International Journal of Scientific Research & Engineering Trends
Volume 5, Issue 3, May-Jun-2019, ISSN (Online): 2395-566X

Poisson distribution Based Weighted Probabilities Model for Outlier Detection

Rashmi Rohitas rashmi.rohitas0001@gmail.com
Prof. Ruchi Dronawat dron.ruchi@gmail.com
Department of Computer Science & Engineering
Sagar Institute of Research and Technology
Bhopal, M.P., India

Abstract- **Outliers may be due to random variation or may indicate something scientifically interesting in this. In this 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. This approach gives a bird eye over data files that concentrate over dynamic nature of log files. Proposed methodology applies Poisson distribution approach for probabilistic classification on high dimension and unstructured file. As audit log is an unstructured file having. Large number of attribute with high dimension distribution probability is better for high dimension and unstructured data class.**

Keywords - **Data mining, Classification, Poisson Distribution, machine learning Positive feature , Negative feature**

I. INTRODUCTION

Outlier detection is currently very active area of research in data set mining community. Finding outliers in a collection of patterns is a very well- known problem in data mining field. An outlier is a pattern which is dissimilar with respect to the rest of patterns in the dataset. 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. An outlier can identify an intruder in a system with malicious intent early detection is essential.

It is essential in tasks such as monitoring the use of credit cards or mobile surveillance to detect a sudden change in the pattern of use that may indicate fraudulent use of stolen card stolen antenna or phone time. The detection of outliers out this mission by analyzing and comparing usage statistics time series. For the treatment of the application, such as loan payments and social security benefits applications, and the system can detect outliers detect any anomaly in the application prior to approval or payment. Outlier detection controlled the function of the circumstances of the expected demand in a timely manner to ensure payment fraud not retreat. O Distributors can refrain from using raw materials typical detection to monitor the actions of certain markets or detection methods and trends that might indicate buy or sell opportunities. New transfer system can detect the change and reporting to ensure that the first resource with the latest news. In the database, you can outliers refers to cases of fraud or set the writer can only go into default or misinterpreting the missing value code or a way to detect anomalies is vital to the cohesion and integrity of the base of data. This can be incentive outliers in the data for a variety of reasons, such as malicious activity, for example, credit card fraud, and cyber infiltration and terrorist activity or system collapse.

Issues

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

Outliers are patters in data that do not conform to a well-defined notion of normal behavior or data object that deviates significantly from the normal objects as if it were generated by a different mechanism Ex. Unusual credit card purchase, Credit Card Fraud, Cyber Intrusion, Terrorist Activity or Break Down of system.

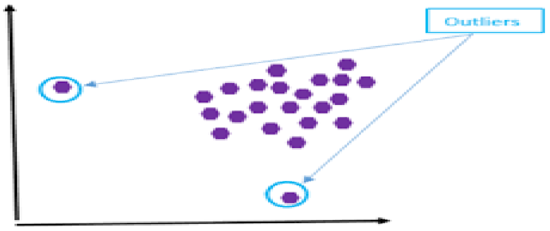

Figure 1 Illustrates outliers in a simple 2- dimensional data set.

In this figure 1 the data has two normal regions. In which two blue points that are in circle have abnormal behavior hence define as oultiers

II.PROPOSED WORK

In this dissertation 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 Data file.

III. PROPOSED ARCHITECTURE

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

1. **Data Correlation and Centralization -** Initially Data file is situated at different remote location and trace the activity according to their time zone. Data correlation is responsible to correlated these remote Data file at a centralized location depend upon GPS time band.
2. **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.
3. **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 behaviour and use for outlier detection.

Table 1 Comparative Accuracy

Data Set	Bayesian Approach (Existing)	Poisson distribution (Proposed)
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

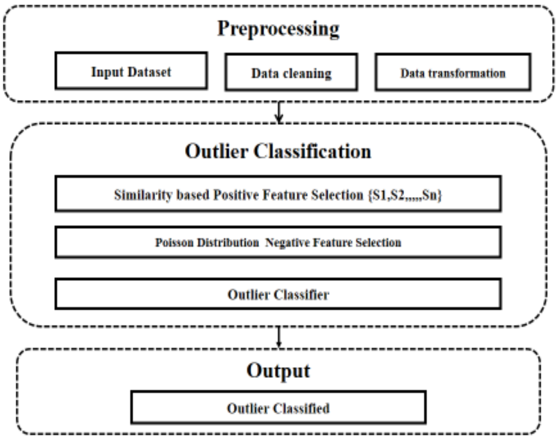

Figure 2 Proposed Architecture.

1. 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 Data 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 Data 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. Parameter The mean is λ.The variance is λ.

$$p(x,y) = \frac{e^{-y} y_x}{x!} \text{..................Eq (2)}$$

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 non occurrence of any other interval. The probability that two or more occurrences in a very small gap is near to 0.

IV. PROPOSED ALGORITHM

The Proposed Solution is going to Provide Efficient Scheme. Proposed Approach use Poisson Distribution that Works effectively over Datas Files and increase Performance.

Step 1- Data Set of Data Files is selected then data is reduced according to rules set or behavior of data instances

Step 2- Possitive Features S1 , S2 & S3 are selected from Data files

Step 3- Possitive Features S1, S2 & S3 negative features are selected.

Step 4- poisson distributions are applied over Selection.

Step 5 - Outliers are classified after applying Poisson distribution

Step 6- All Outliers are grouped in a single group

Step 7- Grouped outliers are detected as outliers

IV. 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.

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

Similarly the True negative rate is shown in the below formula

$$TNR=\frac{TN}{TP+TN}=P\ (\leftharpoonup A|\leftharpoonup I) \qquad \text{....} (2)$$

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.

