REVIEW OF ANOMALY DETECTION IN NETWORK USING MACHINE LEARNING APPROACH

Dr. Amit Bhusari [1], Prabhanjan Chaudhari [2], Dipali Bhusari [3], Amol Payghan [4]

[1] HOD, Department of MCA, Trinity Academy of Engineering, Yewalewadi, Pune, India
[2] HOD, Department of MCA, Saraswati College, Shegaon, India
[3] Assistant Professor, Trinity Academy of Engineering, Yewalewadi, Pune, India
[4] HOD, Department of BCA, College of Management & Computer Science, Yavatmal, India

CorrespondingAuthor
Dr. Amit Bhusari, aabhusari@gmail.com

DOI
10.29121/shodhkosh.v4.i2.2023.5686

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ABSTRACT

In current era Real time networks using Information and Communication Technology (ICT) has a dominating impact on society, its economy and also security. More generally, ICT reinforce computers, mobile communication devices and networks. Widespread use of ICT is challenged by a group of people with malicious intent, whom we called as network intruders, cyber criminals, etc. Owing to these detrimental cyber attackers and cyber-crimes are of the international priorities and subject to trending research area. Anomaly detection is an important data analysis task which is useful for identifying the network intrusions. This paper is an attempt for reviewing the methods of an anomaly detection. The paper also discusses research challenges with the datasets used for network intrusion detection.

Keywords: Real Time Network, ICT, Anomaly

1. INTRODUCTION

Cyberattacks and networking threats has given edge to emerge new technologies such as Cloud, Fog, Edge computing and Internet of Things (IoT). These attacks are able to penetrate network-related environments, Cloud-based servers, and damage economic source information. Network anomaly detection systems (NADS) play an crucial role in every network defencing system as they monitor network packets to prevent potential threats and users' behavioural abnormalities. Anomaly Detection is the technique of detecting suspicious observations which can raise "anomalous" threat and compromise network security e.g. a credit card fraud, failing machine in a server, a cyber-attack etc.

How to cite this article (APA): Bhusari, A., Chaudhari, P., Bhusari, D., and Payghan, A. (2023). Review of Anomaly Detection in Network Using Machine Learning Approach. *ShodhKosh: Journal of Visual and Performing Arts, 4*(2), 4836–4841. doi: 10.29121/shodhkosh.v4.i2.2023.5686

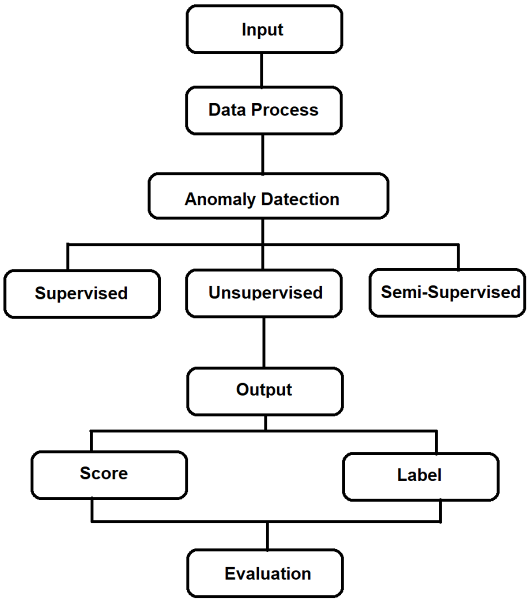

Figure 2.1 Anomaly Detection Architecture

An anomaly can be broadly categorized into three categories–

- **Point Anomaly:** A tuple in a dataset is said to be a Point Anomaly if it is far off from the rest of the data.
- **Contextual Anomaly:** An observation is a Contextual Anomaly if it is an anomaly because of the context of the observation.
- **Collective Anomaly:** A set of data instances help in finding an anomaly.

Machine learning has different approaches for countering the anomalies in networks. Following are the types of anomaly detection methods used in machine learnings.

Unsupervised Anomaly Detection

- Artificial Neural Networks (ANNs)
- Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
- Isolation Forest
- Gaussian Mixture Models (GMM)
- Supervised Anomaly Detection
- Support Vector Machines (SVM)
- Random Forests
- k-Nearest Neighbours (k-NN)

Semi-Supervised Anomaly Detection

Pre-trained Models

Transfer Learning

Anomaly detection is an important way for data analysis that detects anomalous or suspicious data from a given dataset. It is a part of data mining research as it involves discovering enthralling and rare patterns in data. Following are some popular algorithms used by machine learnings which uses large data sets for its effective analysis and pattern analysis.

- **Linear regression**

Linear regression is a type of supervised learning algorithm used for predicting and forecasting values that fall within a continuous range, such as sales numbers or housing prices. It is a statically based technique and is commonly used to justify a relationship between an input variable (X) and an output variable (Y) that can be represented by a straight line. In simple terms, linear regression takes a set of data points with known input and output values and finds the line that best fits those points.

