cases, the data is broken into substructures (segments for sequences, subgraphs for graphs, etc.), and the outlier detection technique finds the substructure, I , such that $C(D)-C(D-I)$ is maximum. This technique has been applied to sequences, graph data and spatial data. A key challenge of such techniques is to find the optimal size of the substructure which would result in detecting outliers.

Advantages and Disadvantages of information theoretic technique

1. They can operate in an unsupervised setting.
2. They do not make any assumptions about the underlying statistical distribution for the data.

Study of Outlier Detection Techniques for Low and High Dimensional Data

3. The performance of such techniques is highly dependent on the choice of the information theoretic measure. Often, such measures can detect the presence of outliers only when there is significantly large number of outliers present in the data.
4. Information theoretic techniques applied to sequences and spatial data sets rely on the size of the substructure, which is often nontrivial to obtain.
5. It is difficult to associate an outlier score with a test instance using an information theoretic technique.

IV. OUTLIER DETECTION TECHNIQUES FOR HIGH DIMENSIONAL DATA

There are many applications in high-dimensional domains in which the data can contain dozens or even hundreds of dimensions. The outlier detection techniques we have reviewed in the preceding sections use various concepts of proximity in order to find the outliers based on their relationship to the other points in the data set.

A. Sparse Cube Method

Aggarwal et al. conducted some pioneering work in high-dimensional outlier detection [22]. They proposed a new technique for outlier detection that finds outliers by observing the density distributions of projections from the data. This new definition considers a point to be an outlier if in some lower-dimensional projection it is located in a local region of abnormally low density. Therefore, the outliers in these lower-dimensional projections are detected by simply searching for these projections featuring lower density. To measure the sparsely of a lower-dimensional projection quantitatively, the authors proposed the so called Sparsely Coefficient. The computation of Sparsely Coefficient involves a grid discretization of the data space and making an assumption of normal distribution for the data in each cell of the hypercube. Each attribute of the data is divided into ϕ equi-depth ranges. In each range, there is a fraction $f = 1/\phi$ of the data. Then, a k -dimensional cube is made of ranges from k different dimensions. Let N be the dataset size and $n(D)$ denote the number of objects in a k -dimensional cube D .

B. Example-based Method

Recently, an approach using outlier examples provided by users are used to detect outliers in high-dimensional space [22]. It adopts an outlier examples subspaces outliers' manner to detect outliers. Specifically, human users or domain experts first provide the systems with a few initial outlier examples. The algorithm finds the subspaces in which most of these outlier examples exhibit significant outlier-ness. Finally, other outliers are detected from these subspaces obtained in the previous step. This approach partitions the data space into equi-depth cells and employs the Sparsely Coefficient proposed in [22] to measure the outlier-ness of outlier examples in each subspace of the lattice. Since it is untenable to exhaustively search the space lattice, the author also proposed to use evolutionary algorithms for subspace search. The fitness of a subspace is

the average Sparsely Coefficients of all cubes in that subspace to which the outlier examples belong. All the objects contained in the cubes which are sparser than or as sparse as cubes containing outlier examples in the subspace are detected as outliers.

C. HighDoD

Zhang et al. proposed a novel dynamic subspace search algorithm, called HighDoD, to efficiently identify the outlying subspaces for the given query data points. The outlying measure, OD, is based on the sum of distances between a data and its k nearest neighbors [23]. This measure is simple and independent of any underlying statistical and distribution characteristics of the data points. The following two heuristic pruning strategies employing upward-and downward closure property are proposed to aid in the search for outlying subspaces: If a point p is not an outlier in a subspace s , then it cannot be an outlier in any subspace that is a subset of s . If a point p is an outlier in a subspace s , then it will be an outlier in any subspace that is a superset of s . These two properties can be used to quickly detect the subspaces in which the point is not an outlier or the subspaces in which the point is an outlier.

D. SOF Method

In [24], a novel technique based on genetic algorithm is proposed to solve the outlying subspace detection problem and well copes with the drawbacks of the existing methods. A new metric, called Subspace Outlying Factor (SOF), is developed for measuring the outlying degree of each data point in different subspaces. Based on SOF, a new definition of outlying subspace, called SOF Outlying Subspaces, is proposed. Given an input dataset D , parameters n and k , a subspace s is a SOF Outlying Subspace for a given query data point p if there are no more than $n-1$ other subspaces s' such that $SOF(s'; p) > SOF(s; p)$. The above definition is equivalent to say that the top n subspaces having the largest SOF values are considered to be outlying subspaces.

