Detection of Unauthorized Access in Bot-IoT Using Machine Learning Algorithm

Mr. S. Arivarasan, Assistant Professor(IT), Department of Information Technology, V.S.B Engineering College, Karur, Tamilnadu, India e-mail: arivarasan.vsbengineering2021@gmAil.com

Ms. R. Kausalya, UG Scholar Department Of Information Technology V.S.B Engineering College, Karur, Tamilnadu, India e-mail: rkausalya07@gmail.com

Ms. K. Priyadharshini, UG Scholar Department of Information Technology, V.S.B Engineering College, Karur, Tamilnadu, India e-mail: dharshpriya2000@gmail.com

Ms. D. Sandhiyai, UG Scholar Department of Information Technology, V.S.B Engineering College, Karur, Tamilnadu, India e-mail: sandhiyatechnogirl124@gmail.com

Ms. S. Samrin, UG Scholar Department of Information Technology, V.S.B Engineering College, Karur, Tamilnadu, India e-mail: s.samrinsheik@gmail.com

ABSTARCT-

Identification of anomaly and malevolent traffic in the Internet of things (IoT) network is essential for the IoT security to keep eyes and block unwanted traffic flows in the IoT network. For this purpose, several machine learning (ML) technique models are presented by many researchers to block malevolent interchange flows in the IoT network. However, due to the inappropriate feature selection, several ML models prone misclassify mostly malicious traffic flows. Nevertheless, the significant problem still needs to be studied more in-depth that is how to select effective features for accurate malicious traffic

Firstly, a novel feature selection metric approach named CorrAUC proposed, and then based on CorrAUC, a new feature selection algorithm name Corrauc is develop and design, which is based on wrapper technique to filter the features accurately and select effective features for the selected ML algorithm by using AUC metric. Then, we applied integrated TOPSIS and Shannon Entropy based on a bijective soft set to validate selected features for malicious traffic identification in the IoT network. We evaluate our proposed approach by using the Bot-IoT dataset and four different ML algorithms. Experimental results analysis showed that our proposed method is efficient and can achieve >96% results on average.

1. INTRODUCTION

The connected world which began with representative services such as connected cars, networked unmanned aerial vehicles, multi-robot systems, and the Internet of things (IoT), results in networks with inherent dynamics. The network entities of such services generally have high mobility, which causes frequent changes in member nodes associated with these networks. The links between the network entities may be under unstable channel conditions with high link failure rates. In wireless ad hoc networks, for example, source nodes are connected to terminal nodes via mobile intermediate nodes. The network topology can be determined in a distributed manner based on decisions of mobile intermediate nodes for transmission ranges. In order to overcome network dynamics, each mobile intermediate node may strategically change its transmission range. Such distributed strategies for network formation can be essential in circumstances where only limited infrastructures can be available, e.g., disaster networks or military networks

The potential of the IoT to drive a sustainable everyday life is more than probable. This fact is easily evidenced through its current application domains such as agriculture, energy saving at home or in industrial settings and the pollution and traffic control within the cities. One example of such potential is the Google's Nest Thermostat, perhaps the most famous IoT gadget during 2014. Their designers disclosed that it can become carbon neutral in a period of just eight weeks after its first usage. Carbon neutrality refers to the greenhouse gases that were created by manufacturing and distributing the device are offset by the energy savings one obtains from using it3 However, it is still controversial how other myriads of IoT devices (everyday consumer appliances, fitness trackers or kitchen appliances) can be also labeled as green devices along their life-cycle: from manufacturing to disposal4. These new devices are designed to replace old-fashioned ones. Therefore, their inclusion will rise to an augment of electronic waste that probably will end in the landfill. This paper describes the implementation of an approach that addresses this latter IoT challenge. Our proposal lies in two pillars: First, it is focused on embedding intelligence through open hardware electronics within everyday appliances of shared use (e.g. beamers, coffee-makers, printers, screens, portable fans, kettles, etc.). Our aim is transforming these electronic devices into Internet- connected eco-aware everyday things rather than replacing them by new ones. As a proof of concept, in the presented work we have focused on electronic coffee machines located in four different work-laboratories. The second pillar, it is to design and implement a RESTful infrastructure that enables to these eco-aware appliances to reduce their energy waste. It is devised to intelligently process in the back-end the most efficient operation mode at any time for each shared de- vice and to give back such information to them, i.e. the appliances are able to operate autonomously in an eco-friendly manner. We have named this architecture ARIIMA (the capital letters of the six former words in the paper title) as an analogy with the predictive model used to forecast the appliance's usage, ARIMA model. The presented paper makes a reality the theoretical design reported in a previous author article [7] by implementing it.