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

Similarly False negative rate calculated by

$$FNR=\frac{FN}{TP+FN}=P\ (\leftharpoonup A|\ I) \qquad \text{.......}(4)$$

V. COMPARATIVE RESULT ANALYSIS

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 Data file.

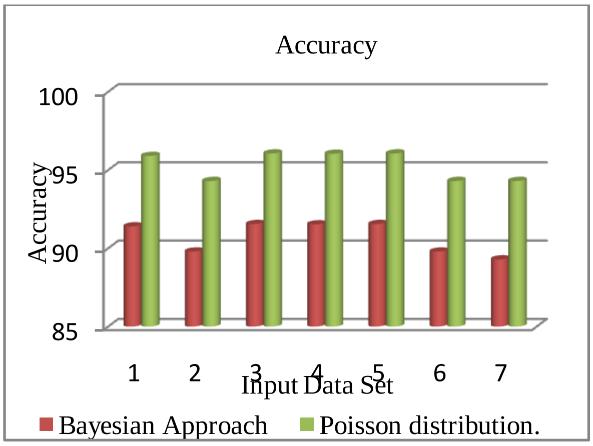

Figure 3 Comparisons of Bayesian approach& Poisson distribution.

Table 1 shows the results of the experiment. In experiment we classify the data with the help of Poisson distribution, the Poisson distribution separately with considering different generating function from experiment 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%.

VI. 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. Some time it seems to be that, it is difficult to find if an outlying point is bad data. Outlier in Data file is use to detect the culprit there is a need to improve the investigation mechanism.

Data file provide troubleshooting, security and pro-active system administration that provide significant help in caching suspicious end user and in process of cyber forensic. An outlier may indicate bad data. There are number of methods use to find the outlier detection .In this paper In this dissertation, 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. As the results show that the proposed method Data gives the improved results which are better from the previous work. Proposed frame work encourages the web investigator to navigate the end user behavior and assist to enforce the effective security policy.

REFERENCE

[1]. Tomono, A. ; Uehara, M. ; Shimada, Y. Improvement and Evaluation of a Method to Manage Multiple Types of Logs IEEE 2011, P 601-607

[2]. Karen Kent and Murugiah Souppaya, “Guide to Computer Security Log Management”, Computer Security Division Information Technology Laboratory National Institute of Standards and Technology Gaithersburg, 2006

[3]. Hulitt, E. ; Vaughn Jr., Rayford B. “Information system security compliance to FISMA standard: A quantitative measure” IEEE 2008 p 799-806

[4]. Keun-gi Lee ; Savoldi, A. ; Gubian, P. ; Kyung Soo Lim ; Seokhee Lee ; Sangjin Lee “Methodologies for Detecting Covert Database” IEEE Conference Publications 2008 p 538-541

[5]. Gardazi, S.U. ; Shahid, A.A. ; Salimbene, C. “HIPAA and QMS Based Architectural Requirements to Cope with the OCR Audit Program” IEEE Conference Publications 2012 pp 246-253

[6]. Karagiannis, D. ; Mylopoulos, J. ; Schwab, M. “Business Process-Based Regulation Compliance: The Case of the Sarbanes-Oxley Act” IEEE Conference Publications 2007 pp 315-321

[7]. Sheth, C. ; Thakker, R. “Performance Evaluation and Comparative Analysis of Network Firewalls” IEEE Conference Publications 2011 pp 1-5

[8]. Nikhil Kumar Singh, Deepak Singh Tomar, Bhola Nath Roy, “Gathering & Analysing the Suspicious end user activity in Log Files”,The Proceedings Of National Conference On Recent Trends & Challenges In Internet Technology ,Manit,Bhopal,2010

[9]. M. Bishop, “A Standard Audit Trail Format”, National Information Systems Security Conference, Baltimore, MD, 1995

[10]. Guideline on Auditing and Log Management - National Computer Board Mauritian Computer Emergency Response Team Enhancing Cyber Security in Mauritius july 12 version 1.1 issue no 5

[11]. Leite, J.P. “Analysis of log files as a security aid IEEE Conference Publications 2011 pp 1-6

[12]. Jaya thilake, D. “Towards structured log analysis” ” IEEE Conference Publications 2012 pp 259-264

[13]. Robert Rinnan “Benefits of Centralized Log file Correlation” Master’s Thesis, Master of Science in Information Security30 ECTS, Department of Computer Science and Media Technology Gjøvik University College, 2005

[14]. Herrerias, J., Gomez. R., Log Analysis Towards an Automated Forensic Diagnosis System IEEE Conference Publications 2010 pp 659-664

[15]. Abad C., Taylor J., Sengul C., Yurcik W., Zhou Y., & Rowe K., “ Log correlation for intrusion detection: A proof of concept”. In ACSAC 03: Proceedings of the 19th Annual Computer Security Applications Conference, 255, Washington, DC, USA. IEEE Computer Society.2005

[16]. Net Forensics inc. “Tech brief. Rationalizing security events with three dimensions of correlation”.2005

[17]. Guarded Net, “Four correlation technologies for improved incident recognition and accelerated response”,2005

[18]. Zhi-Yong Li A Network Security Analysis Method Using Vulnerability Correlation” IEEE Conference Publications 2009 pp 17 – 21

[19]. Guillame-Bert M., Crowley J.L., “New Approach on Temporal Data Mining for Symbolic Time Sequences: Temporal Tree Associate Rules” IEEE Conference Publications 2011 pp 748 – 752

[20]. Ye Changguo, Wei Nianzhong , Wang Tailei , Zhang Qin, Zhu Xiaorong The Research on the Application of Association Rules Mining Algorithm in Network Intrusion Detection IEEE Conference Publications 2009 pp 849 – 852