Geeta Das, 2021, 9:4
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

Design and Implementation of Data Mining Techniques Based on Machine Learning for IDS

Geeta Das, Prof. Khushbu Rai
Department of Computer Science and Engineering,
Lakshmi Narain College of Technology & Science,
Madhya pradesh India
das.milli7049@gmail.com, Khush.20oct@ gmail.com

Abstract- In today's world, the internet is critical for continuous communication, but its efficiency can decrease the effect known as incursions. Intrusion is any activity that has a negative impact on the targeted system. Because of the fast growth of the Internet, network security has become an increasingly significant problem. The major security defensive mechanism against such malicious assaults is the Network Intrusion Detection System (IDS), which is extensively utilised. By identifying user behaviour patterns from network traffic data, data mining and machine learning technologies have been widely used in network intrusion detection and prevention systems. Data mining for intrusion detection is mostly based on association rules and sequence rules. We present a Length-Decreasing Support to identify intrusion based on data mining, which is an enhanced Data mining Techniques based on machine learning for IDS, in light of the Autoencoder algorithm classical method's bottleneck of frequent itemsets mining. The proposed approach appears to be effective based on test findings.

Keywords: Intrusion detection system, data mining, machine learning.

I. INTRODUCTION

An intrusion detection system (IDS) is a piece of software that monitors and defends a network against attackers. Various application aspects for CNs have evolved as a result of the fast growth of Internet-based technologies.

Business, finance, industry, security, and healthcare are just a few of the LAN & WAN applications that have gained popularity in recent years. All of these uses make area networks a tempting target for misappropriation, putting the community in jeopardy [7].

Malicious operators or hackers use an establishment's internal systems to acquire data, exploit software weaknesses, and exploit administrative problems, then return the system to default settings [8]. New things such as viruses and worms are imported as the Internet becomes more prominent in society. Users can use it to erase passwords and unencrypted text since it is deadly.

As a result, users require security in order to protect their systems against intrusions. Firewall technology is a well-known security method for securing both private and public networks. System-related activities, medical apps, credit card fraud, and insurance firms all utilise IDS [8].

An IDS's purpose is to identify malicious traffic. The IDS does this by keeping track of all incoming and outgoing traffic. The implementation of an IDS may be done in a variety of ways.

Two of these are the most popular: Detecting anomalies: The identification of traffic abnormalities is the basis for this approach. The observed traffic's departure from the usual profile is calculated. Based on the metrics used to measure traffic profile deviation, many implementations of this approach have been offered.

Misuse/Signature Detection: This approach examines network traffic for patterns and signatures of previously identified attacks. The signatures of

International Journal of Science, Engineering and Technology

An Open Access Journal

known attacks are generally stored in a database that is continually updated. The way this approach handles intrusion detection is similar to how anti-virus software works.

II. INTRUSION DETECTION SYSTEM (IDS)

IDS indicates that a theft alarm has been activated. A home lock system, for example, protects a home from theft. The thief alarm detects the lock being broken and triggers an alarm if someone tries to steal into the house by breaking the lock system. Additionally, firewalls are quite good at filtering traffic coming in from the Internet [8].

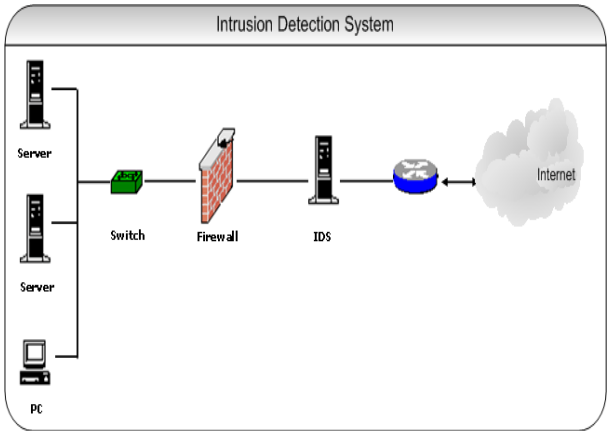

Fig 1. Intrusion Detection system.

