ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

Outlier Detection for High Dimensional Data Using Graph Based Models

R.Revathi[1], Dr. Antony Selvadoss Thanamani[2]

M. Phil Scholar, Department of Computer Science, NGM College, Pollachi, India[1]

Associate Professor and Head, Department of Computer Science, NGM College, Pollachi, India[2]

ABSTRACT: Outlier detection is the process of detecting and subsequently excluding outliers from a given set of data. The outliers may be instances of error or indicate events. The task of outlier detection aims at identifying such outliers in order to improve the analysis of data and further discover interesting and useful knowledge about unusual events within numerous applications domains. In this paper we report on outlier detection techniques such as pre-labeled data and parameters of data distribution. Furthermore, we highlight the types of data sets which includes (simple, multiple attributes, high dimensional, sequential, spatial and streaming), characteristics of outlier and their applications. Graph-based methods make use of a powerful tool data image and map the data into a graph to visualize the single or multi-dimensional data spaces. Outliers are those points that are present in particular positions of the graph. These methods are suitable to identify outliers in real-valued and categorical data.

KEYWORDS: Outlier, Outlier detection techniques, data sets, graph-based method.

I. INTRODUCTION

Outliers are often considered as an error or noise, they may carry important information. Detected outliers are candidates for aberrant data that may otherwise adversely lead to model misspecification, biased parameter estimation and incorrect results. In high dimensional space, the data is sparse and the notion of proximity fails to retain its meaningfulness [1]. Outlier Detection over streaming data is active research area from data mining that aims to detect object which have different behaviour, exceptional than normal object[2]. Outlier detection has been a widely researched problem in several knowledge disciplines, including statistics, data mining and machine learning. It is also known an anomaly detection, deviation detection, novelty detection and exception mining in some literature [12].

II. OUTLIERS

A large number of techniques have been developed for building models for outlier and anomaly detection. However, the real world data set, data stream presents a range of difficulties that bound the effectiveness of the techniques. The assumed behavior of outliers is they were different from other members of cluster or they does not belong to any cluster, or belong to very small clusters, or forced to a cluster [10]. The clustering techniques are highly helpful to detect the outliers they are called cluster based outlier detection.

Outliers are patterns in data that do not conform to a well defined notion of normal behavior. Figure 1 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 the regions, e.g., points o1 and o2 and points in region O3, are outliers. x,y,N1,N2,o1,o2,O3 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.

Outliers often occur due to the following reasons, which make occurrence of an outlier typically being an indication of an error or an event [5].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

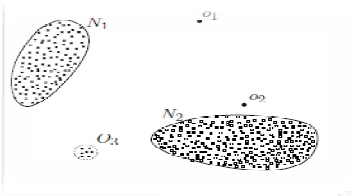

Figure 1: Example of outliers in a 2-dimensional data set

- *Error:* Outlier are known as anomalies, discordant observations, exceptions, faults, defects, aberrations, noise, damage or contaminants. They may occur because of human errors, instrument errors, mechanical faults or change in the environment.
- *Event:* As stated in outliers may be generated by a "different mechanism", which indicates that this sort of outliers belong to unexpected patterns that do not conform to normal behavior and may include interesting and useful information about rarely occurring events within numerous application domains.

The characteristics of outliers are,

- Outliers can be identified as either global or local outliers.
- A data point can be considered as an outlier in two manners, scalar (binary) or outlierness.
- Whether a data point is an outlier is determined by the values of its attributes such as univariate and multivariate.
- Outlier detection techniques can be designed to identify different number of outliers at a time.

III. OUTLIER DETECTION APPROACHES

A. *USE OF PRE-LABELLED DATA*

Supervised vs Unsupervised: Outlier detection approaches can generally be classified into three basic categories, i.e., supervised, unsupervised and semi-supervised learning approaches. This categorization is based on the degree of using pre-defined labels to classify normal or abnormal data.

Supervised learning approach: These approaches initially require the learning of normality and an abnormality models by using pre-labelled data, and then classify a new data point as normal or abnormal depending on which model the data point fits into. It is usually applied for many fraud detection and intrusion detection applications.

Unsupervised learning approach: These approaches are more general because they do not need pre-labelled data that are not available in many practical applications. Similarly, distance-based methods identify outliers based on the measure of full dimensional distance between a point and its nearest neighbours.

Semi-supervised learning approach: These approaches only require training on pre-labelled normal data to learn a boundary of normality, and then classify a new data point as normal or abnormal depending on how well the data point fits into the normality model.

