The Korean Institute of Communications and Information Sciences

Available online at www.sciencedirect.com

ScienceDirect

ICT Express 7 (2021) 177–181

www.elsevier.com/locate/icte

Feature selection for intrusion detection system in Internet-of-Things (IoT)

Pushparaj Nimbalkar[*], Deepak Kshirsagar

Department of Computer Engineering and IT, College of Engineering Pune, India

Received 31 January 2021; received in revised form 25 March 2021; accepted 30 April 2021
Available online 7 May 2021

Abstract

Internet of Things (IoT) is suffered from different types of attacks due to vulnerability present in devices. Due to many IoT network traffic features, the machine learning models take time to detect attacks. This paper proposes a feature selection for intrusion detection systems (IDSs) using Information Gain (IG) and Gain Ratio (GR) with the ranked top 50% features for the detection of DoS and DDoS attacks. The proposed system obtains feature subsets using insertion and union operations on subsets obtained by the ranked top 50% IG and GR features. The proposed method is evaluated and validated on IoT-BoT and KDD Cup 1999 datasets, respectively, with a JRipclassifier. The system provides higher performance than the original feature set and traditional IDSs on IoT-BoT and KDD Cup 1999 datasets using 16 and 19 features, respectively.

Keywords: Denial-of-service; Internet of Things; Feature selection; Intrusion detection system

1. Introduction

Internet of Things (IoT) is a globally adopted technology in automated network systems. The next stage of the Information Technology (IT) rising and interconnectivity is the IoT, from little toy to homemade application to the smart city in IoT. IoT is a mixture of cloud-connected embedded systems used by the consumer to access IT-related services utilizing the combination of electronics-related things and internet protocol.

In IoT systems, protocols used may have security vulnerabilities [1] that can impact the whole system. IoT devices are vulnerable targets for cybercriminals and attackers because of their lack of fundamental security protocols. That implies that they can be hacked and attacked by botnets, which are used to initiate DDoS against organizations.

The noisy captured network traffic in IoT consists of a large number of traffic features. The machine learning models require more time to build models and affect the performance of IDS due to the presence of a large number of features in IoT network traffic. Therefore, feature selection is required for intrusion detection in IoT that builds the models in minimum time and achieves higher performance.

The contributions of this paper are summarized as follows:

1. This paper proposes a feature selection method using Information Gain (IG) and Gain Ratio (GR) with top 50% ranked features for IDS in IoT.
2. The proposed feature selection method is tested on BoT-IoT dataset and validated on the prominent KDD Cup 1999 dataset.
3. The proposed feature selection method provides higher performance with JRip classifier on BoT-IoT and KDD Cup 1999 datasets using obtained features in minimum model build time.
4. The proposed system is also compared to the existing systems on BoT-IoT and KDD Cup 1999 datasets.

2. Literature review

DDoS attacks features and principal component analysis (PCA) [2] is presented to detect DDoS attacks in IDS. The system achieved higher precision of 92% with Mahalanobis Distance (MD) using reduced features on KDD Cup 1999 dataset. The network IDS presented in [3] with K-means clustering achieved higher detection of 96.8% with K-means clustering using manually selected 8–16 features. The work [4]

[*] Corresponding author.
E-mail addresses: nimbalkarpr19.comp@coep.ac.in (P. Nimbalkar), ddk.comp@coep.ac.in (D. Kshirsagar).

Peer review under responsibility of The Korean Institute of Communications and Information Sciences (KICS).

https://doi.org/10.1016/j.icte.2021.04.012

2405-9595/

used PCA in anomaly IDS with softmax regression and provided a detection rate of 99.31% with 1.116% False Alarm Rate (FAR) using ten reduced features.

The work [5] used particle swarm optimization (PSO) for feature selection and provided recall, precision, and accuracy of 99.5%, 99.6%, and 99.6% respectively on the NSL-KDD dataset with deep neural network (DNN) using selected features. The work [6] presented multi-layered framework using deep learning-based IDS. The system obtained 26 selected features using Cohen's Kappa coefficient and Mathew correlation methods and produced higher accuracy of 98.27% for the detection of DoS attacks with Recurrent Neural Network (RNN).

The study [7] proposed IDS with an ensemble classifier using the feature selection method. The system used correlation coefficient (CC) for feature selection and an ensemble classifier of Naive Bayes, decision tree (DT), and artificial neural network (ANN) for intrusion detection. The system achieved 98.54% accuracy for detecting DoS attacks on UNSW-NB 15 datasets with an ensemble classifier using selected features. The system [8] achieved higher accuracy of 89.76% with a higher FAR of 1.68 using top-ranked 13 IG features with C5 classifier. The work [9] proposed the use of IG in feature selection for IDS. The system produced higher accuracy of 93.23% with 6.77% FAR with C4.5 classifier using top ten ranked features of IG.