2 .RELATED WORK

From the last decade, security and trust problems develop a scorching topic, and many researcher endeavour's hard to overcome this problem and proposed numerous effective models along with the future Internet [1], [2], IoVs [3], wireless sensor network (WSN) [4], [5] and IoTs. However, some most viewed and cited studies related to feature selection for malicious Bot-IoT in IoT networks are discussed in this section. In our recent study work [12], for the optimum feature selection problem in Instant Messaging (IM) applications traffic classification, a feature selection technique is proposed based on mutual information (MI) analysis technique. Most of the public IDS are based on pattern matching and are static approches. Attacks are becoming sophisticated and attackers can circumvent rules-based detection techniques. Thus, machine learning based approaches are thought to be promising.

In recent years, as more and more IoT devices are actually deployed, IDS in IoT environments has attracted attentions from many researchers and developers, the researchers addressed specific types of threats target specific devices. The work proposed a detection system for sinkhole which is the attack on routing devices. The detection rate for sinkhole attacks is up to 92% and 72% on fixed and mobile scenrio, respectively. The work tried to prevent the three different levels of battery exhaustion attack on BLE based mesh network.
However, from the experimental results analysis, the proposed approach achieves beneficial performance results by using the selected feature set for the IM application traffic identification.

The technique of feature selection is handy for enhancing ML performance. However, feature selection is a process to select the optimum features set from several features set and removed the features that don't carry enough identification information for the identification or removing the redundant feature. SE in 2018 [8] studied mostly cited research studies related to feature selection technique, especially the correlation coefficient technique, and propose a new feature selection technique named Fast Based Correlation Features (FCBF) algorithm for the improvement of the performance of IoT network in

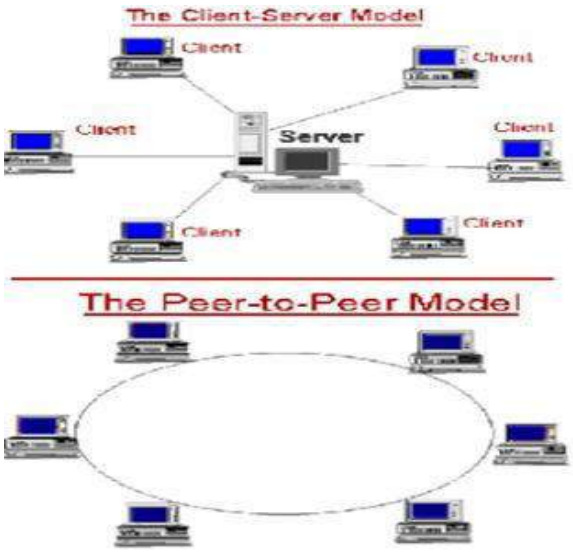
the industrial environment. The main contribution of their study is to split the feature space into several equal parts with equal size. Using the proposed approach, they showed enhanced results of correlation ML of every running node in the IoT network. They showed that their experimental results are effective, and the proposed approach is able to achieve effective performance results in terms of accuracy and execution time, which is very important for accurate identification. Similarly, in 2018 Meidan Yair at el. [7] studied the detection of attacks in IoT network and proposed a new technique to overcome the problem of attacks which is initiated by the Internet of Things (IoT) devices and then for the identification of anomalies in IoT traffic they used auto encoder. The dataset that they used in their study for the evaluation of their proposed approach is botnet attacks Bashlite and Mirai based on the Internet of Things. However, the utilized datasets are also included on several infected devices in the IoT network. They showed in their study that the proposed approach is able to detect cyber-attacks in IoT network devices with high-performance results. In their experimental analysis, they showed that their proposed technique is effective for IoT performance enhancement and anomaly detection in IoT networks. More in-depth, numerous IoT security technique can be applied for the accurate cybersecurity purpose in IoT security environment, for instances, cyber-attacks identification in [9], [10], effective management scheme [11], [12], evidence framework etc. However, the above numerous techniques proposed by many researchers are effective, but it is important to select the most effective feature set that carries accurate information for the Bot-IoT attack detection in the IoT environment. The necessary key process of feature selection technique includes on different important steps such as trace traffic, to trace the original traffic, subset generation, to generate features set from the trace traffic