External operators, for example, can connect to the intranet via a modem connected to the company's private network; however, this type of access is not protected by a firewall. [8].

An intrusion prevention system (IPS) analyses network traffic in order to detect and prevent vulnerable traffic. The two types of defence systems are network (NIPS) and host defence systems (HIPS).

These systems examine network traffic and take mechanical steps to keep the network and system safe. The IPS problem is riddled with false positives and suggestions. False positives are occurrences that set off an IDS alarm but do not result in a successful attack.

A false negative is an occurrence that does not set off an alarm during an assault. Inline functionality such as single points of failure, signature modifications, and encoded traffic are examples of inline functionality that might cause problems. The system's or n/performance is assessed using IDS.

III. RELATED WORK

When compared to prior single learning model techniques, **W. Zhong, et al [1]** solution .'s can enhance the detection rate of intrusive attacks. When several machines are deployed, the model creation time of BDHDLS is significantly decreased thanks to a parallel training method and large data techniques.

M. P. Bharati et al. [2] For CSE-CIC-IDS-2018, we used an Intrusion Detection System using Machine Learning Based (Random Forest) that provided an outstanding score of 99 percent accuracy.

R. Thomas et al. [3] Using the NSL-KDD data set, we undertake a detailed evaluation of numerous investigations relevant to Machine Learning-based IDS. We propose a general process flow for anomaly-based IDS and discuss the components of this process flow in the context of previous research. Then we suggest some fascinating study topics for the future.

R. Sriavstava et al. [4] This article discusses how cyber security intrusion detection technologies may be combined with Machine Learning and Data Mining approaches.

M. Alloghani et al. [5] Because neural networks are built on multi-layer perceptrons and are the foundation of intelligence, phishing detection will be automated and turned into an artificial intelligence task in the future.

IV. MACHINE LEARNING

The study of computer algorithms that improve themselves over time is known as machine learning. Data mining algorithms that find general principles in big data sets to information filtering systems that automatically learn users' preferences are only some of the applications. Machine learning approaches, in contrast to statistical techniques, are ideally adapted to learning patterns without any prior understanding of what such patterns could be.

The two most common machine learning issues are clustering and classification.

IDSs have been subjected to techniques that solve both of these issues. 1) Techniques of classification: The goal of a classification job in machine learning is

International Journal of Science, Engineering and Technology

An Open Access Journal

to assign each occurrence of a dataset to a certain class. An IDS that uses classification attempts to categorise every traffic as either normal or malicious. The goal is to reduce the amount of false positives (traffic classified as harmful that isn't malicious) and false negatives (classification of malicious traffic as normal).

V. PROPOSED METHODOLOGY

Previously, a hybrid clustering-based method was used, in which the appropriate number of clusters was determined first, then clustering. The optimal number of clusters was estimated using the genetic method, and the data was clustered using Autoencoder algorithm. We opted to utilise more complex clustering techniques, which provide better results than the previous work, due to the constraints of the previous study.

In the suggested technique, we employ the information-gathering strategy to pick attributes. The optimal number of clusters was then calculated using DE, and the data was clustered using the Fuzzy C-means method.

DE is a novel heuristic approach that has three advantages: it finds the original global minimum independent of the initial parameter values, it is quick, and it allows for specific control parameters. The DE algorithm is a population-based algorithm that, like a genetic algorithm, employs crossover, mutation, and selection operators.

The Autoencoder algorithm works by following a set of rules. Based on the distance between the cluster core and the data point, each data point corresponds to a cluster centre. Nearby, you may learn more about the cluster centre and its members. The total membership of each data point must clearly equal one. Update each periodic membership and cluster centre after the formulation:

$$\mu_{ij} = 1 / \sum_{k=1}^{c} (d_{ij} / d_{ik})^{(2/m-1)}$$

$$v_j = (\sum_{i=1}^{n} (\mu_{ij})^m x_i) / (\sum_{i=1}^{n} (\mu_{ij})^m), \forall j = 1, 2, ...c$$

Wherever,
'n' is no. of data points.
'vj' denotes jth cluster center.
'm' is fuzziness index m € [1, ∞].
'c' denotes no. of cluster center
'μij' denotes membership of ith data to jth cluster center
'dij' denotes Euclidean distance amongst ith data & jth cluster center.