The system [10] obtained six reduced features using the multi-objective feature selection technique. The system provided 99.90% accuracy with an extreme learning machine (ELM) classifier on CICIDS 2017 dataset. The study [11] proposed Long Short-Term Memory (LSTM) networks in IDS to detect cyber attacks and achieved an accuracy of 99.91% and 98.22% on ISCX and AWID datasets, respectively, with LSTM using Stochastic Gradient Descent (SGD) parameter optimization.

The work [12] presented layer architecture used the top 10 ranked features of the GR method and tested them on a simulated dataset. The architecture achieved higher performance with the J48 classifier than other tree and rule-based classifiers for detecting DoS attacks. The work [13] presented a multilayer framework to detect DDoS attack with a decision tree. The system manually selected eight features and achieved an accuracy of 99.98% on a simulated dataset to detect ICMP, TCP, and UDP flood attacks. The study [14] used nature-inspired algorithms for feature selection on NS-3 simulated dataset with Forecasting and Chaos approach. The system achieved a detection rate (DR) of 94.3% for the detection of transport and application layer DoS attacks. The work [15] used the wrapper feature selection method for feature selection in IDS. The system produced a higher accuracy of 97.39% with support vector machine (SVM) on the honeypot Cowrie dataset, including Spying, SSH, and XOR DDoS attacks.

The literature review observed that KDD Cup 199, CICIDS and ISCX datasets are non-IoT intrusion datasets, i.e., network intrusion datasets that include HTTP DoS attacks. AWID dataset consists of Madiun Access Contol (MAC) Layer attacks related to the IEEE 802.11. The Cowrie dataset consists of DDoS attacks related to authentication in IoT. IoT-BoT is the latest dataset, especially for IoT, with distinct features compared to network intrusion datasets that include application and transport layer DoS and DDoS attacks. This study proposes the feature selection-based IDS to detect DoS and DDoS attacks in IoT.

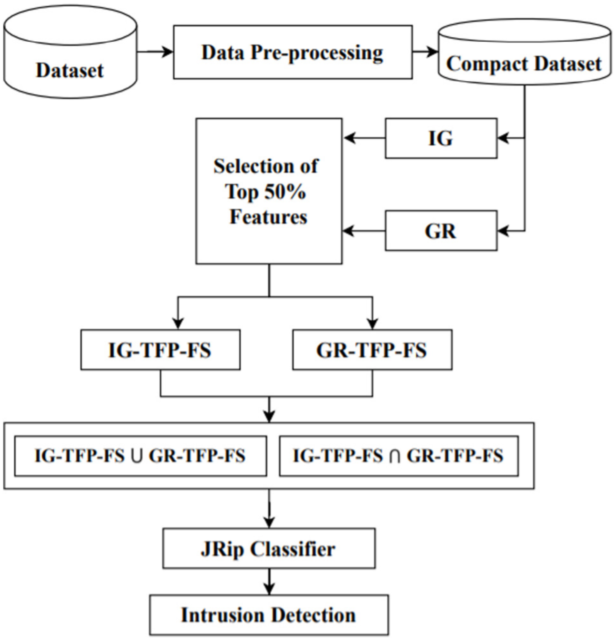

Fig. 1. DoS and DDoS Detection system in IoT.

3. Proposed system

The proposed intrusion detection system in IoT mainly consists of data pre-processing, feature selection, and rule based JRip classifier as shown in Fig. 1.

The captured network traffic is not suitable for machine learning models due to the presence of noise. It consists of NaN and missing values. It also consists of some of the features presenting the same information in text and numeric form. Therefore, initially, data pre-processing is performed on the noisy network traffic. The features presenting the same information in text form are removed, and NaN and missing values are replaced with zero. In this way, data pre-processing achieves the compact dataset for feature selection and intrusion detection.

