International Journal of Engineering Technology and Applied Science

(ISSN: 2395 3853), Vol. 6 Issue 11 November 2020

An Approach to Classify Outlier by Using Poisson distribution

[1]Rani Kushwaha, [2]Praveen Kataria

Department of Computer Science & Engineering

All Saints College Of Technology, Gandhi Nagar, Bhopal

salonikush95@gmail.com

Abstract : The outlier detection plays an important role in data mining in order to collect the important information. An outlier may indicate bad data. There are number of methods use to find the outlier detection. Some time it seems to be that, it is difficult to find if an outlying point is bad data. Outliers may be due to random variation or may indicate something scientifically interesting in this paper there is a discussion of various methodology by which the outlier detection can be possible. In this paper proposed approach use evolutionary based similarity to computing likelihood values for each local data behavior in the feature space and applies Poisson distribution approach for probabilistic classification on probabilistic-based learning framework.

Keywords:- Data Mining , outlier detection, Poisson distribution, Machine learning

I. INTRODUCTION

Detection of outliers is mission critical in many of the critical security environments such as the Standard Operating indicates abnormal conditions and that a clear deterioration of the performance may be, as the default aircraft engines or a problem with the rotation of the flow in the tube. Undefended can indicate anomalous object in an image, such as landmines. Undefended may determine the intruder inside the system with malicious intent quickly disclosure is necessary. Outlier detection can detect an error in the factory production line constantly monitor the specific characteristics of the products and comparison of data in real time, either with a standard product characteristics or those of failure.

t is urgent tasks such as monitoring the use of credit cards or mobile phone surveillance to detect sudden change in the pattern of usage that may indicate the use of forged or stolen card they are broadcast on the stolen phone time. Outlier detection achieves this by analyzing and comparing usage statistics time series. Processing applications, such as loan payments or social security benefits request, the system can detect outlier detect any abnormality in the application before approval or payment. Outlier detection can also control the circumstances of the claim interest in a timely manner to ensure payment of the fraud did not retreat. Can equity or commodity traders use or disclosure of the extreme values of individual methods of procedures to monitor markets and detect trends that might indicate buy or sell opportunities. News delivery system can detect the change news and ensure the seller is the first with the news. In the database, it may outliers refers to cases of fraud or may refer to just enter a defect or misinterpretation of the missing value code either employee revealed by anomalies is vital for consistency and integrity database.

Outliers can cause the data for a variety of reasons, such as malicious activity, for example, credit card fraud, cyber infiltration, terrorist activity or system failure.

Issues

- Various Constraints in Resources
- Large amount of Communication Cost
- Distributed Data Streaming
- Identification of Outlier Source

II. APPLICATIONS

There a number of applications available in which the outlier detection take place for efficient results. Some of them discussed here.

Detection of Fraud: There is a need to detect the fraudulent in order to apply the efficient credit cards system, or usage of cards or mobile phones.

Processing of Loan Application: this type of application are use to detect fraudulent applications or can be find the potentially problematical customers from bulk data

Detection of Intrusions: Here the detection of unauthorized access in an environment of computer networks can show.

Monitoring of Activities: there should be monitoring approach to get the mobile phone fraud by the suspicious activity

Performance Analysis of Network: The network performance of large number of computers needs to monitor problem like bottlenecks in system

Medical Science: It can be applicable in the medical science. In medical equipments it can be apply in order to monitor the heart-rate.

Research field of Pharmaceutical: in this area the structure of molecules are the very important parts to analysis in order to find the new molecular structures

III. DIFFERENT APPROACHES

There are three fundamental approaches to the problem of outlier detection:

