Indian Journal of Cryptography and Network Security (IJCNS)
ISSN: 2582-9238 (Online), Volume-1 Issue-2, November 2021

A Network Intrusion Detection System Based on Categorical Boosting Technique using NSL-KDD

Shiladitya Raj, Megha Jain, Pradeep Chouksey

Abstract: ***Massive volumes of network traffic & data are generated by common technology including the Internet of Things, cloud computing & social networking. Intrusion Detection Systems are therefore required to track the network which dynamically analyses incoming traffic. The purpose of the IDS is to carry out attacks inspection or provide security management with desirable help along with intrusion data. To date, several approaches to intrusion detection have been suggested to anticipate network malicious traffic. The NSL-KDD dataset is being applied in the paper to test intrusion detection machine learning algorithms. We research the potential viability of ELM by evaluating the advantages and disadvantages of ELM. In the preceding part on this issue, we noted that ELM does not degrade the generalisation potential in the expectation sense by selecting the activation function correctly. In this paper, we initiate a separate analysis & demonstrate that the randomness of ELM often contributes to some negative effects. For this reason, we have employed a new technique of machine learning for overcoming the problems of ELM by using the Categorical Boosting technique (CATBoost).***

Keywords: IDS, Network Security, Intrusion Detection, Malicious traffic, Network Traffic Classification.

I. INTRODUCTION

The challenges of network security have also usual greater attention with the exponential growth of the Internet. A significant concern in the area of network security is the study on the detection of an anomaly within a network. Network data are analyzed by IDSs and network anomaly is detected. In general, IDSs may be separated into 2 classes: signature & anomaly systems [1]. IDS based on the Signature [2,3] are structure to detect intrusion by the development of libraries with anomaly behaviour characteristics that matching network data, like Snort intrusion detection systems [3]. These IDSs have a high detection rate, but new network attacks can be difficult to identify. Intrusion detection systems based on Anomaly build models based on normal behaviour & conduct IDS based on which effects are usually devoted.

Manuscript received on 20 October 2021 | Revised Manuscript received on 24 October 2021 | Manuscript Accepted on 15 November 2021 | Manuscript published on 30 November 2021.
* Correspondence Author
Shiladitya Raj, M.Tech, Department of Computer Science, Lakshmi Narain College of Technology Excellence Bhopal (M.P.), India. Email: Raj.shiladitya@gmail.com
Megha Jain*, Assistant Professor, Department of Computer Science, Lakshmi Narain College of Technology Excellence Bhopal (M.P.), India. Email: meghaj@lnct.ac.in
Dr. Pradeep Chouksey, Professor, Department of Computer Science, Lakshmi Narain College of Technology Excellence Bhopal (M.P.), India. Email: pradeepc@lnct.ac.in

Such IDSs are very well-detected for unidentified forms of an anomaly, but their total detection rate is low & the false alarm rate is high.

Researches have been trying to put on several data mining (DM) or machine learning (ML) approaches to IDS to increase the detection rate of IDSs & decrease false alarm rate (FAR).

The vast volume of network data as well as the unbalanced distribution of normal & anomaly activities, therefore, contribute to poor identification & high false alarm in most IDS. An efficient IDS is proposed for sampling and feature selection by using hybrid data optimization. Data sampling is meant to eliminate outliers from the data set & to decrease the detrimental effects on the intrusion detection of unbalanced data delivery.

IDS is differentiated primarily in three forms about identification, configuration and cost. For tracking and recording the incoming and outgoing traffic of networking equipment, NIDS are positioned at various positions in the network. HIDS works on the network's multiple hosts or attached computers. The server or machine is designed and the host is then called [4]. The anomaly-based IDS is utilized to track & compare network traffic with the previously implemented baseline [5].