The compact dataset obtained using data pre-processing is further used for feature selection and classification into normal and attack. Machine learning provides filter, wrapper, and embedded feature selection methods. The system uses IG and GR from the collection of filter-based feature selection techniques. The system selects the top 50% ranked features of the total number of features present in the compact dataset. The unique feature selection techniques obtain a subset of features, namely Information Gain-Top Fifty Percent-Feature Subset (IG-TFP-FS) and Gain Ratio-Top Fifty Percent-Feature Subset (GR-TFP-FS), by selecting a top, 50% ranked features. New Reduced Feature Subsets (RFS), namely RFS-1 and RFS-2, are obtained using intersection and union operations, respectively, on the features' obtained subsets. The new subsets of features, namely RFS-1, and RFS-2 are provided to JRip rule-based classifier to select a single feature subset that includes the minimum number of features. The JRip classifier with ten-fold cross-validation (CV) measures the system's performance. It selects a single subset of features based on improved accuracy (ACC), detection rate (DR), and model built-up time (B. Time) compared to the original feature set.

4. System implementation and result analysis

The proposed system described in Section 3 is implemented and tested using the Waikato Information Research Environment (Weka 3.8.3) on a 32 GB RAM workstation fitted with an Intel Xeon CPU E3-1271 v3 @ 3.60 GHz CPU. The Scikit-learn library in Python is used for pre-processing of the data. The proposed system is tested on latest [16] IoT-BoT dataset. The dataset consists of 43 features excluding labels. The dataset consists of DoS, DDoS, Keylogging, Data exfiltration, OS and Service Scan attacks. The proposed system detects DoS and DDoS attacks in IoT. Therefore the system uses 20% of the dataset for testing and includes UDP, TCP, and HTTP-based DoS and DDoS attacks. The derived dataset consists of a total number of 715 848 records, and includes 477 and 715 371 records of normal and attack respectively.

In data pre-processing, the features such as pkSeqID, Stime, saddr, daddr, and ltime that bypasses the system are removed manually. The features such as flgs, and Proto present in the dataset that have the same meaning in text form are also removed manually. The script written in Python is used to replace the missing and NaN values present in the benchmark dataset. Finally, the system uses the obtained compact dataset using data pre-processing consisting of 36 features. Further, the compact dataset is used for feature selection.

The study [17] shows that filter-based feature selection algorithms are faster than wrapper techniques. Filter-based feature selection algorithms present in the Weka tool are used, and empirical analysis is performed on IoT-BoT dataset. It shows that the top-ranked 50% features of IG and GR provide higher accuracy than other filter techniques. Therefore, the system selects IG and GR and obtains IG-TFP-FS and GR-TFP-FS subsets by choosing the top 50% features to the total number of 36 features present in the dataset. The subsets IG-TFP-FS and GR-TFP-FS consist of top-ranked 18 features, as shown in Table 1.

The system performs intersection and union operations on top 50% ranked features of IG and GR, i.e., IG-TFP-FS and GR-TFP-FS, and obtained new subsets of features, namely RFS-1 and RFS-2. The new feature subsets, namely RFS-1 and RFS-2, obtained using intersection and union operations on IG-TFP-FS and GR-TFP-FS consist of 16 and 20 features as shown in Table 1.

Table 1
Feature subsets obtained for DoS and DDoS detection.

Method	Feature numbers
IG-TFP-FS	4, 6, 10, 18, 20, 23, 24, 25, 26, 27, 28, 29, 30, 31, 33, 34, 35, 36
GR-TFP-FS	4, 9, 10, 18, 19, 20, 23, 24, 25, 26, 27, 28, 29, 30, 31, 33, 35, 36,
RFS-2	4, 6, 9, 10, 18, 19, 20, 23, 24, 25, 26, 27, 28, 29, 30, 31, 33, 34, 35, 36
RFS-1	4, 10, 18, 20, 23, 24, 25, 26, 27, 28, 29, 30, 31, 33, 35, 36

Table 2
Result analysis on IoT-BoT dataset with JRip.

Method	ACC(%)	DR(%)	ICI (%)	B. Time (s)
All F	99.9992	99.3711	0.0008	80.94
IG-TFP-FS	99.9993	99.5799	0.0007	36.34
GR-TFP-FS	99.9992	99.3711	0.0008	42.45
RFS-2	99.9992	99.3711	0.0008	40.87
RFS-1	99.9993	99.5799	0.0007	34.31

The empirical analysis of rule-based classifiers available in the Weka is performed on IoT-BoT dataset. It shows that JRip classifier provides higher accuracy and detection rate of 99.9992% and 99.9937% with 80.94 s model built up time compared to other rule-based classifiers. Therefore, The new feature subsets, namely RFS-1 and RFS-2, consist of 16 and 20 features, respectively; as shown in Table 1, the JRip classifier is provided to select a single subset for feature selection. The model's performance is calculated in terms of ACC, DR, and model built up time with ten-fold CV. The system selects a single feature subset from RFS-1 and RFS-2 based on improved accuracy (ACC), detection rate (DR), and model built-up time (B. Time) compared to 36 features present in the compact dataset, IG-TFP-FS, and GR-TFP-FS. Finally, the system selects the subset RFS-1 that consists of 16 features for DoS and DDoS detection.

