International Journal of Engineering & Technology, 7 (3.24) (2018) 479-482

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A detailed analysis of CICIDS2017 dataset for designing Intrusion Detection Systems

Ranjit Panigrahi[1]*, Samarjeet Borah[2]

[1]Department of Computer Applications, SMIT, Sikkim Manipal University, Majitar, India
[2] Department of Computer Applications, SMIT, Sikkim Manipal University, Majitar, India
**Corresponding author E-mail:ranjit.panigrahi@gmail.com*

Abstract

Many Intrusion Detection Systems (IDS) has been proposed in the current decade. To evaluate the effectiveness of the IDS Canadian Institute of Cybersecurity presented a state of art dataset named CICIDS2017, consisting of latest threats and features. The dataset draws attention of many researchers as it represents threats which were not addressed by the older datasets. While undertaking an experimental research on CICIDS2017, it has been found that the dataset has few major shortcomings. These issues are sufficient enough to biased the detection engine of any typical IDS. This paper explores the detailed characteristics of CICIDS2017 dataset and outlines issues inherent to it.Finally, it also presents a combined dataset by eliminating such issues for better classification and detection of any future intrusion detection engine.

Keywords: *CICIDS2017, Intrusion Detection Systems, IDS, Class Imbalance Problem, Recent Dataset for IDS*

1. Introduction

The current advances in the field of information and communication technology worldwide, poses a great challenge for network engineers. A big challengefor today's network engineers and researchers is the identification of malevolent activities in a host, which eventually propagate to the other hosts over a network. An untrusty program which forcefully take part in this event of disaster and is popularly called intrusion. The factual meaning of intrusion is illegal access to the system, or the network resources[1,2,3,4,5]. The Intrusion Detection Systems (IDS) play a vital role in minimizing such kind of activities. Most of the IDSs are either follow anomaly detection or misuse detection mechanism. Misuse detection mechanism are quite popular in industries for designing effective commercial IDS whereas anomaly detections are limited to academics for research and developments[6]. But overall an IDS needs existing information to detect future attacks. That is the reason IDSs used to be trained on an effective dataset.

Many intrusion detection models have already been proposed by researchers claiming an accuracy of 98%+ with very limited false alarm below 1%. This high rate of accuracy attracted researchers and industry to invest money and effort to deliver effective products for the users. But, only few models are actually accepted by the industries to develop a real-world IDS. To find the reason regarding this, we minutely observe recent IDS models, training and testing datasets and algorithms used to generate samples from such datasets. During our research, we found that a very recent dataset named CICIDS2017[7] provided by Canadian Institute of Cybersecurity contains most update attack scenarios. This state of art dataset not only contains uptodate network attacks but also fulfils all the criteria of real-world attacks. While exploring the characteristics of this dataset we found few shortcomings. One of the shortcomings and quite visible is that the dataset is huge and spanned over eight files as five days traffic information of Canadian Institute of Cybersecurity. A single dataset would be feasible for designing an IDS. Further, the dataset contains many redundant records which seems to be irreverent for training any IDS. Though, the dataset contains recent attack scenario but at the same time we also found the dataset is high class imbalance[8,9] in nature. A class imbalance dataset may mislead the classifier and biased towards the majority class[10,11]. We also tried to resolve these shortcomings and presented a subset of CICIDS2017 dataset to the research community for designing and testing of their detection models.

The contribution of this paper is –

(a) Identifies and provides effective solutions to the shortcoming of CICIDS2017 dataset.

(b) Relabeling CICIDS2017 dataset to reduce high class imbalance problem.

The rest of this paper is asfollow. Section 2 shows the detailed description and characteristics of the dataset of subject concern; Section 3 outlines shortcomings and Section 4 provides solution to those shortcomings of CICIDS2017 dataset followed by conclusion at Section 5.

2. Descriptions of CICIDS2017 dataset

Since the inception of CICIDS2017 dataset the dataset started attracting researchers for analysis and developments of new models and algorithms[12,13,14]. According to the author[7] of CICIDS2017, the dataset spanned over eight different files containing five days normal and attacks traffic data of Canadian Institute of Cybersecurity. A short description of all those files are presented in Table 1.