Data mining and knowledge discovery have gained immense interest in the IT sector. Data mining-based IDS may proficiently know user-interest data and can forecast effects that could later be used. For further research such as predictive analytics, DM is used in IDS as a way of mining features that occur on network traffic data [6]. This is a form of supervised ML algorithm where classification is constructed from data samples or that is applied to forecast unknown class, label classes. There, data sets, whose groups are already defined, are used for training. Multi-classification algorithms are developed by integrating two or more of them.

II. LITERATURE REVIEW

Shen et al. [7] Genetic models, i.e. Multidimensional Assessment (MA) & Secondary Features Extraction and Sampling (SFES) were suggested. The entire database is divided by the MA algorithm into different subsets. By assessing each function independently within different groups, the efficiency of the proposed methodology is improved. At the same time, non-balance, as well as uncertainty, is reduced by the SFES model of the classification algorithm. Classification techniques are usually also supported here. The key emphasis here is to optimise the classification equilibrium in all classes and provide for these systems improved classification efficiency.

Yaseen et al. [8] Multilevel hybrid system dynamically by SVM & ELM adoption. They suggested modifications to SVM & ELM using K-mean algorithms that would minimise the time taken to work with the classifier as well as improve the IDS efficiency to maximise the efficiency of the current classifier model. They utilized KDDcup99 as well as achieved 95.75% accuracy with 1.870 percent of FAR.

Goeschel [9] The method suggested to reduce the false-positive rates and improves IDS performance is novel by placing SVM, DT, or NB together. SVM is initially trained to recognise the attack or regular traffic case. In the next iteration, the DT uses J48 to process first-phase attack-related data to categorise the attack. In the last stage, NB and the DT were used to identify other unclassified attacks.

Gupta et al. [10] NIDS was developed with the aid of data mining methods to protect confidentiality or integrity. Two methods, i.e. clustering or linear regression, were applied in which data analysis was performed using 2 methods, that is data transformation or data normalization. LR gives 80 percent accuracy, while KM gives 67.5 per cent accuracy.

Varma et al. [11] have shown the need for a set of very important functions within the standard IdS preprocessing phase among the already basic features. A brief analysis of different methods of feature selection is provided in this paper by highly focusing on machine learning methods. The outcomes obtained by applying soft computational methods, like rough set theory & ACO, are higher, as opposed to the IDS selection algorithms.

Li et al. [12] The novel raised variant of KNN- TCM has already been suggested, called 'KNearest Neighbor Transductive confidence Machines.' They take the KDD cup99 data set into consideration for anomaly detection by using FS. The chi-square feature ranking system was used to assess the most relevant features. According to performance, the proposed algorithm applies with all characteristics (99.48 percent accuracy & 1.74 percent false positive rate) & data set of top 6 relevant characteristics (99.32 percent accuracy & 2.810 percent FPR).

He et al. [13] understand the essence of the issues in the learning of imbalanced data. They also analysed efficiency under imbalanced learning areas of learning algorithms and proposed new methods for solving the problem of imbalanced learning. They explored briefly the possibilities and obstacles in this area.

Dhote et al. [14] focused on the analysis of three main strategies classified among separate internet traffic categories. Both are referred to as port-based techniques, both based on payload and statistical methods. In this study, which is commonly classified into filters, Wrapper and embedded approaches, feature selection algorithms are also described. These approaches have their benefits and drawbacks along with a quick analysis of the approach for FS that is to be applied now. FS strategies for different ML algorithms are studied. This leads to the study of the recent work performed in this field.

Shetty [15] Genetic Algorithm has been described to build the latest KDD Cup99 data set Network Log Header. It uses 21 of KDD Cup99 Data Set's 41 functions. The use of GA to produce new data connected to new attacks have been simulated in the new headers. On the recently created network data 2 clustering methods, i.e. KM and Kmedoid, have been used and the results correlated with accuracy, detection rate or FPR.

Srivastav & Rama Krishna [16] Layered framework for the neural method is demonstrated to create an effective frame for interruption recognition. Systems are contrasted with current interruption recognition methods, which either use neural networks or take into account the complex architecture. They examined KDD cup99, and the result indicates that the suggested system has a high ID rate as well as a low false alert rate.

