ISSN No. 0976-5697

Volume 6, No. 3, May 2015 (Special Issue)

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

A Survey for Outlier Detection and its Strategies

Ch. Nagamani
Research scholar,
Department of Computer Science
Nagarjuna university
Andhra Pradesh
India.
nagamani502@gmail.com

Dr. Ch. Suneetha
Associate Professor
Department of Computer Applications
RVR&JC college of Engineering &Technology
Andhra Pradesh
India.
suneethachittineni@gmail.com

Abstract: Outlier detection is the most important research problem in data mining that aims to detect outliers from high volumes of data. The Outlier detection problem has sophisticated applications in the field of Fraud detection for Credit cards, Military supervision for enemy activities, E-mail spam detection etc. Most such applications are high dimensional domains in which the data can contain hundreds of dimensions . Most approaches use the concept of proximity in order to find outliers based on relationship to the rest of the data. But it fails when data comes with high dimensions. In order to find out those outliers, we introduce a survey of sophisticated techniques for outlier detection. In this paper, we identified a well defined mechanisms to handle outliers, their motivations and distinguish them.

Keywords: Outlier detection, Proximity, Data mining, High Dimensionality, Fraud detection.

I. INTRODUCTION

Outlier detection is a very important concept in the data mining. It is useful in data analysis for Decision Support Systems(DSS).The definition of Grubbs (Grubbs, 1969) and quoted in Barnett & Lewis (Barnett and Lewis, 1994)[15]:

An outlying observation, or outlier [1], is one that appears to deviate markedly from other members of the sample in which it occurs.

A further outlier definition from Barnett & Lewis
(Barnett and Lewis, 1994) is:

An observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data.

Aggarwal (Aggarwal and Yu, 2001) notes that outliers may be considered *as noise points [2] lying outside a set of defined clusters or alternatively outliers may be defined as the points that lie outside of the set of clusters but are also separated from the noise. These outliers behave differently from the norm.*

In Real databases, outliers may indicate fraudulent cases or they may just denote an error by the entry clerk or a misinterpretation of a missing value code, either way detection of the anomaly is vital for data base consistentecy. All noise points are not becomes the outliers. Some data points are very different from normal objects.

1.1 List of applications that utilize outlier detection is:

-Fraud detection [4] **:** The purchasing order of people who steal credit cards may be different from that of the owners of the cards. By this way we can easily detect the outliers. The identification of such buying pattern changes could effectively prevent fraudulent from a long period of fraud activity. Similar approaches can also be used for other kinds of commercial fraud such as in mobile phones, insurance claim, financial transactions etc .

- *Intrusion detection[2]* : It is similar to Fraud detection but it applies on network based computing systems. In that systems, frequent attacks may result in systems being disabled, even completely crashed . The identification of such intrusions could find out malicious programs in computer operating system and also detect unauthorized access with malicious intentions to computer network systems and so effectively snoop out that hackers.

-*Environmental Change[4]*:. Many unexpected events that occur in the natural environment such as a typhoon, floods, drought and fire, often have an adverse impact on the normal life of human beings. The identification of that typical behaviors could accurately predict whether it is being or not. And it is helpful to take effective measures on time.

- *Medical & public health diagnosis[4]*: The health records of Patient with unusual symptoms or test results may indicate potential health problems for a patient. The identification of such unusual records could be useful to know whether the the patient really has potential diseases and so take effective medical treatment in time.

-*Locating and Tracking*: When users visiting a particular place, they need to locate a particular place and track a destination. It is a known fact that raw data may contain error, which makes localization results not accurate and useful. Filtering such erroneous data could improve the estimation of the location of objects and make tracking easier.

© 2015-2019, IJARCS All Rights Reserved

CONFERENCE PAPER
4th National Conference on Recent Trends in Information Technology 2015 on 25/03/2015
Organized by Dept. of IT, Prasad V. Potluri Siddhartha Institute of Technology, Kanuru, Vijayawada-7 (A.P.) India

1.2 Challenges in Outlier Detection:

According to Charu. C. Agarwal[2] the major challenge is to Identifying and analyzing the unseen area in outlier detection. An outlier is a pattern that does not conform to expected normal behavior. So the simple approach is to define an area indicating normal behavior and declare any data object which does not belong to this normal area as an outlier. But there are several challenges in implementing this approach.

- To Define a normal behavior considering every possible behavior pattern is very difficult to analyze the data.
- The currently defined normal behavior may not correctly represent the normality in future as sometimes the normal behavior keeps evolving.
- Sometimes there is a thin boundary layer between the outlying and normal behavior of data objects. So detecting outliers with maximum efficacy.
- The behavior of an outlier are different for different applications. Every application has its own set of requirements and constraints. So , it is crucial task to choose an elegant approach to detect the outliers.