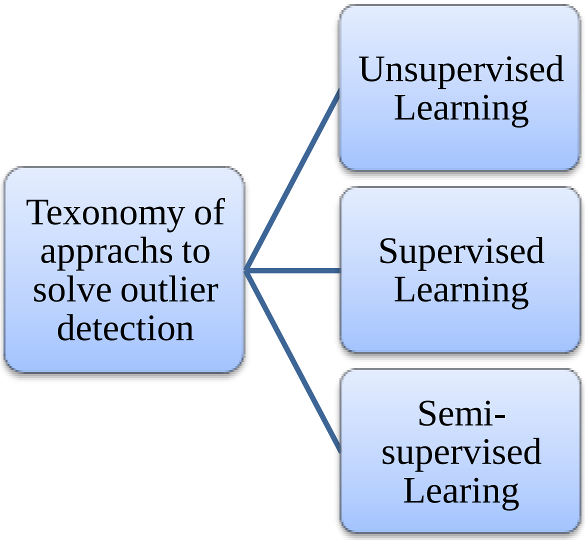

Figure 1 Methods to Solve Outliner detection

Unsupervised Learning: It is use to find outliers without having previous knowledge. This is basically a similar gathering to approach learning uncensored. Focus on fixed data such as distribution operations, says the remote points, and flags as possible outliers

Supervised Learning: This approach is similar to the classification under the supervision and does not require data described previously, in normal or abnormal category .. There are three types of normal data with pre-marked in remote areas outliers. The whole region outside regular class is a class of outliers. And it can be classified as regular points and one class or be divided into three distinct categories according to the requirements of the rules of normal / abnormal simple classification to provide or to provide seed unnatural and layer 3 as normal.

Semi Supervised Learning: The author Fawcett, Provost and Japkowicz has been proposed an approach known as the novelty recognition. It is similar to the task of detecting the role or supervised by the recognition can be considered semi-supervised regular class taught but the algorithm learns to identify abnormalities. Approach requires disaggregated data in advance, but usually only learn the data significantly. It is suitable for fixed or variable data because they are only aware of the class that provides natural lifestyle. You can find out a progressive model up as the new data, and develop a model to improve the fit as each new issue is available. Aim is to determine the limit of normal.

IV. LITERATURE SURVEY

Work outlier detection [6] is to identify data objects that are markedly different or incompatible with the normal play data. Most existing solutions are building a model from normal data and identify outliers that do not fit very well represented in the model. However, in addition to normal data, there are also negative examples aberrant or limited in many applications, and data can be damaged so that the detection data outliers are mislabelled. This makes detection much more difficult than traditional outliers. This paper presents an approach for outlier detection novel to deal with imperfect data labels and integrate abnormal examples restricted in learning. To process the data with imperfect labels, we introduce the probability values for each data entry denoting the degree of membership of an example of normal and abnormal, respectively classes. Our proposed approach works in two steps. In the first step, we generate a set of pseudo training data by calculating probability values for each example according to local behaviour

Introducing core (k) Means clustering method based on kernel and LOF to calculate the probability values method. In the second step, we integrate the probability values generated and limited abnormal SVDD examples in learning to build a more accurate for general detection of outliers classifier. The integration of local and global detection of outliers, our proposed method explicitly handles with imperfect data improves performance labels and detection of outliers. Many experiments with real data has shown that our proposed approach can achieve a better compromise between the detection rate and false alarm rate compared to detection of atypical prior art approaches.

The author has discussed on outlier detection that can happened too much from the other events. Identify outliers can lead to meaningful and useful knowledge discovery. Protuberant means that what has happened in that time is not a normal activity. He studied on a large scale research to detect outlier in the past decade. However, most current research focuses on the algorithm based on the specific knowledge; relative to the outlier detection approach is still rare. This article focuses mainly on the different types of approaches to the discovery of anomalous values and compares it is subject and cones. In this article, we distribute typical detection approach basically both parties to approach the traditional values and extreme outliers spatial approach. The traditional approach identifies anomalous outliers in real estate transactions data set, which can be classified in the statistical approach, the approach of the distance, the deviation approach, and the approach of intensity. Spatial approach anomaly detection and outliers based on the spatial data set differs from transaction data, which can be classified in the approach and methodology away graph. Finally, the comparative approach outlier detection has been shown

Outlier detection [8] area has recently become the focus of research in data mining. Here, we propose the concept of effective DenOD outlier detection (density based on outlier detection) based on the method of uncensored for the detection of leaks in the cloud work environment. Unattended outlier detection techniques play a major role in various applications such as intrusion detection network, and detecting errors and fraud detection. Jamal method is that unattended, it does not require any set of training data or any other type of prior knowledge. This technique may help detect new attacks accuracy and without any prior knowledge. DenOD will implement the IDCC (Intrusion Detection cloud computing) framework that has three legs on the Cloud contract, IDS (Intrusion Detection System) and end-user. This technique is able to detect all types of attacks, as well as for the detection of defective services in the cloud work environment.