3. EXISTING SYSTEM

The beginning of network coding was for throughput gain in a multicast scenario. It is shown that network coding can achieve maximum throughput via the max-flow min-cut theorem, and it is further proved that linear network coding can achieve the upper bound of capacity. Many works in network coding have been studied for random linear network coding (RLNC) as it is a simple, randomized encoding approach that is decentralized. As well as throughput gain, it has been shown that network coding also enhances robustness against packet loss in lossy wireless networks. Another advantage of network coding is that there is a lower complexity requirement for network formation compared to a conventional store-and-forward approach. In a conventional store-and-forward approach, it is difficult to find the optimal routing path that can achieve the capacity upper bound. Even though network coding can reduce complexity in general, it is known that finding an optimal solution in network coding with multiple multicasts is an NP-hard problem. Hence, suboptimal but practical solutions are often studied.

4. PROPOSED SYSTEM

The proposed system consists of two phases: initialization and adaptation. In the initialization phase, the optimal policy for each intermediate node is found and the state can be initialized. The optimal actions can lead the network formation result to certain topologies, referred to as stationary networks. Hence, the design the initial network to be close to the stationary network. In the adaptation phase, each node adaptively and optimally changes its transmission range based on the optimal policy for the current state induced by network dynamics. In our proposed system one of the advantages are our simulation results confirm that the proposed strategy builds a network which adaptively changes its topology in the presence of network dynamics. Moreover, the proposed strategy outperforms existing strategies in terms of system good put and successful connectivity ratio.



5. METHODOLOGY AND ALGORITHM

ALGORITHM: OPTIMAL POLICY

1 Initialize: s sy // build a stationary network

2 while network is active do

3 // receive and combine packets store received packets in buffer

4 if Li 6= ; then

5: //if the buffer is not empty build a network coded packet based on (4)

6: // update network topology check the current states

7: find the optimal action: a (s)

8: update the transmission range: a+1 a+a

9: broadcast the network coded packet

10: + 1

Accuracy: In attacks detection, it can be described as the correctly identified samples of traffic in overall identified samples traffic. However, using performance measurement metrics, the accuracy can be defined mathematically

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

Specificity: In this research study, we used the specificity metrics which can defined as the ability of a machine learning classifiers to detect negative results. The mathematical equation of specificity.

$$\text{Specificity} = \frac{TN}{(FP + TN).}$$

Introductory definitions: In this subsection, the introductory definition and basic operations of the soft set are discussed with details.

a) Soft Set [48]: If U is the universal set and S is it's parameter then U be P(U) and X will subset of S, for example,X ⊂ S. At that point, pair (F, X) will be soft set over U, and function F will F : X→ P(U).

b) Bijective Soft Set: If (F,S) is a soft set and U is the universal set and it's parameters is S respectively then (F,S) isknown as Bijective soft set if the below two given condition are true:

i. UB F(β) = U

ii. For two features;

βi, βj, βj ∈ S, βi = βj, F(βi) Ñ F(βj =)

2.Method

Input: Set of features of dataset Output: Desired Selected Effective feature set

a) Identify a features set based on Bot-IoT attacks and normal traffic in IoT network environment.

b) The soft set will be developed from the identified set of features from each feature, which is the most effective and discard others. However, these function concepts are a theoretical concept that is effective for a better understanding.

c) After the second step completion, feature set values are represented in the soft set and bijective soft set respectively for the decision making.

d) Generate feature preference for the expert and then make a decision matrix as EPDM = [ρi j] a×b, where i = 1, ..., a and j = 1, ...b; ρi j. M indicate the number of experts, while n indicated numbers of features.

This section includes traces traffic, evaluation criteria, and analysis of experimental results.

Table 1: Characteristics of HIT Trace 1 dataset.

Application	Duration	# instances	Date
WTCP	1 hr	20512	28 April 16
WUDP	1 hr	16400	28 April 16
P2P	1 hr	1501	27 Dec 15
IM	1 hr	7911	27 Dec 16
IMAP	1 hr	15832	27 Dec 15
FTP	1 hr	25251	27 Dec 15

5.1 Datasets: In this paper, we select two sets of network traces for our experimental study. One dataset is our set of traces collected in our lab, while the other set is an open network trace dataset. The selected two traces are different network environment datasets. We applied our proposed feature selection algorithm on both datasets, respectively, notion only one dataset for better understanding of the composition of Internet traffic. We used two different network environment datasets, because these datasets are different from each other; for example, in our trace dataset, we capture mostly WeChat instant messaging application's traffic, while in NIMS dataset GTALK IM application's traffic is traced.