Key objective of fuzzy c-means algo is to minimize:

$$J(U,V) = \sum_{i=1}^{n} \sum_{j=1}^{c} (\mu_{ij})^m \left\| x_i - v_j \right\|^2$$

Where,
'||xi – vj||' is Euclidean distance amongst ith data & jth cluster center.

The dataflow diagram of IDS is displayed in Fig.3. The flow diagram shows he steps involved in the implementation of this research work.

IDS can be secluded in following factors:

- Dataset
- Feature Selection
- Training Phase
- Testing Phase
- Classifier

1. Classifier (E):
Intrusion detection systems are tested using classifiers. Whether the algorithm's output is precise or not. ID mapping is used by our classifier to ensure that the output of the class is accurate.

The dataset is compressed into seven attribute datasets after deleting the attribute, one of which is the attribute ID no. For the same example, K refers to the ID number in the CSE-CIC-IDS2018 dataset. To cross-reference and check for accuracy, the ID number of the reference output is used.

Output evaluation involving a variety of variables—

The term "true positive" (TP) refers to a positive no. Classifier tuples that have been properly tagged.

False positive (FP): It refers to the number of negative tuples mistakenly identified by classifiers.

Positive tuples that were mistakenly categorised as negatives are known as False Negatives (FN).

International Journal of Science, Engineering and Technology

An Open Access Journal

Precision is defined as the ratio of true positives to false positives.

Precision = (TP/FP)

Recall is defined as the ratio of true positives to the sum of false positives and false negatives.

Recall = TP/(FP + FN)

Accuracy (ACC): This is the classifier's overall accuracy.

(Precision/Recall) = ACC

Table 1. Association of accuracy of Existing approach and Proposed approach and used dataset CSE-CIC-IDS2018 on AWS.

Algorithm	Precision	Recall	Accuracy
Existing approach	79.345	81.53636	81.48483
Proposed approach	98.242343	95.56643	97.88475

In table 1, we can see the comparison of existing & propose research work i.e, Autoencoder algorithm respectively. The comparison shows that propose work has improved accuracy than that of the previous work along with improved precision & recall.

In figure 2 & fig 3, graph shows fitness values of both research works. The graph shows that proposed approach has higher fitness value than that of the existing approach .

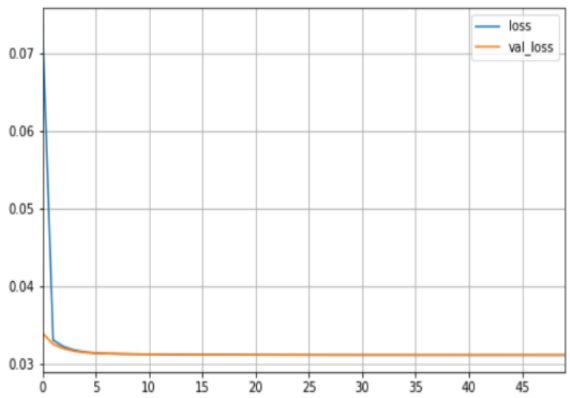

Fig 2. Graph of fitness of CSE-CIC-IDS2018 dataset.

Fitness function used in these algos optimum value of K is create before performing clustering step under training & testing phase.

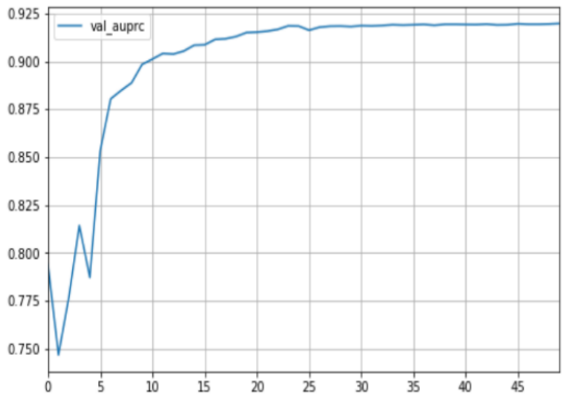

Fig 3. Graph of fitness of CSE-CIC-IDS2018 dataset.