B. *USE OF PARAMETERS OF DATA DISTRIBUTION*

Parametric vs Non-parametric: Unsupervised learning approaches can be further grouped into three categories, i.e., parametric, non-parametric and semi-parametric methods, on the basis of the degree of using the parameters of the underlying data distribution.[9]

Parametric method: These methods assume that the whole data can be modelled to one standard statistical distribution (e.g., the normal distribution), and then directly calculate the parameters of this distribution based on means and covariance of the original data.

Non-parametric method: These methods make no assumption on the statistic properties of data and instead identify outliers based on the full dimensional distance measure between points. Outliers are considered as those points that are distant from their own neighbours in the data.

Semi-parametric method: These methods do not assume a standard data distribution for data, but instead map the data into a trained network model or a feature space to further identify if these points deviate from the trained network model or are distant from other points in the feature space, on the basis of some classification techniques such as neural network and support vector machine.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

IV. TYPES OF DATA SET

Here, we describe several common types of data sets based on the characteristics and attributes of data such as simple and complex data sets as follows:

A. *Simple Dataset*
The simple data set belongs to a commonly used data set, where the data has no complex semantics and usually is represented by low- dimensional real-valued ordering attributes. Most existing outlier detection techniques are applicable for such simple data sets.

B. *High Dimensional Data Set*
This data set contains a large number of data and each data point also has a large number of attributes. As stated before, detecting multivariate outliers is more complicated, thus many outlier detection techniques may be susceptible to the problem of the curse of dimensionality in high-dimensional data sets, especially high computation complexity and no sufficient similarity measures.

C. *Mixed-Type Attributes Data Set*
In some practical applications, the data contains the mixture of continuous (numeric) and categorical attributes. The latter usually has non-numeric and partial ordering values, e.g., city names, or type of diseases. This makes it very difficult to measure the similarity between points by commonly used measure methods.

D. *Sequence Data Set*
In the sequence data set, the data is naturally represented as a sequence of individual entities, such as symbols or letters. Also, the data has not the same length and no priori known distribution.

E. *Spatial Data Set*
Attributes of spatial data set are distinguished as spatial and non-spatial attributes. Spatial attributes contain location, shape, directions and other geometric or topological information. They can determine spatial neighbourhoods in terms of spatial relationships such as distance or adjacency. On the other hand, non-spatial attributes include the intrinsically information of data characteristic, which are used to compare and distinguish spatial points in the spatial neighbourhood.

F. *Streaming Data Set*
A data stream is a large data that is arriving continuously and fast in the ordered sequence. They usually are unlimited in size and occur in many real-time applications. For example, a huge amount of data in average of daily temperature are collected to the base station in wireless sensor networks continually. Thus, an efficient outlier detection technique is required to deal with the data streams in an online fashion.

G. *Spatio-Temporal Data Set*
Due to the fact that many geographic phenomena are evolving over time, the temporal aspect and spatial-temporal relationships existing among spatial data also should be considered in detecting outliers for real-life applications, e.g., geographic information systems (GIS), robotics, mobile computing, traffic analysis etc.

V. OFDM DETECTION METHOD FOR SIMPLE DATASET

A. *GRAPH –BASED METHOD*

Graph-based methods make use of a powerful tool data image, i.e., map the data into a graph to visualize the single or multi-dimensional data spaces. Outliers are expected to those points that are present in particular positions of the graph. Laurikkala et al. [6] propose an outlier detection approach for univariate data based on box plot, which is a simple single-dimensional graphical representation and includes five number values: lower threshold, low quartile, median, upper quartile and upper threshold. Figure 2 shows an example of a box plot. Using box plot, points that lie outside the lower and upper threshold are identified as outliers. Also, these detected outliers can be ranked by the occurrence frequencies of outliers. Thus, the box plot effectively identifies the top n outliers with the highest occurrence frequencies and then discards these outliers. The approach is applicable for real-valued, ordinal and categorical data. However, it is too subjective due to excessively rely on experts to determine several specific points plotted in the graph, e.g., low and upper quartile. Scatter plot [8] is a graphical technique to detect outliers in two-dimensional data sets. It reveals a basic linear relationship between the axis X and Y for most of the data. An outlier is defined as a data point that deviates significantly from a linear model. Figure 3 shows an example of a scatter plot. In addition, spin plot [15] can be used for detecting outliers in 3-D data sets.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

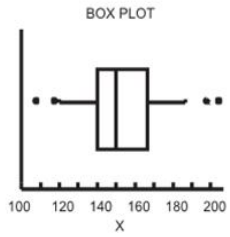

Fig 2.An example of a box plot

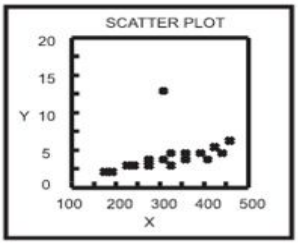

Fig 3.An example of a scatter plot

B. *Evaluation of Graph-based Techniques*

Graph-based approaches have no assumptions about the data distribution and instead exploit the graphical representation to visually highlight the outlying points. They are suitable for identifying outliers in real-valued and categorical data. However, they are limited by the lack of precise criteria to detect outliers. In particular, several specific points in the graph are determined subjectively by experts, which is also a very time-consuming and difficult process.