III. RESEARCH METHODOLOGY

PROBLEM STATEMENT: One of the key challenges in the intrusion detection method consists of creating helpful behaviour trends or statistics to analyses typical conducts from unusual behaviour by alert collected network dataset. To overcome this problem, previous IDSs normally measure the data set based on the DBN algorithm from security experts or formulate intrusion detection method. After all the data volume, the rise varies quickly. DBN is now an annoying and repetitive task in evaluating and extracting attack signatures or detection rules out of complex data & large network data volumes.

PROPOSED METHODOLOGY: In this paper, the author proposes a network IDS based on Category Boost (CATBoost) classifier. The main advantage of CATBoost is that normalization is not needed as is done with SVM, etc. Trees also do well, if the data is what I call “lumpy”, i.e. non-monotonic. Previously, Extreme Learning Machine was employed on the same dataset i. e, NSLKDD but some disadvantages were found in this technique like One is that ELM's randomness creates more uncertainty both in approximation as well as learning. The other is that ELM has also an inadequate activation function for a general degradation phenomenon. Therefore, CATBoost was taken into consideration because of the regularization function.

CATBoost is the CATBoost classifier to predict categorical features. CATboost is a design for gradient boosts that some decision trees as fundamental predictors. Assume we are observing a Data sample D = $\{(X_J, y_J)\}_{j=1,\ldots,m,}$ in which $X_J = x_j^1, x_j^2, \ldots x_j^n$ is an N-function variable, of solution feature $y_J \in$ R, which may be binary (i.e. yes or no) or encoded as a mathematical feature (0 or 1). The samples (X_J, y_J) are distributed separately by unspecified distribution p according to p(. , .). The purpose of the study task is to train an H: $R^n \rightarrow$ R function that reduces the expected loss (1).

$$L(H) := EL(y, H(X)) \qquad (1)$$

In which the smooth loss feature is L(., .) or (X,y) the evaluation data obtained from training data D is.

The process to gradient boosting concepts iteratively a categorization of approximations Ht : $R^m \rightarrow$ R, t = 0, 1, . . . in a greedy fashion. From preceding approximation H^{t-1} , H^t is gotten in an additive procedure, such that $Ht = H^{t-1} + \alpha g^t$,

with a stage size α & function $g^t : R^n \rightarrow R$, that is base predictor, is selected from a set of functions G to decrease or min expected loss defined in (2)

$g^t = \arg\min_{g \in G} L(H^{t-1}+g) = \arg\min_{g \in G} EL(y, H^{t-1}(X) + g(X))..(2)$

Fig 1 is a schematic diagram of CATBoost showing the significant steps in the implementation of this research.

