IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

SURVEY OF RECENT ADVANCES IN OUTLIER DETECTION TECHNIQUES USING DATA MINING

[1]Pranali K.Bhowate,

[1]Assistant Professor
[1]Department of Computer science and Engineering,
[1]P.R.Pote (Patil) Institute of Engineering and Research, Amravati,India

Abstract: Outlier detection is currently very active area of research in data set mining community. Finding outlier in a collection of patterns is a very well-known problem in the data mining field. An outlier is a pattern which is dissimilar with respect to the rest of the patterns in the dataset. In this paper, I present a survey of outlier detection techniques to reflect the recent advancements in this field. The survey will not only cover the traditional outlier detection methods for static and low dimensional datasets but also for dynamic/streaming and high-dimensional datasets.

***Index Terms* - Data Mining, Outlier Detection, Low-dimensional datasets, High-dimensional Datasets**

I. INTRODUCTION

Outlier is a point of data that does not belongs to group of data also it is a data point that does not conform to the normal points characterizing the data set [1].Outlier detection is an integral part of data mining and has attracted much attention recently [2].It is very vigorous problem to Find anomalous points among the data points is the basic idea to find out an outlier. Outlier detection signals out the objects mostly deviating from a given data set.

Data mining, in general, deals with the discovery of non-trivial, hidden and interesting knowledge from different types of data. With the development of information technologies, the number of databases, as well as their dimension and complexity, grow rapidly. It is necessary what we need automated analysis of great amount of information. The analysis results are then used for making a decision by a human or program. One of the basic problems of data mining is the outlier detection [4].

Outlier detection as a branch of data mining has many important applications and deserves more attention from data mining community [3]. Detecting outliers has important applications in data cleaning as well as in the mining of abnormal points for fraud detection, stock market analysis, intrusion detection, marketing, network sensors [1].

Outliers may be erroneous or real in the following sense. Real outliers are observations whose actual values are very different than those observed for the rest of the data and violate plausible relationships among variables. Erroneous outliers are observations that are distorted due to misreporting or miss recording errors in the data-collection process. Outliers of either type may exert undue influence on the results of statistical analysis, so they should be identified using reliable detection methods prior to performing data analysis [4].

For example, outlier detection can help identify suspicious fraudulent transaction for credit card companies. It can also be utilized to identify abnormal brain signals that may indicate the early development of brain cancers. Due to its inherent importance in various areas, considerable research efforts in outlier detection have been conducted in the past decade. A number of outlier detection techniques have been proposed that use different mechanisms and algorithms [5]. This paper presents a comprehensive review on the major techniques of outlier detection.

The datasets could be static with a small number of attributes where outlier detection is relatively easy. Nevertheless, the datasets could also be dynamic, such as data streams, and at the same time have a large number of attributes. Dealing with this kind of datasets is more complex by nature and requires special attentions to the detection the methods to be developed.

The scope of this survey will be clearly specified as first will review the conventional outlier detection techniques that are suitable for low-dimensional data, followed by some outlier detection techniques for high-dimensional data.

II. APPLICATIONS DOMAIN

Quality control applications
Financial applications
Web log analytics
Intrusion detection applications
Medical applications

Text and social media applications
Earth science applications

.

III. CLASSIFICATION OF OUTLIERS

There have been a lot of research work in detecting different kinds of outliers from various types of data where the techniques outlier detection methods utilize differ considerably. In the reminder of this subsection, we will discuss briefly different types of outliers. First, outliers can be classified as point outliers and collective outliers based on the number of data instances involved in the concept of outliers.

3.1.Point outliers. In a given set of data instances, an individual outlying instance is termed as a point outlier. This is the simplest type of outliers and is the focus of majority of existing outlier detection schemes [6]. A data point is detected as a point outlier because it displays outlier-ness at its own right, rather than together with other data points. the extent to which each single data is deviated from the other data in the data set.

3.2.Collective outliers. A collective outlier represents a collection of data instances that is outlying with respect to the entire data set. The individual data instance in a collective outlier may not be outlier by itself, but the joint occurrence as a collection is anomalous [6]. Usually, the data instances in a collective outlier are related to each other.