- **Logistic regression**

Logistic regression, also known as "logic regression," is a type of supervised learning algorithm primarily used for binary classification tasks. It is useful when we want to determine whether an input belongs to one class or another. Logistic regression predicts the probability that an input can be categorized into a single primary class. However, in practice, it is commonly used to group outputs into two categories: the primary class and not the primary class.

- **Naive Bayes**

Naive Bayes is a also of category supervised learning algorithms used to create predictive models for binary or multi-classification tasks. It is based on Bayes' Theorem and operates on conditional probabilities, which estimate the likelihood of a classification based on the combined factors while assuming independence between them.

- **Decision tree**

A decision tree is a supervised learning algorithm used for classification and predictive modelling tasks. It resembles a flowchart, starting with a root node that asks a specific question about the data. Based on the answer, the data is directed down different branches to subsequent internal nodes, which ask further questions and guide the data to subsequent branches. This process continues until the data reaches an end node, also known as a leaf node, where no further branching occurs.

- **Random forest**

A random forest algorithm is an ensemble of decision trees used for classification and predictive modelling. Instead of relying on a single decision tree, a random forest combines the predictions from multiple decision trees to make more accurate predictions. In a random forest, numerous decision tree algorithms (sometimes hundreds or even thousands) are individually trained using different random samples from the training dataset. This sampling method is called "bagging." Each decision tree is trained independently on its respective random sample.

K-nearest neighbour (KNN)

K-nearest neighbour (KNN) is a supervised learning algorithm commonly used for classification and predictive modeling tasks. The name "K-nearest neighbour" reflects the algorithm's approach of classifying an output based on its proximity to other data points on a graph.

- **Support vector machine (SVM)**

A support vector machine (SVM) is a supervised learning algorithm commonly used for classification and predictive modeling tasks. SVM algorithms are popular because they are reliable and can work well even with a small amount of data. SVM algorithms work by creating a decision boundary called a "hyperplane." In two-dimensional space, this hyperplane is like a line that separates two sets of labeled data.

- **Apriori**

Apriori is an unsupervised learning algorithm used for predictive modeling, particularly in concern of association rule mining. The Apriori algorithm was discovered in 1990s as a way to map association rules between item sets. It is commonly used in pattern recognition and analysis.

- **Gradient boosting**

Gradient boosting algorithms employ an ensemble method, which means they create a series of "weak" models that are iteratively improved upon to form a strong predictive model. The iterative process gradually reduces the errors made by the models, leading to the generation of an optimal and accurate final model.

2. LITERATURE SURVEY

In this section we have tried to analyse various research work done in the field of anomaly detection using machine learnings.

Catania, C in [2013] described signature based IDS which can be used to automatically detect anomaly. This paper has well described use of IDS to detect anomaly.

O.Y.AL-Jarrah [2014] defines the defence mechanism of cyber intrusion by showing enterprise networks can be protected from attacks and also report (Random Forest-Forword), KDD99 NSL-KDD99 collated and measured effectivity.

Johann Stanek [2015] in this year suggested the way of securing important information of end user over the internet.by the help of DARPA98,NSL-KDD.Shows the importance of security issue.

Noor Moustafa [2015] By the help of machine learning and native base algorithms and decision trees tried to convey the style to apply DARPA 99 data set for network anomaly detection.

Ravi Kiran Varma in[2016] suggested ways of network traffic objectives which used for real time attack detection and IDS algorithm for real time IDS to generate programming output.

Soo-Yeon Ji in[2016] shows a levelled method of intrusion detection by the help of NSL-KDD(having 96% accuracy) for abnormal network traffic. The result is very correct as it denoted but till the end users Data is not secure by using this technique.

Mehdi Hosseinzadeh Aghdam [2016] convey the attribute selection for intrusion detection taction by the use of colony optimization and DARPA98, introduced methodology has less complexity of computational power. Proposed more correctness.

M A Jabbar [2017] described anomaly detection techniques like MR along Data Mining (DM) used (RFAODE) and CAIDA dataset and tried to resolve attribute dependency by 90% accuracy, also combines Forest (FR),(AODE).

Wesam Bhaya in [2017] defines a technique to find DDOS attack cluster studies of huge data by the help of DARPA 200,CAIDA2007,CAIDA2008 set of data.

Suleman Khalid in [2017] explained the way to network security using distributed machine learning over a efficient gateway network used by ISCX-2012,and Linear and Sigmoid Kernel functions.

Tarfa Hamed [2018] described recursive featured addition (NIDS based attribute selection) and bigram approaches, modelling, implementing, developing and testing on ISCX-2012 set of data, Used different metrics.