VI. APPLICATIONS

Fraud detection:[10]The purchasing behaviour of people who steal credit cards may be different from that of the owners of the cards. The identification of such buying pattern changes could effectively prevent thieves 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:[3] Frequent attacks on computer systems may result in systems being disabled, even completely collapsing. 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 keep out hackers.

Environmental monitoring:[4]Many unusual events that occur in the natural environment such as a typhoon, flooding, drought and fire ,often have an adverse impact on the normal life of human beings and allow people to take effective measures on time.

Medical and public health:[7] Patient records with unusual symptoms or test results may indicate potential health problems for a particular patient. The identification of such unusual records could distinguish instrumentation or recording errors from whether the patient really has potential diseases and so take effective medical measures in time.

Localization and tracking:[13] Localization refers to the determination of the location of an object or a set of objects. The collection of raw data can be used to calibrate and localize the nodes of a network while simultaneously tracking a moving target. Filtering such erroneous data could improve the estimation of the location of objects and make tracking easier.

Logistics and transportation:[14] Logistics refers to manage and control the follow of products from the source of production to the destination. It is very essential to ensure product safety and product reliability issues during this process. Tracking and tracing shipments could find out potential exceptions, e.g., inappropriate quantity and quality of the product, and notify all trading partners in time.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

VII. CONCLUSION

In this paper, we present an outlier detection techniques for specific application domains and type of data sets consists of simple and complex data. There is no single universally applicable or generic outlier detection approach. Thus, the developers should consider whether an outlier detection technique is suitable for a data set depending on several important aspects, i.e., the use of pre-labelled data, use of parameters of data distribution, the type and dimension of detected outliers, the degree of being outliers, the number of detected outliers at once. Also, Graph based method using outlier detection techniques can be used map the data into a graph to visualize the single or multi-dimensional data spaces.

REFERENCES

1. Aggarwal, C. C. Yu, P. S. (2001). Outlier Detection for High Dimensional Data. Proceedings of the ACM SIGMOD Conference 2001.
2. Bakar, Z. A., Mohemad, R., Ahmad, A., & Deris, M. M.(2006), "A comparative study for outlier detection techniques in data mining", In Proc. 2006 IEEE Conf. Cybernetics and Intelligent Systems, pp. 1–6, Bangkok, Thailand.
3. D.J. Marchette (2001) Computer intrusion detection and network monitoring: a statistical viewpoint. New York: Springer
4. G. M. Davis, K. B. Ensor (2006) Outlier detection in environmental monitoring network data: an application to ambient ozone measurements for Houston, Texas. Journal of Statistical Computation and Simulation, vol. 76, no. 5, pp 407-422
5. J. Han and M. Kamber (2001) Data mining: concepts and techniques. Morgan Kaufmann
6. J. Laurikkala, M. Juhola, E. Kentala (2000) Informal identification of outliers in medical data. In: Proceedings of IDAMAP
7. J. Lin, A. E. Fu and H. V. Herle (2005) Approximations to magic: Finding unusual medical time series. In: Proceedings of Symposium on Computer-Based Medical systems. Washington, DC, USA, pp 329-334
8. M. Markos, S. Singh (2003) Novelty detection: a review-part 1: statistical approaches in Signal Processing, vol. 83, pp 2481-2497.
9. P. M. Valero-Mora, F. W. Young, M. Friendly (2003) Visualizing categorical data in ViSta. Computational Statistics & Data Analysis, vol. 43, pp 495-508.
10. R. J. Bolton, D. J. Hand (2001) Unsupervised profiling methods for fraud detection. In: Proceedings of CSCC
11. R. Milen, A. Sohal and S. Moss (1999) Quality management in the logistics function: an empirical study. Journal of Quality & Reliability Mangement, Volume 16, pp 166-180
12. V. Hodge and J. Austin, "A Survey of Outlier Detection Methodologies", Artificial Intelligence Review, Vol. 22, pp. 85-126, 2003.
13. V. J. Hodge, J. Austin (2003) A survey of outlier detection methodologies. Artificial Intelligence Review, vol. 22, pp 85-126
14. W. Du, L. Fang and P. Ning (2005) LAD: localization anomaly detection for wireless sensor networks. In: Proceedings of Parallel and Distributed Processing Symposium.
15. Y. Panatier (1996) Variowin. Software for spatial data analysis in 2D. Springer-Verlag, New York