6. RESULT ANALYSIS AND DISCUSSION

Though the results of the five applied machine learning classifiers are different with respect to accuracy, recall, and precision using HIT Trace 1 dataset and NIMS dataset, some information can be obtained from experimental study for IM traffic classification:

(i) From this study, it is clear that our proposed algorithm selects effective features set for IM traffic classification using two different network environment datasets in terms of classification accuracy, recall, and precision metrics.

(ii) From the experimental results, all the applied machine learning classifiers give very effective performance results for all application applications are classifications, but only FTP and Telnet classified a little bit low in both utilized datasets as compared to other applications.

(iii) In this research study, our proposed algorithm gives effective features sets and it is evident that all the features carry enough identification information for IM traffic classification.

(iv) Through accuracy results, the classification performance can be easily evaluated for the instant messaging (IM) traffic classification. But, in some cases, some classifiers get high identification performance results and in some cases they do not get very effective results. It is due to imbalance traffic composition found in the datasets.

(v) We discuss that all the applied ML classifiers give very effective performance results. However, decision tree and Random Forest ML classifiers give very accurate performance results as compared to other machine leaning classifiers.

7. LITERATURE SURVEY

[7.1] TITLE: Decoding Delay Controlled Completion Time Reduction in Instantly Decodable Network Coding.

DESCRIPTION: For several years, the completion time and the decoding delay problems in Instantly Decodable Network Coding (IDNC) were considered separately and were thought to act completely against each other. Recently, some works aimed to balance the effects of these two important IDNC metrics but none of them studied a further optimization of one by controlling the other. This paper investigates the effect of controlling the decoding delay to reduce the completion time below its currently best-known solution in both perfect and imperfect feedback with persistent erasure channels. To solve the problem, the decoding delay-dependent expressions of the users' and overall completion

times are derived in the complete feedback scenario. The gap in performance becomes more significant as the erasure of the channel increases.

[7.2] TITLE: Virtual Overhearing: an Effective Way to Increase Network Coding Opportunities in Wireless Ad-Hoc Networks

DESCRIPTION: Overhearing is of great importance to wireless network coding in that it can be exploited to obtain the side information needed for packet decoding. Recently, a new technique called virtual overhearing (VOH) was proposed to allow a node to obtain the packet sent by another node that is multiple hops away for free. This can overcome the limitation of overhearing and be used to discover more coding opportunities. In this paper, we take advantage of VOH and propose two modes of exploiting VOH to increase coding opportunities in wireless ad-hoc networks. First, we make use of VOH to increase the chance of finding a route with coding opportunities for a new incoming flow. Second, and more importantly, we make use of VOH to create coding opportunities between two established flows which are currently unmixable. Note that most previous studies only attempt to find coding opportunities rather than create them. Based on these two modes of VOH usage, we design two routing protocols: distributed coding-aware routing with virtual overhearing (DCAR-VOH), and its enhanced version DCAR-VOH.

[7.3] TITLE: Random Linear Network Coding for Wireless Layered Video Broadcast: General Design Methods for Adaptive Feedback-free Transmission

DESCRIPTION: The problem of broadcasting layered video streams over heterogeneous single-hop wireless networks using feedback-free random linear network coding (RLNC). We combine RLNC with unequal error protection (UEP) and our main purpose is twofold. First, to systematically investigate the benefits of UEP+RLNC layered approach in servicing users with different reception capabilities. Second, to study the effect of not using feedback, by comparing feedback free schemes with idealistic full-feedback schemes. To these ends, we study 'expected percentage of decoded frames' as a key content-independent performance metric and propose a general framework for calculation of this metric, which can highlight the effect of key system, video and channel parameters. We study the effect of number of layers and propose a scheme that selects the optimum number of layers adaptively to achieve the highest performance. Assessing the proposed schemes with real H.264 test streams, the trade-offs among the users' performances are discussed and the gain of adaptive selection of number of layers to improve the trade-offs is shown.