Chritopher B.Freas in [2018] measured huge threats to a large data also heavy attack response on busy network and used QAD machine learning algorithm and KDD99 and CICIDS2017 dataset.

Nasrin Sultana [2018] described the effectiveness of software networking technology for intrusion monitoring and detection.

Saddam Hossen [2018] explained Machine learning algorithm for examining secure network detection system.

Sidey C Smith [2018] described importance of inconsistent packet header detection with failed network traffic balancing helps to app which working in network attack detection.Gotam Singh Lalotra and Vinod Kumar [2021] put forward iReTADS mechanism to reduce the network traffic using a data summarization technique and also provide network security through an effective real-time neural network.

Shangbin Han, Qianhong Wu and Yang [2022] proposed anomaly detection technique without compromising data quality.

Leandros Maglaras [2019] suggested intrusion detection technique which is combination of various views depending on decision tree, REP tree, forest PA, rule based models, Jrip algorithm with dataset of CICIDS2017.

Novin Anggis [2019] described the approaches for betterment of ada-boost based intrusion detection system result on CICIDS2017 via Synthetic Minority Oversampling Technique Principal Component Analysis.

Kazi Abu Taher in [2019] defined a approach to differentiate network traffic is it malicious or suspicious.it is got like Artificial Natural Network (ANN) based machine learning with Support Vector Machin (SVM)also tried to minimize ratio of cyber-attacks and enhance efficiency.

3. CONCLUSION

This paper is an attempt to address the anomaly detection schemes. As machine learning uses different data set to compare and analyse the pattern, we require large data set for using heuristics approaches of machine learning algorithms as discussed above. In real time network getting and handling such real time data set is challenging.

Also in real time network the anomalies can be launched by any means which can compromise the quality of data. Hence anomaly detection in real time network is need of time and researchers has to find countering schemes for this.

CONFLICT OF INTERESTS

None

ACKNOWLEDGMENTS

None

WORKS CITED

Mohiuddin Ahmed, Abdun Naser Mahmood, Jiankun Hu "A Survey of Network Anomaly detection technique" Journal of Network and Computer Applications 60 (2016) pp19–31

S. S. Panwar and Y. P. Raiwani, “Performance Analysis of NSL-KDD Dataset Using Classification Algorithms with Different Feature Selection Algorithms and Supervised Filter Discretization,” in Intelligent Communication, Control and Devices, Springer, 2020, pp. 497–511.

Gotam Singh Lalotra, Vinod Kumar, Abhishek Bhatt, Tianhua Chen, Mufti Mahmud, "iReTADS: An Intelligent Real-Time Anomaly Detection System for Cloud Communications Using Temporal Data Summarization and Neural Network", Security and Communication Networks, vol. 2022

R. Abdulhammed, M. Faezipour, H. Musafer, and A. Abuzneid, “Efficient network intrusion detection using pca-based dimensionality reduction of features,” in 2019 International Symposium on Networks, Computers and Communications (ISNCC), 2019, pp. 1–6.

S. C. Smith, I. I. Hammell, and J. Robert, “The use of Snap Length in Lossy Network Traffic Compression for Network Intrusion Detection Applications,” J. Inf. Syst. Appl. Res., vol. 12, no. 1, p. 17, 2019.

D. A. Cieslak, N. V Chawla, and A. Striegel, “Combating imbalance in network intrusion datasets.,” in GrC, 2006, pp. 732–737.

R. Bala and R. Nagpal, “A REVIEW ON KDD CUP99 AND NSL-KDD DATASET,” Int. J. Adv. Res. Comput. Sci., vol. 10, no. 2, p. 64, 2019.

V. Kumar, D. Sinha, A. K. Das, S. C. Pandey, and R. T. Goswami, “An integrated rule based intrusion detection system: Analysis on UNSW-NB15 data set and the real time online dataset,” Cluster Comput., pp. 1–22, 2019.

P. Negandhi, Y. Trivedi, and R. Mangrulkar, “Intrusion Detection System Using Random Forest on the NSLKDD Dataset,” in Emerging Research in Computing, Information, Communication and Applications, Springer, 2019, pp. 519–531.

T. Merino et al., “Expansion of cyber attack data from unbalanced datasets using generative adversarial networks,” in International Conference on Software Engineering Research, Management and Applications, 2019, pp. 131–145.

H. P. Vinutha and B. Poornima, “Analysis of NSL-KDD Dataset Using K-Means and Canopy Clustering Algorithms Based on Distance Metrics,” in Integrated Intelligent Computing, Communication and Security, Springer, 2019, pp. 193–200.