Figures 4 & 5 shows the comparison graphs of the time complexity of the both research works. Time complexity is a concept in computer science that measures the amount of time that a code or algo takes to process or execute as a function of the amount of input..

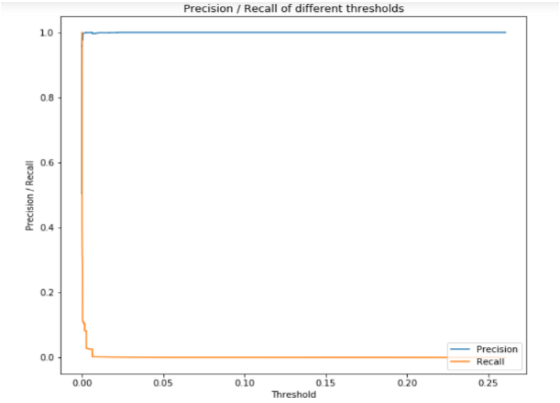

Fig 4. Pression and recall.

From these figure we can see that the time taken by IDEFCM is less than IGKM which is more fast & efficient.

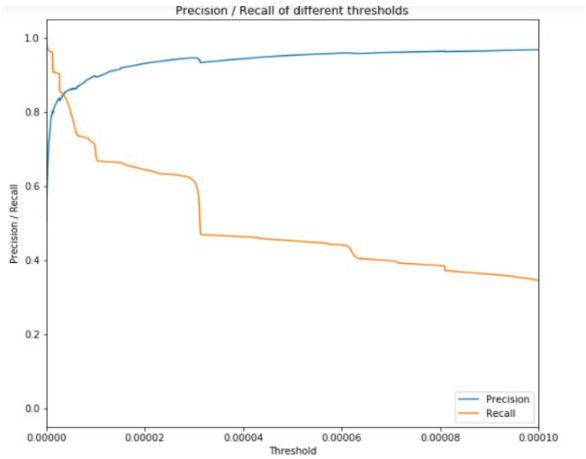

Fig 5. Pression and recall with different parameter.

VI. CONCLUSION

Intrusion crimes are becoming more common by the day. As a result, the optimal intrusion detection system must be identified when compared to intrusion detection systems that use standard clustering algorithms. In this paper, we've built an intrusion detection system that uses the algorithm to determine the type of intrusion, and group no. (k) isn't preset.

The optimum value of K is determined using the fitness function, which aids in the efficient construction of optimised clusters, therefore enhancing the efficiency of type of attack detection. In comparison to intrusion mechanisms.

REFERENCES

[1] W. Zhong, N. Yu and C. Ai, "Applying big data based deep learning system to intrusion detection," in Big Data Mining and Analytics, vol. 3, no. 3, pp. 181-195, Sept. 2020, doi: 10.26599/BDMA.2020.9020003.

[2] M. P. Bharati and S. Tamane, "NIDS-Network Intrusion Detection System Based on Deep and Machine Learning Frameworks with CICIDS2018 using Cloud Computing," 2020 International Conference on Smart Innovations in Design, Environment, Management, Planning and Computing (ICSIDEMPC), 2020, pp. 27-30, doi: 10.1109/IC SIDEMPC49020.2020.9299584.

[3] R. Thomas and D. Pavithran, "A Survey of Intrusion Detection Models based on NSL-KDD Data Set," 2018 Fifth HCT Information Technology Trends (ITT), 2018, pp. 286-291, doi: 10.1109/CTIT.2018. 8649498.