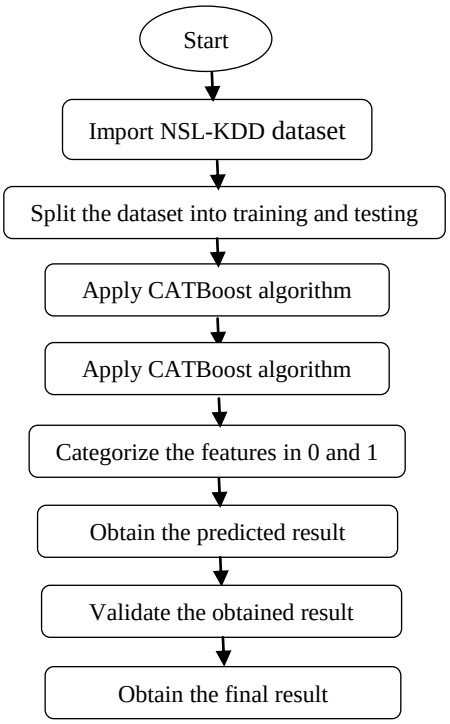

Figure 1: Data Flow diagram of CATBoost

IV. SIMULATION RESULTS

The experiments are carried out on the laptop with the software python-3. The NSL-KDD dataset was included in this test for the use of training and testing.

```
# Compute the error. Results after classification
predictions=elmc.predict(x_test)
accuracy = elmc.score(x_test,y_test)

precision=precision_score(y_test, predictions,average='macro')
recall=recall_score(y_test, predictions,average='macro')

print("Accuracy  : {:.4f}%".format(accuracy*100))
print("Precision : {:.4f}%".format(precision*100))
print("Recall    : {:.4f}%".format(recall*100))
print("--- %s seconds ---" % (time.time() - start_time))
```

```
Accuracy  : 94.8603%
Precision : 74.4291%
Recall    : 66.5068%
--- 17.92579221725464 seconds ---
```

Figure 2: Result visualization of ELM

After the dataset is loaded and each value is assigned a tag, then the data set is divided into two sets, one as the training data & the other as test data. For training purposes, 40 percent of the KDD data set is being used & the remaining 60% is for testing. The algorithm is trained by using a training data set for learning purposes after the division of the KDD data model.

```
# Compute the error. Results after classification
train_predictions=model.predict(train_scaled)
train_accuracy = accuracy_score(y_train, train_predictions)

train_precision=precision_score(y_train, train_predictions,average='macro')
train_recall=recall_score(y_train, train_predictions,average='macro')

print("Train Accuracy  : {:.4f}%".format(train_accuracy*100))
print("Train Precision : {:.4f}%".format(train_precision*100))
print("Train Recall    : {:.4f}%".format(train_recall*100))
print("--- %s seconds ---" % (time.time() - start_time))
```

```
Train Accuracy  : 99.9226%
Train Precision : 79.4254%
Train Recall    : 80.0000%
--- 3.145759344100952 seconds ---
```

Figure 3: Result visualization of CATBoost algorithm

Table 1: Performance Comparison results between exiting ELM & Propose CATBoost

Method	Accuracy	Precision	Recall
ELM	94.86%	74.42%	66.50%
CATBoost	99.92%	79.42%	80.00%

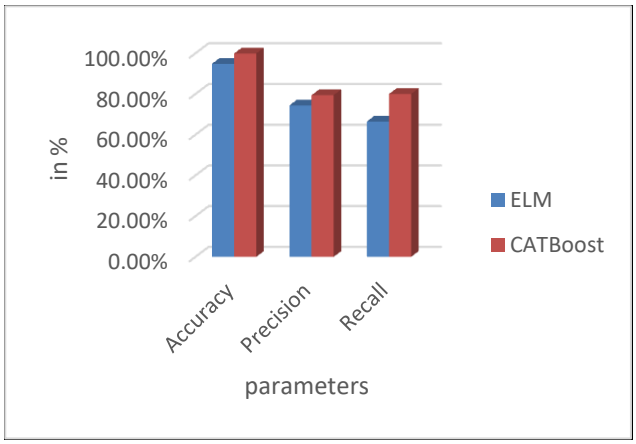

Figure 4: Graphs showing the comparison of ELM and CATBoost algorithms

Compared of both models based on total time as seen in Figure 5, following computational steps of the existing ELM model as well as the proposed CATBoost model.