The implemented system is tested on IoT-BoT dataset, and performance is measured with JRip. Table 2 shows the performance of the system with JRip using different feature subsets obtained during system implementation. Table 2 shows that the obtained feature subset RFS-1 built the model using JRip in 34.31 s, achieved higher ACC and DR of 99.9993%, and 99.5798% with 0.000004194 FAR using 16 features compared to GR-TFP-FS, RFS-2, and 36 features present in the compact dataset. It also shows that the system achieved the same ACC and DR with the minimum model built time using 16 features compared to IG-TFP-FS. It also provided lesser Incorrect Classification Instances (ICI) compared to compact dataset features.

5. System comparison and validation

A comparative analysis of the proposed system is performed with the traditional feature selection based network IDSs on IoT-BoT dataset. The feature selection approaches mentioned in traditional IDSs are applied on IoT-BoT dataset, and the performance is calculated with JRip using obtained

Table 3
Comparison with the traditional IDSs on IoT-BoT.

Study	ACC(%)	DR(%)	FAR	B. Time (s)
All F	99.9992	99.3711	0.000004194	80.94
[8]	99.9990	99.3697	0.000005591	27.56
[18]	99.9992	99.3711	0.000004194	52.08
RFS-1	99.9993	99.5798	0.000004194	34.31

Table 4
Proposed system validation on KDD Cup 1999 dataset.

Study	ACC(%)	DR(%)	B. Time (s)	FAR
All F	99.990	99.9909	16.5	0.000128
[8]	99.990	99.9932	7.61	0.000165
[18]	99.985	99.9943	9.67	0.000311
RFS-1	99.992	99.9943	9.34	0.000110

feature subset. A comparative analysis of the system with traditional network IDSs on IoT-BoT dataset with JRip is as shown in Table 3.

Table 3 shows that the proposed system achieved higher ACC and DR of 99.9993% and 99.5798% respectively, with JRip using obtained feature subset compared to [8,18], and 36 features present in the compact dataset. The system [8] uses the top 13 ranked IG features, and the system obtained reduced features using IG and CR. The proposed feature selection uses a combination of IG and GR with top-ranked 50% features provides higher performance in terms of ACC and DR compared to [8] and [18]. The system also provides a lesser False Alarm Rate (FAR) of 0.000004194 compared to [8].

The proposed system is also validated on KDD Cup 1999 dataset with JRip. The proposed system mainly detects DoS and DDoS attacks. Therefore, only DoS instances present in KDD Cup 1999 dataset are extracted for experimentation and validation. The dataset consists of a total number of 142 404 instances with 41 features. The dataset includes 87 832 and 54 572 of normal and DoS instances, respectively. The dataset consists of Smurf, land, pod, teardrop, Neptune, and back DoS attacks. The proposed system described in Section 3 obtains 19 features for the detection of DoS attacks on KDD Cup 1999 dataset. The comparative analysis of the proposed method with traditional IDSs has been performed on KDD Cup 1999 dataset with JRip, as shown in Table 4.

Table 4 shows that the proposed system achieved higher accuracy of 99.9920% with JRip using 19 features to detect DoS attacks on KDD Cup 1999 dataset compared to [8,18], and original features. It also shows that the feature selection method provides the same detection rate of 99.9943% using the minimum of 19 features and 9.34 s model built time compared to [18] and original features.

6. Conclusion

The presented system obtained 16 and 19 features using the intersection operation on the subsets obtained by the top 50% ranked features of IG and GR for BoT-IoT and KDD Cup 1999 datasets, respectively, for detecting DDoS and DoS attacks. The system achieved higher accuracy and detection rate of 99.9993%, and 99.5798% respectively, with JRip using 16 features on BoT-IoT dataset. The KDD Cup 1999 dataset's system validation also improved accuracy and detection rate of 99.9920% and 99.9943% to detect DoS attack using 19 features with JRip. This presented work will be extended to find optimal features for IDS using a combination of bio-inspired algorithms.

