Data Mining In Network Anomaly Detection

***Abstract*—We are surrounded by uncountable data in every aspects of our life. The value of those are immeasurable. In this era of internet, data is the main key. Our data is floating. In every moment huge volume of data is sending and receiving. Even though there are lots of mechanism is working for saving those, but during the transferring process data can be looted. There are lots of ways to find or create vulnerabilities. From the very beginning of internet, the intruders are always there. Lots of Intrusion Detection System (IRS) have been introduced to trace them. To build these Data Mining technique played emerging role to heal the vulnerabilities. The IRS need to be up to date as the technology also developing. As always, lots of new data mining frameworks are inventing and those are playing amazing role to detect anomalies.**

This paper reviews the Data Mining frameworks, Algorithms of recent researches along with those accuracy and flexibility in anomaly or intrusion detection which may help the new researcher to understand the current scenario for their work in this field.

***Index Terms*—anomaly-detection, data-mining, network, intrusion.**

I. Introduction

Anomaly detection in network is a very old topic. From the very beginning of the internet it working for data safety. A network has lots of penetrable point. By using those the intruders take access of data. This is the main challenge of anomaly detection to differentiate these intruding behaviors in a network. Anomaly refers to the unexpected behavior or pattern in a system. Which is not an easy term. Sometimes the normal behavior can be detecting as anomaly. As the technology is developing, the risk of intrusion is also increasing. For this, this sector got tremendous interest to researchers. Anomaly detection often known as outlier detection, fraud detection, astonishment deviation and so on based on the research aspects. [1] [2]

Data mining is using vastly in anomaly detection. Lots of research have been done in past and lots are on process. All the four classes of data mining are using efficiently in intrusion detection. Association rule, classification, clustering, regression did well in past researches. Some hybrid approaches did spectacular performance. In recent researches, these hybrid techniques growing interest among researchers. [3] In lots of sector anomaly detection technique playing important role. In health sector, anomaly in health data can be detected. Then in medicine formula the anomalous pattern can be detected. Even in the uncertain health data anomaly detection method did amazing role. [4] In health insurance fraudulent anomaly detection can use. [5] Fuzzy association rules used to detect anomaly in Enterprise Resource Planning (ERP). Proved that, in lower confidence fuzzy association rule did best in accuracy. Around 93% accuracy with 0.3 confidence achieved. [6] In shipping route anomaly, data mining also being used. In that unsupervised approach using clustering technique, the traffic in turning route have been detected. This can use to maintain maritime traffic. [7] Detecting urban anomalies which can reduce accident and death is proposed using decomposing approach in sensitive data. Local Outlier Factor applied to detect anomaly which did well. [8] Data mining and anomaly detection is a great combination. Both have been used in various crucial aspect with great result indeed. In network there are lot of layer. In multi layer, any layer can be intruded. And the data mining is a great healer to that. Hybrid approaches even can detect fraud who use fake account or fake identity.

In this paper we will describe about several anomaly detection techniques which are used in recent works with great accuracy. The researcher who want to do their work in this field, they can get a great visualization from this paper.

II. Data Mining In Anomaly Detection

Several data mining techniques are using to determine anomalies. Some recent detection techniques are analyzed and describe as below:

A. Support Vector Machine (SVM)

SVM is a classification technique where the algorithm finds the best hyperplane which separate two classes. Discrimination line boundary assigned from +1 to -1 for two classes. SVM determine the distance of two hyperplane. In KDDCUP dataset SVM did great. Got 99.29% accuracy with 99.9% of true positive rate and 0% of false positive rate. [9]

B. KNN

Nearest Neighbor is a technique of differentiate same type of data in n classes. User defined the number of K. KNN did very well in intrusion detection. A time sensitive intrusion detection using KNN is proposed and evaluated over stream data. Model applied on KDDCUP99, TAO, SUTTLE dataset. Where from several approaches KNN based proposed approach did well with high time and space efficiency. [10]