O. E. Elejla, M. Anbar, B. Belaton, and S. Hamouda, “Labeled flow-based dataset of ICMPv6-based DDoS attacks,” Neural Comput. Appl., vol. 31, no. 8, pp. 3629–3646, 2019

Shangbin Han, Qianhong Wu and Yang Yang “Machine Learning for Internet of Things anomaly detection under low quality data” International Journal of Distributed Sensor NetworksVolume 18, Issue 10, October 2022

N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, “Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset,” Futur. Gener. Comput. Syst., vol. 100, pp. 779–796, 2019.

S. S. Panwar and Y. P. Raiwani, “Performance Analysis of NSL-KDD Dataset Using Classification Algorithms with Different Feature Selection Algorithms and Supervised Filter Discretization,” in Intelligent Communication, Control and Devices, Springer, 2020, pp. 497–511.

T. Bhaskar, T. Hiwarkar, and K. Ramanjaneyulu, “Adaptive Jaya Optimization Technique for Feature Selection in NSL-KDD Data Set of Intrusion Detection System,” Available SSRN 3421665, 2019.

P. Negandhi, Y. Trivedi, and R. Mangrulkar, “Intrusion Detection System Using Random Forest on the NSLKDD Dataset,” in Emerging Research in Computing, Information, Communication and Applications, Springer, 2019, pp. 519–531.

S. Dwivedi, M. Vardhan, S. Tripathi, and A. K. Shukla, “Implementation of adaptive scheme in evolutionary technique for anomaly-based intrusion detection,” Evol. Intell., pp. 1–15, 2019.

P. S. Chaithanya, M. R. G. Raman, S. Nivethitha, K. S. Seshan, and V. S. Sriram, “An Efficient Intrusion Detection Approach Using Enhanced Random Forest and Moth-Flame Optimization Technique,” in Computational Intelligence in Pattern Recognition, Springer, 2020, pp. 877–884.

N. Sultana, N. Chilamkurti, W. Peng, and R. Alhadad, “Survey on SDN based network intrusion detection system using machine learning approaches,” Peer-to-Peer Netw. Appl., vol. 12, no. 2, pp. 493–501, 2019.

W. A. H. M. Ghanem and A. Jantan, “Training a Neural Network for Cyberattack Classification Applications Using Hybridization of an Artificial Bee Colony and Monarch Butterfly Optimization,” Neural Process. Lett., pp. 1– 42, 2019.

R. Panigrahi and S. Borah, “A detailed analysis of CICIDS2017 dataset for designing Intrusion Detection Systems,” Int. J. Eng. Technol., vol. 7, no. 3.24, pp. 479–482, 2018.

V. Gustavsson, “Machine Learning for a Network-based Intrusion Detection System: An application using Zeek and the CICIDS2017 dataset.” 2019.

N. Bakhareva, A. Shukhman, A. Matveev, P. Polezhaev, Y. Ushakov, and L. Legashev, “Attack Detection in Enterprise Networks by Machine Learning Methods,” in 2019 International Russian Automation Conference (RusAutoCon), 2019, pp. 1–6.

Z. Chiba, N. Abghour, K. Moussaid, A. El Omri, and M. Rida, “An Efficient Network IDS for Cloud Environments Based on a Combination of Deep Learning and an Optimized Self-adaptive Heuristic Search Algorithm,” in International Conference on Networked Systems, 2019, pp. 235–249.

S. Chen, G. I. Webb, L. Liu, and X. Ma, “A novel selective naïve Bayes algorithm,” Knowledge-Based Syst., p. 105361, 2019.

K. J. Mathai, “Performance Comparison of Intrusion Detection System Between Deep Belief Network (DBN) Algorithm and State Preserving Extreme Learning Machine (SPELM) Algorithm,” in 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), 2019, pp. 1–7.

M. Aloqaily, S. Otoum, I. Al Ridhawi, and Y. Jararweh, “An intrusion detection system for connected vehicles in smart cities,” Ad Hoc Networks, vol. 90, p. 101842, 2019.

Z. El Mrabet, H. El Ghazi, and N. Kaabouch, “A Performance Comparison of Data Mining Algorithms Based Intrusion Detection System for Smart Grid,” in 2019 IEEE International Conference on Electro Information Technology (EIT), 2019, pp. 298–303.

M. Mazini, B. Shirazi, and I. Mahdavi, “Anomaly network-based intrusion detection system using a reliable hybrid artificial bee colony and AdaBoost algorithms,” J. King Saud Univ. Inf. Sci., vol. 31, no. 4, pp. 541–553, 2019.

A. Shokoohsaljooghi and H. Mirvaziri, “Performance improvement of intrusion detection system using neural networks and particle swarm optimization algorithms,” Int. J. Inf. Technol., pp. 1–12, 2019.

W. Książek, M. Abdar, U. R. Acharya, and P. Pławiak, “A novel machine learning approach for early detection of hepatocellular carcinoma patients,” Cogn. Syst. Res., vol. 54, pp. 116–127, 2019