By considering the above mentioned challenges, a detailed specification of the problem is required in order to detect the outliers in specified application domain. Because development of a framework based on a general idea about the problem is a difficult task.

1.3 Related work:

Many works have been published on outlier detection in various application systems such as Database systems, Network systems, Web based systems etc. To detect the outliers, most researchers published various papers and defined a taxonomy for anomalies found through outlier detection, while some other papers make mention of work conducted on fraud detection for credit cards and cellular phones by the use of various techniques such as Descriptive and Predictive techniques. Recently, J.B. Macqueen published a paper on Clustering sentence – level text using fuzzy relational clustering algorithm for outlier detection . This strategy consists of unsupervised learning, during which data are automatically assigned to one of several clusters according to certain shared characteristics. A tuple is more likely to be tagged as an outlier the further it falls from the rest of the sample. Several clustering techniques are available, including the following[7]:

– Hierarchical clustering, which produces a hierarchy of clusters within the dataset. This technique's results are usually presented in a dendrogram.

– Partitioning methods[7], in which the dataset is successively partitioned. Objects are clustered into different groups and therefore each object's deviation from the cluster's centers must be kept to a minimum.

– Density-based clustering[10], in which clusters are defined by object density. Objects in low density regions are considered anomalous.

Other clustering procedures include Fuzzy, Neural networking, evolutionary algorithms, and Entropy-based methods etc.

Knorr and Ng[9] were introduced Distance based outlier techniques . Charu C, Agarwal proposed various Cluster based approaches for outlier detection. H.D kuna[9] proposed a hybrid approach for outlier detection in audit logs for application systems. Alessia and sankar paul proposed a rough set approach for detecting outliers[10]. Later most of the techniques applied on outliers by researchers.

II. DATA MINING APPLIED TO OUTLIER DETECTION:

Data mining[7] is defined as a process of extracting useful information from large databases. As a step of KDD process which produces a knowledge for analysis. Currently, data mining plays a major role in outlier detection. It is composed of a wide selection of techniques that use several types of algorithm to classify and define outliers according to their specific characteristics, as stated by Zhang, among others. We must note that these techniques have evolved in both efficiency and efficacy.

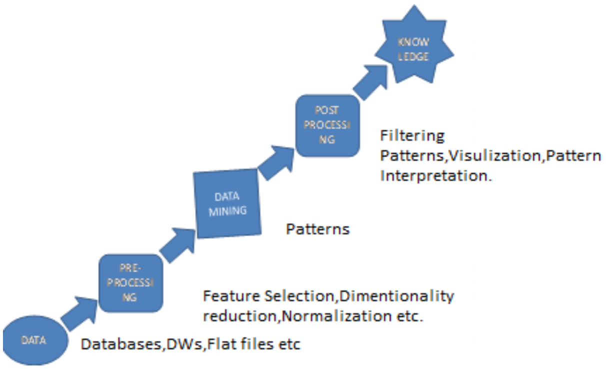

KNOW LEDGE
POST PROCESSI NG
Filtering Patterns,Visulization,Pattern Interpretation.
DATA MINING
Patterns
PRE-PROCESSI NG
Feature Selection,Dimentionality reduction,Normalization etc.
DATA
Databases,DWs,Flat files etc

Figure: Data mining as a step of KDD process.

Defining Outliers(Kamber) :

Outliers[6] are nothing but patterns in data that do not conform to a well defined notion of normal behavior. Figure illustrates outliers in 2-dimensional data set. The data has two normal regions, R1 ,R2 and R3, since most observations lie in these three regions. Data Points that are far away from the regions, e.g., points O1 and O2, and points in region R3, are outliers. X Y R1 R2 O1 O2 R3.

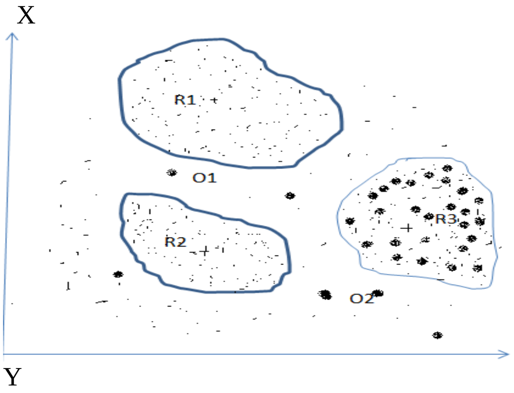


Figure 1. O utliers in a 2-dimensional data set.

outliers might be induced in the data for a variety of reasons, such as malicious activity, e.g., credit card fraud, cyber-intrusion, terrorist activity or breakdown of a system, but the common point of all is that they are interesting to the analyst. The “interestingness” or real life relevance of outliers is a key feature of outlier detection.