C. NN

In Neural Network there are lots of interconnected weighted node which take input, calculate and pass the output through another end. Its like human brain. The hidden layer is defined by the user. It is capable to earn great accuracy. One class Neural Network (OC-NN) performed great to detect anomaly from a complex dataset. The main advantage of it that, the hidden layer can be constructed for detecting specific

anomalies. Applied on four real world and synthetic data set. Where among several approaches (OC-NN) did better. [11]

III. Hybrid Approach to detect Anomaly

A. Nave Bayes + Random Forest

For cloud computing, a system developed which is also time sensitive to detect anomaly. For preprocessing data, a time-based window algorithm used. The preprocessed data then pass through Nave Bayes classifier to detect anomaly. The random forest based ensemble classifier then detect the type of the attack. Which achieve 97% accuracy with 0.21 false positive rate. Detecting several attacks like DOS, Port scan, Ping Scan, Brute Force the accuracy obtained 99.2%, 96.7%, 97.8%, 94.5%. With false positive rate of 0.12, 0.24, 0.22, 0.26. [12]

B. DM-FCC

Density maximization based fuzzy c means clustering technique developed. Which is applied on ADFA dataset. To increase efficiency Membership Matrix Generation (MMG) used. To measure distance Dissimilarity Distance Function (DDF) applied. For 1000 instances of data the model got 32% accuracy and after increasing the instances to 20000 it acquired 95.2992% accuracy. [13]

C. Auto-encoder + Nearest Neighbor

An unsupervised intrusion detection framework named as AEKNN. It operates on two stages. Where is training stage the auto-encoder take place in normal data and in testing stage the KNN do the rest to detect anomaly in compressed hidden layer vector. The model applied in three UCI dataset and the AUC results got in MNIST dataset 0.9700, CARDIO dataset 0.9643 and MAMOGRAPHY dataset 0.9200. Among LOF (Local Outlier Factor), IF (Isolation Forest), RC (Robust Covariance), HBOS (Histogram based Outlier Score) proposed AEKNN did much better. [14]

D. DAGMM

An unsupervised anomaly detection model called Deep Auto-Encoding Gaussian Mixture Model proposed and developed. Lower dimensional representation done by the Deep auto encoding section and then Gaussian Mixture applied on that. The model applied on four benchmark dataset. Those are from UCI repository KDDCUP, from ODDS repository Thyroid, from ODDS Arrythmia, KDDCUP-Rev. From several methods DAGMM got 14% higher f1 rate from all of these dataset. [15]

E. OnCAD

It's the first online clustering algorithm which can find real time anomaly. To detect real time anomalies it follow two approaches. Temporal proximity and spatial proximity do both observation and real time detection. Compared with online K means, ART2 network, OnCAD achieve higher detection rate of 93%. [16]

F. AE + CAE

Proposed and derive a auto encoder based system to detect anomaly. Where auto encoder used for catching Non linear correlation for improving accuracy of detection. Convolutional Auto encoder used for dimensionality reduction. It's advantage is, it need low training time. The model applied on NSL-KDD dataset which achieved better accuracy and low false positive rate compared to the K-NN,SVM,TANN. It got 96.87% accuracy with very low FP rate of 3.44%. [17]

G. AE

Anomaly detection in super computer is not an easy task. A auto-encoder(type of Neural Network) based anomaly detected system proposed for super computer. By training a set of auto encoders the system learn about normal state of a super computer. Then it can differentiate abnormalities. The model applied on D.A.V.I.D.E an energy efficient super computer in Bologna, Italy. After operating on all 45 cores it achieve accuracy between 88% to 96%. [18]

IV. Conclusion

Anomaly Detection is a vast field and its future work will not be over. Because the technology is upgrading and various intrusion vulnerabilities techniques also increasing. Data mining is the best medicine as it is doing the job from the very beginning. Though there are some deep learning algorithms which also doing well, but Data mining has vast area to heal intrusion. Lots of work is in process. In this paper author tried to draw up the very recent techniques or methods that are being applied to detect intrusion. Also the accuracy for specific dataset and the FP and TP rate also analyzed. The researcher who are thinking to do something new in this area, this piece of review paper can be a torch bearer for them.

References

[1] P. Gogoi, D. K. Bhattacharyya, B. Borah and J. K. Kalita, "A Survey of Outlier Detection Methods in Network Anomaly Identification," in The Computer Journal, vol. 54, no. 4, pp. 570-588, Apr. 2011, doi: 10.1093/comjnl/bxr026.