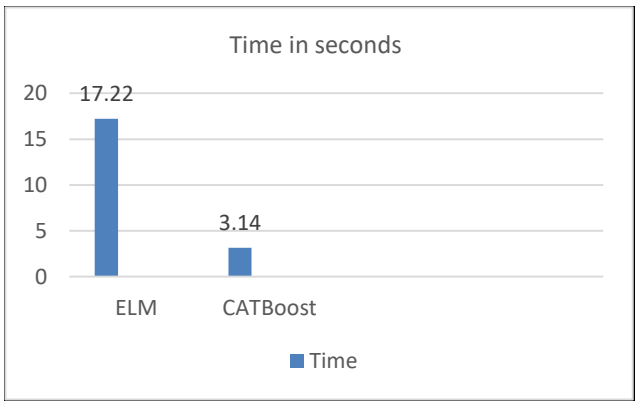

Figure 5 is the comparison chart of the total time taken by both the techniques

V. CONCLUSION

The IDS is developed to provide basic detection strategies to protect the systems in networks connected with the Internet directly or indirectly. Owing to the rise in the volume of sensitive data stored as well as analyzed on the networking systems, demand for IDS & some other security systems has increased enormously over the last decade. IDS aim mainly to overcome safety by detecting intrusion activities that endanger or compromise the system's security, availability, or integrity. To minimize the effect of interference, studies suggest numerous options. In this work we proposed CATBoost algorithm to classify the NIDS. If we equate this model with other parameters, e.g. accuracy, precision, or recall, it can be said that CATBoost Algorithms outstanding than ELM. A high-quality IoT IDS data set is so very important for evaluating as well as validating proposed NIDS. As IoT safety measures are not yet fully established, there is huge scope for future studies, especially in the field of anomaly & intrusion detection with ML & DL methods.

REFERENCES

1. W. Lee, S. J. Stolfo, and K. W. Mok, "A data mining framework for building intrusion detection models," in Proceedings of the 1999 IEEE Symposium on Security and Privacy, pp. 120–132, USA, May 1999.
2. H. P. Sasan and M. Sharma, "Intrusion detection using feature selection and machine learning algorithm with misuse detection," International Journal of Computer Science and Information Technologies, vol. 8, no. 1, pp. 17–25, 2016.
3. J. E. Díaz-Verdejo, P. García-Teodoro, P. Muñoz, G. Maciá-Fernández, and F. De Toro, "A Snort-based approach for the development and deployment of hybrid IDS," IEEE Latin America Transactions, vol. 5, no. 6, pp. 386–392, 2007.
4. P. Mishra, V. Varadharajan, U. Tupakula, and E. S. Pilli, "A Detailed Investigation and Analysis of using Machine Learning Techniques for Intrusion Detection," IEEE Communications Surveys & Tutorials, pp. 1–1, 2018.
5. N. V. Patel, N. M. Patel, and C. Kleopa, "OpenAppID - application identification framework next generation of firewalls," International Conference on Green Engineering and Technologies (IC-GET), pp. 1–5, 2016.
6. V. Bontupalli and T. M. Taha, "Comprehensive survey on intrusion detection on various hardware and software," National Aerospace and Electronics Conference (NAECON), pp. 267–272, 2015.
7. J. Shen, J. Xia, Y. Shan, and Z. Wei, "Classification model for imbalanced traffic data based on secondary feature extraction," IET Communications: IET Journals, vol. 11, no. 11, pp. 1725–1731, 2017.
8. W. L. Al-Yaseen, Z. A. Othman, and M. Z. A. Nazri, "Multi-level hybrid support vector machine and extreme learning machine based on modified K-means for intrusion detection system," Expert Systems with Applications, vol. 67, pp. 296–303, 2017.
9. K. Goeschel, "Reducing false positives in intrusion detection systems using data-mining techniques utilizing support vector machines, decision trees, and Naive Bayes for off-line analysis," SoutheastCon 2016: IEEE, pp. 1–6, 2016.
10. D. Gupta, S. Singhal, S. Malik, and A. Singh, "Network intrusion detection system using various data mining techniques," International Conference on Research Advances in Integrated Navigation Systems (RAINS): IEEE, pp. 1–6, 2016.
11. P. Ravi KiranVarma, V. ValliKumari, and S. Srinivas Kumar, "A Survey of Feature Selection Techniques in Intrusion Detection System: A Soft Computing Perspective," in Advances in Intelligent Systems and Computing, Singapore: Springer Singapore, vol. 710, pp. 785–793, 2018.
12. Y. Li, B. Fang, L. Guo, and Y. Chen, "Network anomaly detection based on TCM-KNN algorithm," 2nd ACM symposium on Information, no. 6, pp. 13-19, 2007.
13. H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263–1284, 2009.
14. Y. Dhote, S. Agrawal, and A. J. Deen, "A Survey on Feature Selection Techniques for Internet Traffic Classification," International Conference on Computational Intelligence and Communication Networks (CICN): IEEE, pp. 1375–1380, 2015.
15. N. P. Shetty, "Using clustering to capture attackers," International Conference on Inventive Computation Technologies (ICICT): IEEE, vol. 3, pp. 1–5, 2016.
16. N. Srivastav and R. K. Challa, "Novel intrusion detection system integrating layered framework with a neural network," in Proceedings of the 2013 3rd IEEE International Advance Computing Conference (IACC), vol. 35, no. 2, pp. 682–689, 2013.