[4] Sriavstava R., Singh P., Chhabra H. (2020) Review on Cyber Security Intrusion Detection: Using Methods of Machine Learning and Data Mining. In: Balas V., Solanki V., Kumar R. (eds) Internet of Things and Big Data Applications. Intelligent Systems Reference Library, vol 180. Springer, Cham. https://doi.org/10.1 007/978-3-030-39119-5_8

[5] Alloghani M., Al-Jumeily D., Hussain A., Mustafina J., Baker T., Aljaaf A.J. (2020) Implementation of Machine Learning and Data Mining to Improve Cybersecurity and Limit Vulnerabilities to Cyber Attacks. In: Yang XS., He XS. (eds) Nature-Inspired Computation in Data Mining and Machine Learning. Studies in Computational Intelligence, vol 855. Springer, Cham. https://doi.org/10.1007/978-3-030-28553-1_3

[6] Helskyaho H., Yu J., Yu K. (2021) Oracle Machine Learning for SQL. In: Machine Learning for Oracle Database Professionals. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-7032-5_3

[7] Monge M., Quesada-López C., Martínez A., Jenkins M. (2021) Data Mining and Machine Learning Techniques for Bank Customers Segmentation: A Systematic Mapping Study. In: Arai K., Kapoor S., Bhatia R. (eds) Intelligent Systems and Applications. IntelliSys 2020. Advances in Intelligent Systems and Computing, vol 1251. Springer, Cham. https://doi.org/10.1007/978-3-030-55187-2_48

[8] Chayal N.M., Patel N.P. (2021) Review of Machine Learning and Data Mining Methods to Predict Different Cyberattacks. In: Kotecha K., Piuri V., Shah H., Patel R. (eds) Data Science and Intelligent Applications. Lecture Notes on Data Engineering and Communications Technologies, vol 52. Springer, Singapore. https://doi.org/10.1007/978-981-15-4474-3_5

[9] Wadiai Y., El Mourabit Y., Baslam M. (2021) Machine Learning for Intrusion Detection: Design and Implementation of an IDS Based on Artificial Neural Network. In: Abraham A., Sasaki H., Rios R., Gandhi N., Singh U., Ma K. (eds) Innovations in Bio-Inspired Computing and Applications. IBICA 2020. Advances in Intelligent Systems and Computing, vol 1372. Springer, Cham. https://doi.org/10.1007/978-3-030-73603-3_19

[10] Hikal N.A., Elgayar M.M. (2020) Enhancing IoT Botnets Attack Detection Using Machine Learning-IDS and Ensemble Data Preprocessing Technique. In: Ghalwash A., El Khameesy N., Magdi D., Joshi A. (eds) Internet of Things—Applications and Future. Lecture Notes in Networks and Systems, vol 114. Springer, Singapore. https://doi.org/10.1007/978-981-15-3075-3_6

[11] Mahmud S., Nuha M., Sattar A. (2021) Crime Rate Prediction Using Machine Learning and Data Mining. In: Borah S., Pradhan R., Dey N., Gupta P. (eds) Soft Computing Techniques and Applications. Advances in Intelligent Systems and Computing, vol 1248. Springer, Singapore. https://doi.org/10.1007/978-981-15-7394-1_5

[12] Wu S., Wang C., Cao H., Jia X. (2020) Crime Prediction Using Data Mining and Machine Learning. In: Liu Q., Mısır M., Wang X., Liu W.

International Journal of Science, Engineering and Technology

An Open Access Journal

(eds) The 8th International Conference on Computer Engineering and Networks (CENet2018). CENet2018 2018. Advances in Intelligent Systems and Computing, vol 905. Springer, Cham. https://doi.org/10.1007/978-3-030-14680-1_40

[13] Gong H.G.J.M., Cui Y., Qian P. (2019) Research on the Key Techniques of Semantic Mining of Information Digest in the Field of Agricultural Major Crops Based on Deep Learning. In: Li D., Zhao C. (eds) Computer and Computing Technologies in Agriculture XI. CCTA 2017. IFIP Advances in Information and Communication Technology, vol 546. Springer, Cham. https://doi.org/10.1007/978-3-030-06179-1_49

[14] Khudadad M., Huang Z. (2018) Intrusion Detection with Tree-Based Data Mining Classification Techniques by Using KDD. In: Gu X., Liu G., Li B. (eds) Machine Learning and Intelligent Communications. MLICOM 2017. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 227. Springer, Cham. https://doi.org/10.1007/978-3-319-73447-7_33