Outliers can also be categorized into vector outliers, sequence outliers, trajectory outliers and graph outliers, etc, depending on the types of data from where outliers can be detected.

3.3.Vector outliers. Vector outliers are detected from vector-like representation of data such as the relational databases[5]. The data are presented in tuples and each tuple has a set of associated attributes. The data set can contain only numeric attributes, or categorical attributes or both. Based on the number of attributes, the data set can be broadly classified as low-dimensional data and high-dimensional data, even though there is not a clear cutoff between these two types of data sets. As relational databases still represent the mainstream approaches for data storage, therefore, vector outliers are the most common type of outliers we are dealing with.

3.4. Sequence outliers. In many applications, data are presented as a sequence. A good example of a sequence database is the computer system call log where the computer commands executed, in a certain order, are stored. A sequence of commands in this log may look like the following sequence: http-web, buffer-overflow, http-web, http-web, smtpmail, ftp, http-web, ssh. Outlying sequence of commands may indicate a malicious behaviour that potentially compromises system security. In order to detect abnormal command sequences, normal command sequences are maintained and those sequences that do not match any normal sequences are labeled sequence outliers. Sequence outliers are a form of collective outlier.[5]

3.5.Trajectory outliers. Recent improvements in satellites and tracking facilities have made it possible to collect a huge amount of trajectory data of moving objects. Examples include vehicle positioning data, hurricane tracking data, and animal movement data [7]. Unlike a vector or a sequence, a trajectory is typically represented by a set of key features for its movement, including the coordinates of the starting and ending points; the average, minimum, and maximum values of the directional vector; and the average, minimum, and maximum velocities Based on this representation, a weighted-sum distance function can be defined to compute the difference of trajectory based on the key features for the trajectory. A more recent work proposed a partition-and-detect framework for detecting trajectory outliers [7]. The idea of this method is that it partitions the whole trajectory into line segments and tries to detect outlying line segments, rather than the whole trajectory. Trajectory outliers can be point outliers if we consider each single trajectory as the basic data unit in the outlier detection. However, if the moving objects in the trajectory are considered, then an abnormal sequence of such moving objects (constituting the sub-trajectory) is a collective outlier.[5]

Graph outliers. Graph outliers represent those graph entities that are abnormal when compared with their squint. The graph entities that can become outliers include nodes, edges and sub-graphs.

IV. OUTLIER DETECTION METHODS FOR LOW DIMENSIONAL

This focus on outlier detection methods for low Dimensional data. These broadly classified as they used, i.e., Distance based outlier detection ,Clustering based outlier detection, Density based outlier detection , Depth based outlier detection. Each of these techniques has its own advantages and disadvantages. In general, in all these methods, the technique to detect outliers consists of two steps. The first identifies an outlier around a data set using a set of inliers (normal data). In the second step, a data request is analyzed and identified as outlier when its attributes are different from the attributes of inliers. All these techniques assume that all normal instances will be similar, while the anomalies will be different.

4.1.Distance based outlier detection:- There have already been a number of different ways for defining outliers from the perspective of distance related metrics. Most existing metrics used for distancebased outlier detection techniques are defined based upon the concepts of local neighborhood or k nearest neighbors (kNN) 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 statistical-based methods. In distance-based methods, distance between data points is needed to be computed.

4.2.Clustering based outlier detection Clustering methods like DBSCAN, BIRCH and CURE may detect outliers. However, since the main objective of a clustering method is to find clusters, they are developed to optimize clustering, and not to optimize outlier detection. The definition of outlier used is subjective to the clusters that are detected by these algorithms. While definitions of distance-based outliers are more objective and independent of how clusters in the input data are identified [8].

4.3. Density based outlier detection 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 [5].Density-based methods feature a stronger modelling capability of outliers but require more expensive computation at the same time. What will be discussed in this subsection are the major density-based methods called LOF method, COF method, INFLO method and MDEF method.