CRediT authorship contribution statement

Pushparaj Nimbalkar: Methodology, Software, Validation, Writing - original draft. **Deepak Kshirsagar:** Conceptualization, Supervision, Investigation, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Shaker Alanazi, Jalal Al-Muhtadi, Abdelouahid Derhab, Kashif Saleem, Afnan N. AlRomi, Hanan S. Alholaibah, Joel J.P.C. Rodrigues, On resilience of Wireless Mesh routing protocol against DoS attacks in IoT-based ambient assisted living applications, in: 2015 17th International Conference on E-Health Networking, Application & Services, HealthCom, IEEE, 2015, pp. 205–210.

[2] Yongsheng Zong, Guoyan Huang, A feature dimension reduction technology for predicting DDoS intrusion behavior in multimedia internet of things, Multimedia Tools Appl. (2019) 1–14.

[3] Lianbing Deng, Daming Li, Xiang Yao, David Cox, Haoxiang Wang, Mobile network intrusion detection for IoT system based on transfer learning algorithm, Cluster Comput. 22 (4) (2019) 9889–9904.

[4] Shengchu Zhao, Wei Li, Tanveer Zia, Albert Y. Zomaya, A dimension reduction model and classifier for anomaly-based intrusion detection in internet of things, in: 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, IEEE, 2017, pp. 836–843.

[5] Bayu Adhi Tama, Kyung-Hyune Rhee, An integration of PSO-based feature selection and random forest for anomaly detection in IoT network, in: MATEC Web of Conferences, Vol. 159, EDP Sciences, 2018, p. 01053.

[6] Muder Almiani, Alia AbuGhazleh, Amer Al-Rahayfeh, Saleh Atiewi, Abdul Razaque, Deep recurrent neural network for IoT intrusion detection system, Simul. Model. Pract. Theory 101 (2020) 102031.

[7] Nour Moustafa, Benjamin Turnbull, Kim-Kwang Raymond Choo, An ensemble intrusion detection technique based on proposed statistical flow features for protecting network traffic of internet of things, IEEE Internet Things J. 6 (3) (2018) 4815–4830.

[8] Vikash Kumar, Ayan Kumar Das, Ditipriya Sinha, UIDS: a unified intrusion detection system for IoT environment, Evol. Intell. (2019) 1–13.

[9] Nickolaos Koroniotis, Nour Moustafa, Elena Sitnikova, Jill Slay, Towards developing network forensic mechanism for botnet activities in the iot based on machine learning techniques, in: International Conference on Mobile Networks and Management, Springer, 2017, pp. 30–44.

[10] Monika Roopak, Gui Yun Tian, Jonathon Chambers, An intrusion detection system against ddos attacks in iot networks, in: 2020 10th Annual Computing and Communication Workshop and Conference, CCWC, IEEE, 2020, pp. 0562–0567.

[11] Abebe Diro, Naveen Chilamkurti, Leveraging LSTM networks for attack detection in fog-to-things communications, IEEE Commun. Mag. 56 (9) (2018) 124–130.

[12] Eirini Anthi, Lowri Williams, Małgorzata Słowińska, George Theodorakopoulos, Pete Burnap, A supervised intrusion detection system for smart home IoT devices, IEEE Internet Things J. 6 (5) (2019) 9042–9053.

[13] Yi-Wen Chen, Jang-Ping Sheu, Yung-Ching Kuo, Nguyen Van Cuong, Design and implementation of IoT DDoS attacks detection system based on machine learning, in: 2020 European Conference on Networks and Communications, EuCNC, IEEE, 2020, pp. 122–127.

[14] Andria Procopiou, Nikos Komninos, Christos Douligeris, ForChaos: Real time application DDoS detection using forecasting and chaos theory in smart home IoT network, Wirel. Commun. Mob. Comput. 2019 (2019).

[15] Rajesh Kumar Shrivastava, Bazila Bashir, Chittaranjan Hota, Attack detection and forensics using honeypot in IoT environment, in: International Conference on Distributed Computing and Internet Technology, Springer, 2019, pp. 402–409.

[16] Nickolaos Koroniotis, Nour Moustafa, Elena Sitnikova, Benjamin Turnbull, Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset, Future Gener. Comput. Syst. 100 (2019) 779–796.

[17] Veeran Ranganathan Balasaraswathi, Muthukumarasamy Sugumaran, Yasir Hamid, Feature selection techniques for intrusion detection using non-bio-inspired and bio-inspired optimization algorithms, J. Commun. Inf. Netw. 2 (4) (2017) 107–119.

[18] Ishfaq Manzoor, Neeraj Kumar, et al., A feature reduced intrusion detection system using ANN classifier, Expert Syst. Appl. 88 (2017) 249–257.