AUTHORS PROFILE

Shiladitya Raj, was born in India on January 8, 1988. He received the Engineering degree of B.E. in Electronics and communication from Swami Vivekananda College of Engineering & Technology Indore in 2009 and Post Graduate degree M.Tech in Computer Science from LNCT Bhopal in 2021.
He is Currently running a successful IT Development and Training Company in Bhopal, named with Saksham Digital Technology from 2018. He had a 11years of experience in IT Industry, in this duration he worked with reputed IT companies and got many Certifications as well. Being android Developer, he developed many Mobile Applications like Cracko, Mobile Tracker, BusRoute, ETC. Now a days he is working on some projects based on Artificial Intelligence and Machine Learning. Also providing training to corporate professionals on Android and IOS Mobile Application Development using trending languages like Flutter, Kotlin, Java, etc. Recently he developed on application where user can prepare for online exams, watch video, read notes and give Test online, with lot of other features. He has a collaboration with **Pearson VUE**, Pearson VUE is the leader in global computer-based testing solutions for academic, government, and professional testing programs, such as skills tests, IT certifications, and real estate licenses.

Megha Jain, was born in India on October 5, 1988. She received the Engineering degree, B.E. in Computer Science and engineering from TRUBA group of Institute Bhopal in 2010 and post graduate degree MTech. in Computer Science and Engineering From SATI Vidisha Bhopal, Madhya Pradesh, India, in 2012 and pursuing P.HD from VIT Bhopal. She is currently a Assistant Professor in the Department of Computer Science and Engineering, at Lakshmi Narain College of Technology, Excellence, Madhya Pradesh, India, She possesses 8 years of teaching experience. She has published more than 10 scientific papers in International and National reputed Journals and conference proceedings in the field of image processing, machine learning. She had been student project mentors under Smart India Hackathon.

Pradeep Chouksey, Born in April 1980 at Bhopal in Madhya Pradesh, he obtained his Post Graduate Degree from Guru Ghasidas Central University (Bilashpur) in 2005. He received his Doctorate Degree in Computer Science from the Samrat Ashok Technological Institute Vidisha with University of Barkatullah, Bhopal in 2011. Prof. Pradeep Chouksey is a man of vision and firm commitment and professional excellence in originations and Institutions to which he has associated himself during his 15 years long professional career.
He is currently a Professor in the Department of Computer Science and Engineering, at Lakshmi Narain College of Technology, Madhya Pradesh, India, and has been the Vice Principal from 2011 to 2019 in reputed institution in M.P. India. He was annual member IACSIT, SDIWC, IAENG, UACEE, CBEES in various countries. He has visited Thailand in international conference. He has successfully supervised 3 Ph. D. research scholars and 6 Ph. D. ongoing under his supervision. He has published 1 book and more than 60 scientific papers in International and National reputed Journals and conference proceedings. in the field of Data Mining, Network Security. Her current research interests Big Data, Deep learning, Machine Learning. He has also published 2 Patents 1 Copyright.