4.4.Depth based outlier detection Search for outliers at the border of the data space but independent of the data space but independent of statistical distributions – Organize data objects in convex hull layers – Outliers are objects on outer layers. Outliers are located at the border of the data space– Normal objects are in the center of the data space. Algorithm used are ISODEPTH,FDC.

V. Outlier Detection Methods for high Dimensional.

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. However, in high-dimensional space, the data are sparse and concepts using the notion of proximity fail to achieve most of their effectiveness. This is due to the curse of dimensionality that renders the high-dimensional data tend to be equi-distant to each other as dimensionality increases. They does not consider the outliers embedded in subspaces and are not equipped with the mechanism for detecting them.[5] Challenges Curse of dimensionality Relative contrast between distances decreases with increasing dimensionality Data is very sparse, almost all points are outliers Concept of neighborhood becomes meaningless. Solutions Use more robust distance functions and find full dimensional outliers Find outliers in projections (subspaces) of the original feature space

Methods for detecting the outlier in high-dimensional data.

5.1.Sparse Cube Method. Aggarwal et al. conducted some pioneering work in high-dimensional outlier detection [9][10]. 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.[5] Therefore, the outliers in these lower-dimensional projections are detected by simply searching for these projections featuring lower density. To measure the sparsity of a lower-dimensional projection quantitatively, the authors proposed the so-called Sparsity Coefficient.

5.2.Example-based Method. Recently, an approach using outlier examples provided by users are used to detect outliers in high-dimensional space It adopts an 00outlier examples ! subspaces ! outliers00 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 Sparsity Coefficient proposed in [9] 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 Sparsity Coefficients of all cubes in that subspace to which the outlier examples belong.[5] 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.

5.3.Outlier Detection in Subspaces. Since outlier-ness of data points mainly appear significant in some subspaces of moderate dimensionality in high-dimensional space and the quality of the outliers detected varies in different subspaces consisting of different combinations of dimension subsets. It uses a wrapper algorithm in which the dimension subsets are selected such that the quality of outlier detected or the clusters generated can be optimized.[5] The originality of this work is to combine the evolutionary algorithm with the data visualization technique utilizing parallel coordinates to present evolution results interactively and allow users to actively participate in evolutionary algorithm searching to achieve a fast convergence of the algorithm.

VI. Conclusions

In this paper, a comprehensive survey is presented to review the existing methods for detecting outliers from various kinds of datasets. The outlier detection techniques that are primarily suitable for relatively low-dimensional data. We have also reviewed some of recent advancements in outlier detection for dealing with more complex high-dimensional data. 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 applications.

REFERENCES

[1] Rajendra pamula,Jatindra Kumar Deka,Sukumar Nandi,” *An Outlier Detection Method based on clustering*”, Second international conference On Emerging application of information technology,2011.

[2]Murugavel. P. et al, *“Improved Hybrid Clustering And Distance-Based Technique for Outlier Removal”,* International Journal on Computer Science and Engineering (IJCSE), 1 JAN 2011

[3]Ms. S. D. Pachgade, Ms. S. S. Dhande*,”Outlier Detection over Data Set Using Cluster-Based and Distance-Based Approach”*,International Journal of Advanced Research in Computer Science and Software Engineering,Volume 2, Issue 6, June 2012

[4] Svetlana Cherednichenko, *“Outlier Detection in Clustering*”,2005

[5] Ji Zhang *, “ Advancements of Outlier Detection: A Survey*”, ICST Transactions on Scalable Information Systems, 04 February 2013

[6]. 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.

[7]. J. Lee, J. Han and X. Li. “*Trajectory Outlier Detection: A Partition-and-Detect Framework*”. ICDE’08, 140-149, 2008.

[8]. Pranjali Kasture, Jayant Gadge,” *Cluster based Outlier Detection*”, International Journal of Computer Applications (0975 – 8887) Volume 58– No.10, November 2012

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

[10] Charu C. Aggarwal and Philip S. Yu. 2005. “*An effective and efficient algorithm for high-dimensional outlier detection*.” VLDB Journal, 14: 211-221, Springer-Verlag Publisher..