Forms of Outliers: Outlier may comes in different forms [4]:

Global outliers- It is defined as Observations inconsistent with the rest of the dataset.

Local outliers- It is defined as Observations inconsistent with their neighborhoods only.

III. DIFFICULTIES IN OUTLIER DETECTION

Outliers[13] are nothing but patterns that are different from normal behavior, which in its simplest form could be represented by a region and visualize all normal observations to belong to this normal region and consider the rest as outliers. Sometimes normal data objects can be appered as outliers, hence the users have a challenge to identify those outliers data from given dataset. This outlier detection approach looks simple but is highly challenging due to following reasons.

Sometimes , it is very difficult to define the normal behavior or a normal region. The difficulties are as under.

- Encircling of every possible normal behavior in the region,
i.e every time it needs to maintain Indefinite boundary between normal and outlier behavior. since at times outlier observation lying close to the boundary could be normal, and vice-versa.
- In addition to that, Adaptation of malicious adversaries to make the outlier observations appear like normal when outliers result from abnormal actions.

IV. TECHNIQUES APPLIED TO HANDLE OUTLIERS

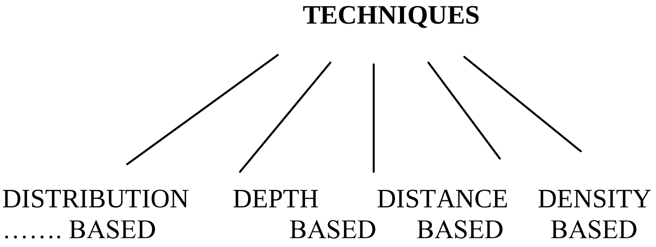
There are THREE fundamental approaches to detect the outliers [1],[2],[4],[8]:

1. Approach 1 - Deals to determine the outliers with no prior knowledge of the data. This is specifically a unsupervised learning approach like a unsupervised clustering method.

2. Approach 2 - Deals to determine the outliers with prior knowledge of data. This is specifically a supervised learning approach like Classification technique[12] and requires pre-labelled data, labled as normal or abnormal data.

3. Approach 3 - This deals with both known data and unknown data. It is known as semi-supervised task .This approach is used to recognise abnormal data from both labled and unlabled data. This approach needs pre-classified data but only learns data marked as normal.

4.1. The data mining based methods for outlier detection include the following[2],[4],[11],[13]:


TECHNIQUES

DISTRIBUTION ……. BASED

DEPTH BASED

DISTANCE BASED

DENSITY BASED

All these methods depends on various types of data available in the real world. The data set may be simple or complex type.

Methods need to be used when data set as simple are

-Parametric, Non parametric, Semi paramaetric methods.

Parametric: {Depth based, Distribution based,
Graph Based}

Non-parametric: { Distance based, Density based,
Clustering based methods}

Semi- parametric :{Neural network based,
Support vector based methods}

Methods used when dataset as Complex {

(High dimentional dataset => Subspace method,
Distance based methods) ,

(Sequence data set => Tree based method, Clustering
based method),

(Spatio Temporal dataset => Distribution based ,
Clustering based methods),

© 2015-2019, IJARCS All Rights Reserved

CONFERENCE PAPER
4th National Conference on Recent Trends in Information Technology 2015 on 25/03/2015
Organized by Dept. of IT, Prasad V. Potluri Siddhartha Institute of Technology, Kanuru, Vijayawada-7 (A.P.) India

(Spatial dataset => Distribution based method ,
Graph based methods)

(Mixed type attributes dataset => Graph based methods),

(Streaming dataset => Graph based, Model based,
Density based methods}.

– Distribution based methods : This method uses the following probability distribution policies, such as a Normal, Poisson or Bionomial distribution. Once the data distribution has been defined, it is statistically tested to determine which points are significantly different from the defined distribution and are therefore considered anomalous.

– Depth based methods[12] define each object as the representation of a point in a k-dimensional space. These points have their own depths, and the shallower ones are classified as outliers.

– Distance based methods[2] used to identify outliers by measuring the distance between a data point and its neighbor objects. Commonly used measures are Euclidean distances, Minkoswki distance etc.