E. Clustering Algorithms

1. **CLIQUE:** CLIQUE [22] is a grid-based clustering method that discretizes the data space into non-overlapping rectangular units, which are obtained by partitioning every dimension into a specific number of intervals of equal length. A unit is dense if the fraction of total data points contained in this unit is greater than a threshold. Clusters are defined as unions of connected dense units within a subspace. CLIQUE first identifies a subspace that contains clusters. A bottom-up algorithm is used that exploits the monotonicity of the clustering criterion with respect to dimensionality: if a k -dimensional unit is dense, then so are its projections in $(k-1)$ -dimensional space. A candidate generation procedure iteratively determines the candidate k -dimensional units C_k after determining the $(k-1)$ -dimensional dense units D_{k-1} . A pass is made over the data to determine those candidates units that are dense D_k . A depth-first search algorithm is then used to identify clusters

in the subspace: it starts with some unit u in D , assign it the first cluster label number, and find all the units it is connected to. Then, if there are still units in D that have yet been visited, it finds one and repeats the procedure. CLIQUE is able to automatically find dense clusters in subspaces of high-dimensional dataset.

2. HPStream: In order to find the clusters embedded in the subspaces of high-dimensional data space in data streams, a new clustering method, called HPStream, is proposed [9]. HPStream introduces the concept of projected clustering to data streams as significant and high-quality clusters only exist in some lowdimensional subspaces. The basic idea of HPStream is that it does not only find clusters but also updates the set of dimensions associated with each cluster where more compact clusters can be found. The total number of clusters obtained in HPStream is initially obtained through k-means clustering and the initial set of dimensions associated with each of these k clusters is the full set of dimensions of the data stream. As more streaming data arrive, the set of dimensions for each cluster evolves such that each cluster can become more compact with a smaller radius.

F. ITB-SP Method

In this a formal definition of outliers and an optimization model of outlier detection are defined, via a new concept of holoentropy that takes both entropy and total correlation into consideration. Based on this model [1], a function for the outlier factor of an object which is solely determined by the object itself and can be updated efficiently is determined. Here two practical 1-parameter outlier detection methods, named ITB-SS and ITB-SP are used, which require no user-defined parameters for deciding whether an object is an outlier. Users need only provide the number of outliers they want to detect. Experimental results show that ITB-SS and ITB-SP are more effective and efficient than mainstream methods and can be used to deal with both large and high-dimensional data sets where existing algorithms fail. Outlier detection for categorical data sets. This problem is especially challenging because of the difficulty of defining a meaningful similarity measure for categorical data.

V. SUMMARY

This section presents a comprehensive survey on the major existing methods for detecting point outliers from vector-like data sets. Both the conventional outlier detection methods that are mainly appropriate for relatively low dimensional static databases and the more recent methods that are able to deal with high-dimensional projected outliers or data stream applications have been discussed. For a big picture of these methods, we present a summary in Table 1. In this table, we evaluate each method against two criteria, namely whether it can detect projected outliers in a high-dimensional data space and whether it can handle data streams. The symbols of tick and cross in the Table 1 indicate respectively whether or not the corresponding method satisfies the evaluation criteria. From this table, we can see that the conventional outlier detection methods

cannot detect projected outliers embedded in different subspaces; They detect outliers only in the full data space or a given subspace. Amongst these methods that can detect projected outliers, only HPStream can meet both criteria. However, being a clustering method, HPStream cannot provide satisfactory support for projected outlier detection from high-dimensional data streams.

Table1. A Summary of Major Existing Outlier Detection Methods

Category	Method	Low Dimensional Data	High Dimensional Data	Data Stream
Classification Based Method	Multiclass	✓	✗	✗
	SVMs	✓	✗	✗
	Kernel Fisher Discriminants	✓	✗	✗
Statistical Methods	Gaussian Models	✓	✗	✗
	Regression Models	✓	✗	✓
	Histograms	✓	✗	✓
	Kernel Functions	✓	✗	✓
Distance Based Methods	DB(k,λ)-outliers	✓	✗	✗
	DB(pct,dmin)-outliers	✓	✗	✗
	KNN method	✓	✗	✗
	KNN sum method	✓	✗	✗
Density Based Methods	Grid ODF	✓	✗	✗
	LOF	✓	✗	✗
	COF	✓	✗	✗
	INFLO	✓	✗	✗
	MDEF	✓	✗	✗
Clustering Based Methods	Incremental LOF	✓	✗	✓
	PAM/CLARA	✓	✗	✗
	CLARANS	✓	✗	✗
	BIRCH	✓	✗	✓
	CURE	✓	✗	✗
	CLQUE	✓	✓	✗
	HPSTREAM	✓	✓	✓
	K-means	✓	✗	✗
	MST clustering	✓	✗	✗
	DBSCAN	✓	✗	✗
	STING	✓	✗	✓
	DCLUS	✓	✗	✓
STREAM	✓	✗	✓	
CLUSTREAM	✓	✗	✓	
Information Theoretic Based Methods	Scalable minimization	✓	✗	✗
	LSA	✓	✗	✗
	ITB-SP	✓	✓	✗
	FPOF	✓	✓	✗
Other	Sparse cube	✓	✓	✗
	Example-based	✓	✓	✗
	High DoD	✓	✓	✗
	SOF	✓	✓	✗

VI. CONCLUSION

In this paper, a comprehensive study is presented to review the existing methods for detecting point outliers from various kinds of vector-like datasets. The outlier detection techniques that are primarily suitable for relatively low-dimensional static data, which serve the technical foundation for many of the methods proposed later, are reviewed first. We have also reviewed some of recent

Study of Outlier Detection Techniques for Low and High Dimensional Data

advancements in outlier detection for dealing with more complex high-dimensional static data and data streams. It is important to be aware of the limitation of this survey. As it has clearly stated in Section 2, we only focus on the point outlier detection methods from vector-like datasets due to the space limit. Also, outlier detection is a fast developing field of research and more new methods will quickly emerge in the foreseeable near future. Driven by their emergence, it is believed that outlier detection techniques will play an increasingly important role in various practical applications where they can be applied.