High data mining flow [9] attract many researchers have a large database of mine that presents different challenges to researchers need base. The data stream is different from the usual data as they are produced continuously from different applications that require different challenges and large-scale drift concept of infinity, for treatment. The so-called object does not respect the normal values of the object data extremist behaviour. Typical uses detection in various applications such as fraud detection, intrusion detection, and monitoring of environmental changes, and medical diagnosis so it is necessary for the detection of outliers of RSS data. And use different methods to detect outlier. Some of them use the K- means algorithm for the detection of outliers in the data tables that help to create a group or groups of similar data points. Classification data flow techniques are very useful for a group of similar data elements in the data stream, as well as for the detection of extreme values them, so they are called outlier detection block list. K- means algorithm algorithm is based on the section which is used to collect data in several groups groups. It is more common and popular algorithm to assemble because of its simplicity and effectiveness. The purpose of this paper is to study different methods for outlier detection, which is used for K- data collection algorithm means that have been identified with other methods. Detection of outliers different applications in this document are processing areas.

Detection of outliers [10] is one of the most important issues of data mining and knowledge discovery in databases. It is to find a methodology for the detection of cases in the set of data that does not fit with the rest of the worksheet. Local outliers factor is the result before the typical detection. In this paper, it is suggested to calculate the difference regulator after outlier factor algorithm is typical detection free parameter values for a new approach. We formulate the result is a new module for each instance when considering the difference in distance order. We then use this value to calculate the typical degree. We use the score for each state to provide a degree of outlier and compared with LOF. Our systems can produce OOF in Θ (N2) without parameter.

V. PROPOSED WORK

In this paper proposed approach use evolutionary based similarity to computing likelihood values for each local data behavior in the feature space and applies Poisson distribution approach for probabilistic classification on log file.

This work proposed a new approach to identify the malicious user or attacker. This can be done by analyzing the log files. There are three basic fundamentals component of the proposed work.

Log Correlation and Centralization: Initially log file is situated at different remote location and trace the activity according to their time zone. Log correlation is responsible to correlated these remote log file at a centralized location depend upon GPS time band.

Positive Feature: This step is used for the applying the Similarity function for extracting associate relevant positive features set as S1, S2,S3 from each acquiesced data set.

Negative Feature: This step use poisson distribution approach to extract negative feature associated with S1, S2 and S3 as N1,N2 and N3. These negative features set Ni have outlier behavior and use for outlier detection.

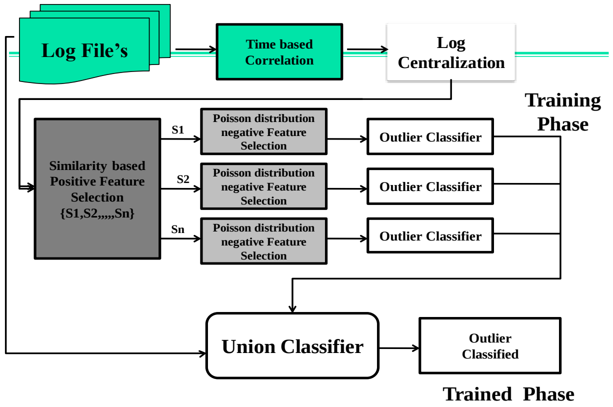
Log File's
Time based Correlation
Log Centralization
Training Phase
Similarity based Positive Feature Selection {S1,S2,,,,,Sn}
S1
S2
Sn
Poisson distribution negative Feature Selection
Poisson distribution negative Feature Selection
Poisson distribution negative Feature Selection
Outlier Classifier
Outlier Classifier
Outlier Classifier
Union Classifier
Outlier Classified
Trained Phase