Table 1: Description of files containing CICIDS2017 dataset

Name of Files	Day Activity	Attacks Found
Monday-WorkingHours.pcap_ISCX.csv	Monday	Benign (Normal human activities)
Tuesday-WorkingHours.pcap_ISCX.csv	Tuesday	Benign, FTP-Patator, SSH-Patator
Wednesday-workingHours.pcap_ISCX.csv	Wednesday	Benign, DoS GoldenEye, DoS Hulk, DoS Slowhttptest, DoS slowloris, Heartbleed
Thursday-WorkingHours-Morning-WebAttacks.pcap_ISCX.csv	Thursday	Benign, Web Attack – Brute Force, Web Attack – Sql Injection, Web Attack – XSS
Thursday-WorkingHours-Afternoon-Infilteration.pcap_ISCX.csv	Thursday	Benign, Infiltration
Friday-WorkingHours-Morning.pcap_ISCX.csv	Friday	Benign, Bot
Friday-WorkingHours-Afternoon-PortScan.pcap_ISCX.csv	Friday	Benign, PortScan
Friday-WorkingHours-Afternoon-DDos.pcap_ISCX.csv	Friday	Benign, DDoS

It can be seen from Table 1 that the dataset contains attack information as five days traffic data. Thursday working hour afternoon and Friday data are well suited for binary classification. Similarly, Tuesday, Wednesday and Thursday morning data are best for designing multiclass detection model. However, it should be noted that a best detection model should be able to detect attacks of any type. Therefore, to design such a typical IDS, the traffic data of all the day should be merged together to form a single dataset to be used by IDS. This is exactly we followed to merge these files.

By merging the files presented in Table 1, we found the whole shape of a dataset that contains 3119345 instances and 83 features containing 15 class labels (1 normal + 14 attack labels). Further, examining the instances of the combined files, it has ben found that the dataset contains 288602 instances having missing class label and 203 instances having missing information. By removing such missing instances, we found a combined dataset of CICIDS2017 having 2830540 instances. At this moment we queried for any possible redundant instances. Surprisingly, no redundant instances were found. The characteristics of combined dataset and the detailed class wise occurrence has been presented in Table 2 and Table 3 respectively.

Table 2: Overall characteristics of CICIDS2017 dataset

Dataset Name	CICIDS2018
Dataset Type	Multi class
Year of release	2017
Total number of distinct instances	2830540
Number of features	83
Number of distinct classes	15

Table 3: Class wise instance occurrence of CICIDs2017 dataset

Class Labels	Number of instances
BENIGN	2359087
DoS Hulk	231072
PortScan	158930
DDoS	41835
DoS GoldenEye	10293
FTP-Patator	7938
SSH-Patator	5897
DoS slowloris	5796
DoS Slowhttptest	5499
Bot	1966
Web Attack – Brute Force	1507
Web Attack – XSS	652
Infiltration	36
Web Attack – Sql Injection	21
Heartbleed	11

Another interesting point that we observed that the dataset fulfills all the criteria[7,15] of a true intrusion detection dataset such as complete network configuration, complete traffic, labelled dataset, complete interaction, complete capture, available protocols, attack diversity, heterogeneity, feature set and meta data.

3. Shortcomings of CICIDS2017

We observed earlier that CICIDS2017 dataset contains few shortcomings and the aim of this paper is to address those shortcomings for better understanding of the future researchers –

4.1. Scattered Presence

We have seen in table 1 that the data of CICIDS2017 dataset is present scatter across eight files. Processing individual files are a tedious task. Therefore, we combined those files to form a single file that contains a total of 3119345 instances of all the files.

4.2. Huge Volume of Data

After combining all the files, we noticed that the combined dataset contains data of all the possible recent attack labels at one place. But, at the same time the size of a combined dataset becomes huge. This huge volume of data itself becomes a shortcoming. The shortcoming is that it consumes more overhead for loading and processing.

4.3. Missing Values

We also observed that the combined CICIDS2017 dataset contains 288602 instances having missing class label and 203 instances having missing information. These unwanted instances have been removed to form a dataset that contains unique 2830540 instances.

4.4. High class imbalance

High class imbalance[8,9,16] is a situation in a dataset where if the dataset is used for training of a classifier or detector, in such a case the detector biased towards the majority class[9,10]. As a result, the detector shows lower accuracy with higher false alarm. In the case of CICIDS2017, the dataset is also prone to high class imbalance. The class imbalance ratio has been presented in Table 4 and its impact can be visualize in Figure 1.