VII. REFERENCES

- [1] Shu Wu, Member IEEE, and Shengrui Wang, Member IEEE, "Information-Theoretic Outlier Detection for Large-Scale Categorical Data," *IEEE transactions on knowledge and data engineering*, vol. 25, no. 3, march.
- [2] V. Chandola, A. Banerjee, and V. Kumar, "Outlier Detection: A Survey," *ACM Computing Surveys*, vol. 41, no. 3, pp. 1-58, 2009.
- [3] V.J. Hodge and J. Austin, "A Survey of Outlier Detection Methodologies," *Artificial Intelligence Rev.*, vol. 22, no. 2, pp. 85-126, 2004.
- [4] V. Barnett and T. Lewis, "Outliers in Statistical Data", John Wiley, 3rd edition, 1994.
- [5] D. Hawkins, "Identification of Outliers", Chapman and Hall, London, 1980.
- [6] V. Barnett, "The ordering of multivariate data", *Journal of the Royal Statistical Society. Series A* 139, 318-354, 1976.
- [7] R. J. Beckman and R. D. Cook. *Outliers. Technometrics* 25, 2, 119-149, 1983.
- [8] J. Han and M Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufman Publishers, 2000.
- [9] E. M. Knorr and R. T. Ng. Algorithms for Mining Distance-based Outliers in Large Dataset. In Proc. Of 24th International Conference on Very Large Data Bases (VLDB'98), New York, NY, pp 392-403, 1998.
- [10] B. Abraham and A. Chuang. Outlier detection and time series modeling. *Technometrics* 31, 2, 241-248, 1989.
- [11] B. Abraham and G. E. P. Box. Bayesian analysis of some outlier problems in time series. *Biometrika* 66, 2, 229-236, 1979.
- [12] A. J. Fox. Outliers in time series. *Journal of the Royal Statistical Society, Series B (Methodological)* 34, 3, 350-363, 1972.
- [13] C. C. Aggarwal. On Abnormality Detection in Spuriously Populated Data Streams. *SIAM International Conference on Data Mining (SDM'05)*, Newport Beach, CA, 2005.
- [14] X. Li and J. Han: Mining Approximate Top-K Subspace Outliers in Multi-Dimensional Time-Series Data. *VLDB*, 447-458, 2007.
- [15] E. M. Knorr and R. T. Ng. Algorithms for Mining Distance-based Outliers in Large Dataset. In Proc. Of 24th International Conference on Very Large Data Bases (VLDB'98), New York, NY, pp 392-403, 1998.
- [16] K. Beyer, J. Goldstein, R. Ramakrishnan and U. Shaft. When is nearest neighbors meaningful? In Proc. of 7th International Conference on Database Theory (ICDT'99), pp 217-235, Jerusalem, Israel, 1999.
- [17] S. Ramaswamy, R. Rastogi, and S. Kyuseok. Efficient Algorithms for Mining Outliers from Large Data Sets. In Proc. of 2000 ACM SIGMOD International Conference on Management of Data (SIGMOD'00), Dallas, Texas, pp 427-438, 2000.
- [18] 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.
- [19] J. Tang, Z. Chen, A. Fu, and D. W. Cheung. Enhancing Effectiveness of Outlier Detections for Low Density Patterns. In Proc. of 6th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'02), Taipei, Taiwan, 2002.
- [20] W. Jin, A. K. H. Tung, J. Han and W. Wang: Ranking Outliers Using Symmetric Neighborhood Relationship. *PAKDD'06*, 577-593, 2006.
- [21] S. Papadimitriou, H. Kitagawa, P. B. Gibbons and C. Faloutsos: LOCI: Fast Outlier Detection Using the Local Correlation Integral. *ICDE'03*, 315, 2003.
- [22] Ji Zhang, "Advancements of Outlier Detection: A Survey," *ICST Transactions on Scalable Information Systems*, Vol.13, no. 1, March 2013.
- [23] F. Angiulli and C. Pizzuti. Fast Outlier Detection in High Dimensional Spaces. In Proc. of 6th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'02), Helsinki, Finland, pp 15-26, 2002.
- [24] J. Zhang, Q. Gao and H. Wang. A Novel Method for Detecting Outlying Subspaces in High-dimensional Databases Using Genetic Algorithm. 2006 IEEE International Conference on Data Mining (ICDM'06), pages 731- 740, Hong Kong, China, 2006.