[7.4] TITLE: Network Coding Based Evolutionary Network Formation for Dynamic Wireless Networks

DESCRIPTION: We aim to find a robust network formation strategy that can adaptively evolve the network topology against network dynamics in a distributed manner. We consider a network coding deployed wireless ad hoc network where source nodes are connected to terminal nodes with the help of intermediate nodes. We show that mixing operations in network coding can induce packet anonymity that allows the inter-connections in a network to be decoupled. This enables each intermediate node to consider complex network inter-connections as a node-environment interaction such that the Markov decision process (MDP) can be employed at each intermediate node. The optimal policy that can be obtained by solving the MDP provides each node with optimal amount of changes in transmission range given network dynamics (e.g., the number of nodes in the range and channel condition). Hence, the network can be adaptively and optimally evolved by responding to the network dynamics. The proposed strategy is used to maximize long-term utility, which is achieved by considering both current network conditions and future network dynamics.

8. CONCLUSION

Detection of attacks in the Internet of things (IoT) network is essential for the IoT security to keep eyes and block unwanted traffic flows. Numerous machine learning (ML) technique models are presented by many researchers to block attack traffic flows in the IoT network. However, due to the inappropriate feature selection, several ML models prone misclassify mostly malicious traffic flows.

Nevertheless, the noteworthy problem still needs to be studied more in-depth, that is how to select effective features for accurate malicious traffic detection in IoT networks. For this purpose, a new framework model is proposed. novel feature selection metric approach named CorrAUC proposed, and then based on CorrAUC, a new feature selection algorithm name Corrauc is develop and design, which is based on wrapper technique to filter the feature accurately and select effective features for the selected ML algorithm by using AUC metric.

7. REFERENCE

[1] M. Shafiq, Z. Tian, A. K. Bashir, A. R. Jolfaei, and X. Yu, "Data mining and machine learning methods for sustainable smart cities traffic classification: A survey," Sustainable Cities and Society, 2020.

[2] Y. Xiao, X. Du, J. Zhang, F. Hu, and S. Guizani, "Internet protocol television (IPTV): the killer application for the next-generation internet," IEEE Communications Magazine, vol. 45, no. 11, pp. 126–134, 2007.

[3] Z. Tian, S. Su, W. Shi, X. Du, M. Guizani, and X. Yu, "A data-driven method for future internet route decision modeling," Future Generation Computer Systems, vol. 95, pp. 212–220, 2019.

[4] Z. Tian, X. Gao, S. Su, J. Qiu, X. Du, and M. Guizani, "Evaluating reputation management schemes of internet of vehicles based on evolutionary game theory," IEEE Transactions on Vehicular Technology, 2019. 68(6): 5971-5980.

[5] Y. Xiao, V. K. Rayi, B. Sun, X. Du, F. Hu, and M. Galloway, "A survey of key management schemes in wireless sensor networks," Computer Communications, vol. 30, no. 11-12, pp. 2314–2341, 2007.

[6] X. Du and H.-H. Chen, "Security in wireless sensor networks," IEEE Wireless Communications, vol. 15, no. 4, pp. 60–66, 2008.

[7] S. Egea, A. R. Mañez, B. Carro, A. Sánchez-Esguevillas, and J. Lloret, "Intelligent iot traffic classification using novel search strategy for fastbased- correlation feature selection in industrial environments," IEEE Internet of Things Journal, vol. 5, no. 3, pp. 1616–1624, 2018.

[8] Y. Meidan, M. Bohadana, Y. Mathov, Y. Mirsky, A. Shabtai, D. Breitenbacher, and Y. Elovici, "N-baiotâ˘A ˇ Tnetwork-based detection of iot botnet attacks using deep autoencoders," IEEE Pervasive Computing, vol. 17, no. 3, pp. 12–22, 2018.

[9] S. Su, Y. Sun, X. Gao, J. Qiu, and Z. Tian, "A correlation-change based feature selection method for iot equipment anomaly detection," Applied Sciences, vol. 9, no. 3, p. 437, 2019.

[10] Q. Tan, Y. Gao, J. Shi, X. Wang, B. Fang, and Z. H. Tian, "Towards a comprehensive insight into the eclipse attacks of tor hidden services," IEEE Internet of Things Journal, 2019. vol. 6, no. 2, pp. 1584-1593, April.

[11] Z. Tian, W. Shi, Y. Wang, C. Zhu, X. Du, S. Su, Y. Sun, and N. Guizani, "Real time lateral movement detection based on evidence reasoning network for edge computing environment," IEEE Transactions on Industrial Informatics, 2019. Vol 15(7): 4285-4294.

[12] X. Du, Y. Xiao, M. Guizani, and H. Chen, "An effective key management scheme for heterogeneous sensor networks," Ad Hoc Networks, vol. 5, no. 1, pp. 24–34, 2007.