Figure 2 Proposed Architecture

Poisson distribution

Poisson distribution approach is use to modulate the number of events in a precise time period. For example to modulate numerous telephone calls at a call centre , analysis of faults in a given surface area, airplane arrivals, or the number of accidents at an intersection. In same way proposed work use Poisson distribution for extracting number of negative feature from log file. Basically proposed work based on the concept that number of negative incident is very minor as compare to positive incident. If any incident if not be occur form log time and but currently in any time scenario it is to be occur frequently then it is suspicious of us.

Poisson not to be considered have a fixed number of trials. Other hand it uses the fixed interval of time or space in which the number of successes is recorded.

Parameters: The mean is λ. The variance is λ.

$$p(x,y) = \frac{e^{-y}y_x}{x!}..................1$$

For X = 0,1,2,…..

Where

X : Representing the number of occurrences of negative feature in a continuous interval.

Y : Expected value of occurrences in this interval.

The probability of an event is the same for both intervals of equal length!!. The expected value of occurrences in a gap is proportional to the length of this interval. The presence or absence in an interval is independent of the occurrence or nonoccurrence of any other interval. The probability that two or more occurrences in a very small gap is near to 0.

VI. RESULT ANALYSIS

The performance of an intrusion detection system can calculate in terms of True Positive rate and false positive rate rate. TP rate is a result or division of abnormal patterns detected by any system and the total abnormal patterns of system. A simple represent in mathematically is shown below.

$$\text{TPR}=\frac{\text{TP}}{\text{TP+FN}} =\text{P (A| I)} \qquad \text{............. (1)}$$

Similarly the True negative rate is shown in the below formula

$$\text{TNR}=\frac{\text{TN}}{\text{TP+TN}}=\text{P } (\leftharpoondown \text{A}| \leftharpoondown \text{I}) \qquad \text{.............}$$

(2)

False Positive rate take place when the any system classify the normal result in a wrong manner. In this experiment, FP rate is calculation can be done by the number of false positives created by the system, divided by the total number of self-antigens.

$$\text{FNR}=\frac{\text{FP}}{\text{FP+TN}}=\text{P } (\leftharpoondown \text{A}| \text{I}) \qquad \text{.............. (3)}$$

Similarly False negative rate calculated by

$$\mathrm{FNR}=\frac{\mathrm{FN}}{\mathrm{TP+FN}}=\mathrm{P}\ (\leftharpoondown \mathrm{A}|\ \mathrm{I}) \qquad \ldots\ldots\ldots\ldots.(4)$$

The comparison of the simulation result is given in fig.3. It gives the comparison of the accuracy rate for the classification of attack using the traditional method namely Bayesian approach with our proposed method Poisson distribution. In simulation the generating function also called the activated threshold value was set to The maximum accuracy rate of our algorithm is possible only by using Poisson distribution method .fig.3 shows when using Bayesian approach& Poisson distribution of the accuracy of attack never reaches even 92.00% but by using Poisson distribution approaches the accuracy rate reaches 96.00%. The x-axes represent the accuracy rate and the y-axes indicate different Log file.

Table 1 Comparative Accuracy

DataSet	Bayesian Approach	Poisson distribution.
1	91.4	95.9
2	89.79	94.3
3	91.55	96.06
4	91.53	96.04
5	91.55	96.06
6	89.79	94.3
7	89.29	94.3