[2] Ajmera Rajesh, Siripuri Kiran "Anomaly Detection Using Data Mining Techniques in Social Networking." International Journal for Research in Applied Science & Engineering Technology (IJRASET) Volume 6 Issue II, February 2018 ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887

[3] Shikha Agrawal, Jitendra Agrawal "Survey on Anomaly Detection using Data Mining Techniques" Procedia Computer Science. Volume 60, 2015, Pages 708-713

[4] Liu, B., Xiao, Y., Cao, L. et al. SVDD-based outlier detection on uncertain data. Knowl Inf Syst 34, 597–618 (2013). https://doi.org/10.1007/s10115-012-0484-y

[5] Guido van Capelleveen, Mannes Poel, Roland M. Mueller, Dallas Thornton, Jos van Hillegersberg "Outlier detection in healthcare fraud: A case study in the Medicaid dental domain" International Journal of Accounting Information Systems Volume 21, June 2016, Pages 18-31

[6] Sarno, R., Sinaga, F. & Sungkono, K.R. Anomaly detection in business processes using process mining and fuzzy association rule learning. J Big Data 7, 5 (2020). https://doi.org/10.1186/s40537-019-0277-1

[7] H. Rong, A.P. Teixeira, C. Guedes Soares "Data mining approach to shipping route characterization and anomaly detection based on AIS data" Ocean Engineering Volume 198, 15 February 2020, 106936

[8] Mingyang Zhang , Tong Li , Hongzhi Shi, Yong Li and Pan Hui "A Decomposition Approach for Urban Anomaly Detection Across Spatiotemporal Data"

[9] Tohari Ahmad and Mohammad Nasrul Aziz. ”DATA PREPROCESSING AND FEATURE SELECTION FOR MACHINE LEARNING INTRUSION DETECTION SYSTEMS”

[10] Guangjun Wu, Zhihui Zhao, Ge Fu, Haiping Wang, Yong Wang, Zhenyu Wang, Junteng Hou, and Liang Huang. ”A Fast kNN-Based Approach for Time Sensitive Anomaly Detection over Data Streams”

[11] Raghavendra Chalapathy, Aditya Krishna Menon, Sanjay Chawla. ”ANOMALY DETECTION USING ONE-CLASS NEURAL NETWORKS”

[12] Mohamed Idhammada, Karim Afdel,Mustapha Belouch. ”Distributed Intrusion Detection System for Cloud Environments based on Data Mining techniques” . Procedia Computer Science Volume 127, 2018, Pages 35-41

[13] Ruby Sharma and Sandeep Chaurasia. ”An Enhanced Approach to Fuzzy C-means Clustering for Anomaly Detection”

[14] J. Guo, G. Liu, Y. Zuo and J. Wu, ”An Anomaly Detection Framework Based on Autoencoder and Nearest Neighbor,” 2018 15th International Conference on Service Systems and Service Management (ICSSSM), Hangzhou, 2018, pp. 1-6, doi: 10.1109/ICSSSM.2018.8464983.

[15] Bo Zongy, Qi Songz, Martin Renqiang Miny, Wei Chengy Cristian Lumezanuy, Daeki Choy, Haifeng Chen. ”DEEP AUTOENCODING GAUSSIAN MIXTURE MODEL FOR UNSUPERVISED ANOMALY DETECTION”

[16] Milad Chenaghlou, Masud Moshtaghi, Christopher Leckie and Mahsa Salehi. ”Online Clustering for Evolving Data Streams with Online Anomaly Detection.”

[17] Z. Chen, C. K. Yeo, B. S. Lee and C. T. Lau, ”Autoencoder-based network anomaly detection,” 2018 Wireless Telecommunications Symposium (WTS), Phoenix, AZ, 2018, pp. 1-5, doi: 10.1109/WTS.2018.8363930.

[18] Andrea Borghesi, Andrea Bartolini, Michele Lombardi, Michela Milano, Luca Benini. ”Anomaly Detection Using Autoencoders in High Performance Computing Systems”