– Density based methods[3] uses local density into account to identify anomalous data. These techniques estimates the density of objects and do not separate examples into categories, i.e., outliers and non-outliers, but instead provide a value for each example to signify how likely that object is to be an actual outlier.

– Clustering based methods[4][7] employ data mining techniques to isolate outliers in a cluster. These methods were not designed for this particular purpose, but rather, to group data that share certain characteristics.

– RNN-based methods[5] are known for their ability to distinguish normal from abnormal cases. Abnormal cases are not reproduced well in the exit layer.

– Support Vector-based methods[5] are usually employed for classification or regression analysis in data mining.

Proximity Measures:

Commonly used measures for proximity[7] are the Ecludean distance,The Minkowski Diastance for high dimentional datasets, Simple Matching Coefficient(SMC), Jaccord Coefficient for binary datasets, Cosine Similarity for document type datasets(binary type), Extended jaccord coefficient for high level document type datasets , Correlation for continuous datasets.

To decide which algorithms would be used for detecting outliers, the following elements are taken into consideration for solving a problem[9].

- Depends on type of data set ,we should choose a proper method and the ability of the algorithm to produce results that are comprehensible for the final user.
- The efficacy in its detection of outliers ,it should be high.
- The false positive rate, i.e., the fraction of data misclassified as outliers should be very low.
- The compatibility among the algorithms with the objectives of the procedure .
- The expected improvement of the efficacy by combining several techniques

V. CONCLUSIONS

We have successfully conducted a survey on outlier detection aimed to know the various strategies for detecting outliers. Most of the data mining researches focus on finding frequent patterns in the data, such as Clustering, Classification, Association rule mining methods are used and other aspects of data mining such as identifying abnormal behavior called “Outlier” and various detection methods such as Both Supervised and Unsupervised methods for finding abnormal behavior comes in datasets. Such techniques have both advantages and disadvantages.

VI. REFERENCES

[1] V. J. Hodge, J. Austin (2003) A survey of outlier detection methodologies. Artificial Intelligence Review, vol. 22, pp 85-126.

[2] C.C.Aggarwal,P.S.Yu,Outlier detection for high dimensional data, in: T.SellisandS.Mehrotra(Eds.),Proceedings of the 2001ACM SIGMOD

[3] C.C.Aggarwal,P.S.Yu ,An effective and efficient algorithm for high- dimensional outlier detection,VLDBJ.14(2005)211–221.

[4] Y. Zhang, N. Meratnia, P. Havinga, A taxonomy framework for unsupervised outlier detection techniques for multi-type data sets, Technical Report TR-CTIT-07-79, Centre for Telematics and Informa- tion Technology, University of Twente, Enschede, 2007.

[5] V. Barnett and T. Lewis (1994) Outliers in statistical data. John Wiley Sons, Reading,New York.

[6] J. Han and M. Kamber (2001) Data mining: concepts and techniques. Morgan Kaufmann.

[7] P-N Tan, M.Steinback, V.Kumar (2005) Introduction to data mining. Addison Wesley.

[8] [8]H.D.Kuna,R.Garcia-Martinez,F.r.Villatoro ,Outlier Detection in audit logs for application systems ,2014.

[9] E.Knorr and R.Ng,Algorithams for Mining Distance based outliers in large data sets,proc.24th international conference very large databases(VLDB 98),pp.392-403,1998.

[10] Alessia Albanese,sankar k.paul , Rough sets,Kernel sets and spatio temporal outlier detection,vol.26,No.1 January 2014.

© 2015-2019, IJARCS All Rights Reserved

CONFERENCE PAPER
4th National Conference on Recent Trends in Information Technology 2015 on 25/03/2015
Organized by Dept. of IT, Prasad V. Potluri Siddhartha Institute of Technology, Kanuru, Vijayawada-7 (A.P.) India

[11] T.cheng and Z. Li, A Multiscale Approch to detect spatial temporal outliers,Trans.GIS,vol.10 no.2,pp.253-263.

[12] R.jornsten, Clusterng and classification based on the L1 data depth,,vol. 90,no.1 pp.67-89,2004.

[13] Kanishka B., N.SRinivastava,Privacy –preserving Outlier detection Through Random Non linear Data Distrortion,IEEE,Vol. 41No.1 ,feb 2011.

[14] A. Lazarevic, A. Ozgur, L. Ertoz, J. Srivastava, and V. Kumar (2003) A comparative study of anomaly detection schemes in network intrusion detection. In: Proceedings of SIAM.

[15] rubbs, Frank (1969) Procedures for detecting outlying observations in samples. Techno-metrics, vol. 11, no. 1, pp. 1-21 .

© 2015-2019, IJARCS All Rights Reserved