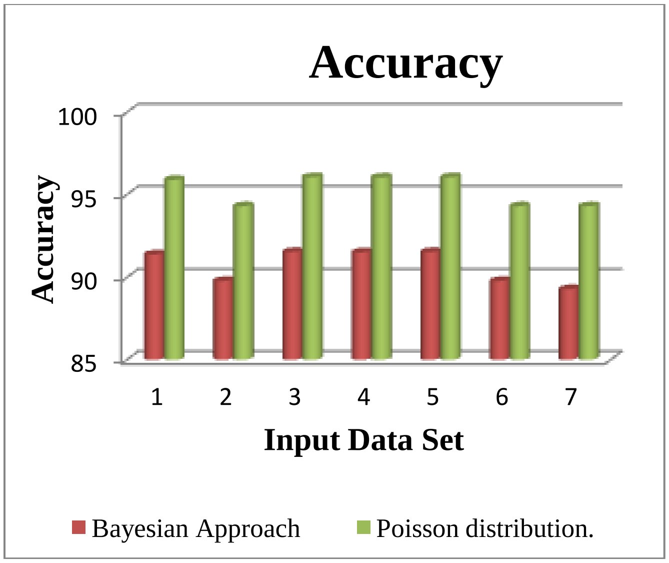

Figure 2 Comparisons of Bayesian approach& Poisson distribution

Table 2 shows the results of the experiment1. In experiment 1 we classify the data with the help of Poisson distribution, the Poisson distribution separately with considering different generating function .from experiment 1 we conclude that the maximum accuracy of the detection of intrusion using Bayesian approach never reaches above 92%, whereas by using Poisson distribution the accuracy becomes even maximum 96%.

VII. CONCLUSION

The outlier detection plays an important role in data mining in order to collect the important information. An outlier may indicate bad data. There are number of methods use to find the outlier detection .In this paper there is a discussion of various methodology by which the outlier detection can be possible. In this paper proposed approach use evolutionary based similarity to computing likelihood values for each local data behavior in the feature space and applies Poisson distribution approach for probabilistic classification on probabilistic-based learning framework.

REFERENCES

[1] Jiang Mon-Fong, Tseng Shian-Shyong and Su Chih-Ming "Two-phase clustering process for outliers detection", *Pattern recognition letters 22*, no. 6, pp.691 -700 2001

[2] Elahi Manzoor, Kun Li, Wasif Nisar, Xinjie Lv and Hongan Wang "Efficient clustering-based

outlier detection algorithm for dynamic data stream", *IEEE*, vol. 5, pp.298 -304

[3] Yogitaa Durga Toshniwala "A Framework for Outlier Detection in Evolving Data Streams by Weighting Attributes in Clustering",*2nd International Conference on Communication, Computing & Security*, pp.214 -222

[4] Yu, D., Sheikholeslami, G. and Zang, "A find out: finding outliers in very large datasets". In Knowledge and Information Systems, 2002, pp.387-412.

[5] Aggarwal, C. C., Yu, S. P., "An effective and efficient algorithm for high-dimensional outlier detection, The VLDB Journal, 2005, vol. 14, pp. 211-221.

[6] Bo Liu; Yanshan Xiao; Yu, P.S.; Zhifeng Hao; Longbing Cao, "An Efficient Approach for Outlier Detection with Imperfect Data Labels," in *Knowledge and Data Engineering, IEEE Transactions on*, vol.26, no.7, pp.1602-1616, July 2014

[7] Chandarana, D.R.; Dhamecha, M.V., "A survey for different approaches of Outlier Detection in data mining," in *Electrical, Electronics, Signals, Communication and Optimization (EESCO), 2015 International Conference on*, vol., no., pp.1-5, 24-25 Jan. 2015

[8] Kumar, M.; Mathur, R., "Unsupervised outlier detection technique for intrusion detection in cloud computing," in *Convergence of Technology (I2CT), 2014 International Conference for*, vol., no., pp.1-4, 6-8 April 2014

[9] Chauhan, P.; Shukla, M., "A review on outlier detection techniques on data stream by using different approaches of K-Means algorithm," in *Computer Engineering and Applications (ICACEA), 2015 International Conference on Advances in* , vol., no., pp.580-585, 19-20 March 2015

[10] Buthong, N.; Luangsodsai, A.; Sinapiromsaran, K., "Outlier detection score based on ordered distance difference," in *Computer Science and Engineering Conference (ICSEC), 2013 International*, vol., no., pp.157-162, 4-6 Sept. 2013