Table 4: Class prevalence ratio of CICIDS2017 dataset

Sl No	Normal / Attack Labels	Number of instances	% of prevalence w.r.t. the majority class	% of prevalence w.r.t. the total instances
1	BENIGN	2359087	1	83.34406
2	Bot	1966	0.000833	0.06946
3	DDoS	41835	0.017734	1.47799
4	DoS GoldenEye	10293	0.004363	0.36364
5	DoS Hulk	231072	0.09795	8.16353
6	DoS Slowhttptest	5499	0.002331	0.19427
7	DoS slowloris	5796	0.002457	0.20477
8	FTP-Patator	7938	0.003365	0.28044
9	Heartbleed	11	0.000005	0.00039
10	Infiltration	36	0.000015	0.00127
11	PortScan	158930	0.067369	5.61483
12	SSH-Patator	5897	0.0025	0.20833
13	Web Attack – Brute Force	1507	0.000639	0.05324
14	Web Attack – Sql Injection	21	0.000009	0.00074

15	Web Attack – XSS	652	0.000276	0.02303

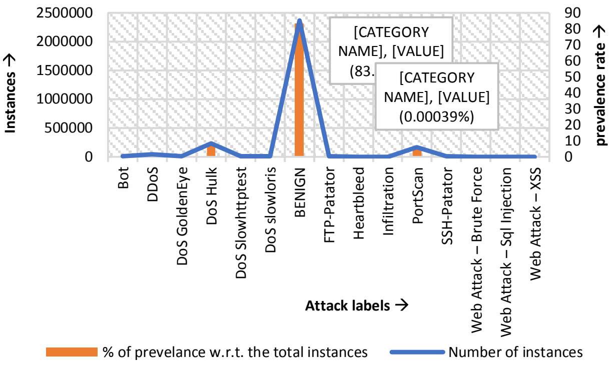

Fig. 1: Graphical representation of majority and minor class prevalence of CICIDS2017 dataset

It can be seen from the table that the prevalence of majority class (Benign) is 83.34% where as for the minority class is 0.00039% (Heartbleed). In such a huge difference of prevalence rate a potential detector may tilt towards benign. The situation becomes worst when a detector is based on a sample of this dataset. It is because, when a random sample on this dataset is ascertained for training and testing of a detector there is a great possibility that instances of a particular attack label such as “Heartbleed” or “Web Attack – Sql Injection” may not found in training set. As a result, the detector will fail to detect such attack when an instance of type such attack arrives. This is a major drawback that we have noticed in CICIDS2017 dataset.

4. Our Solutions

In this section we handled the shortcomings presented in Section 4. The problem of scattered presence has been handled by combining various data files if CICIDS2017. Further, the missing values has also been removed. Though, huge volume of data is a shortcoming for a dataset but at the same time it is inherent to any typical dataset that contains typical information. This shortcoming of high volume can be overcome by sampling the dataset before actual detection process starts. However, it is strongly advised that before sampling the class imbalance problem must be addressed before hand. If the dataset will be balanced the probability of occurrence of instances of all class label will be increased.

There are many ways to handle class imbalance problem of a dataset[8,11,17,18]. One of the major ways to resolve class imbalance issue is relabeling of classes. Relabeling of classes includes splitting of majority classes to form more classes or merging of few minority classes to form a class; thus; improving prevalence ratio and reducing class imbalance issue. In the case of CICIDS2017 it is very difficulty to split majority classes to form discrete classes equivalent to minority classes. It is because, the difference in prevalence is 83.34367%, which seems to be too high. Therefore, we have decided to merge few minority classes to form new attack classes.

To form new classes, we merge few minority attack classes having similar characteristics and behavior. The new label information can be found at [19,20]. After merging similar classes, the class the prevalence ratio of various attack labels seems to be improved. The dataset characteristics after merging of similar data labels has been presented in Table 5.

Table 5: Characteristics of new attack labels with their prevalence rate in CICIDS2017 dataset

Sl No	***New Labels***	***Old Labels***	***Number of instances***	***% of prevalence w.r.t. the majority class***	***% of prevalence w.r.t. the total instances***
1	Normal	Benign	2359087	100	83.34
2	Botnet ARES	Bot	1966	0.083	0.06
3	Brute Force	FTP-Patator, SSH-Patator	13835	0.59	0.48
4	Dos/DDos	DDoS, DoS GoldenEye, DoS Hulk, DoS Slowhttptest, DoS slowloris, Heartbleed	294506	12.49	10.4
5	Infiltration	Infiltration	36	0.001	0.001
6	PortScan	PortScan	158930	6.74	5.61
7	Web Attack	Web Attack – Brute Force, Web Attack – Sql Injection, Web Attack – XSS	2180	0.092	0.07

It can be seen from Table 5 is that, the new labels of the dataset improve prevalence of all attack labels significantly; thus, reducing the class imbalance rate.

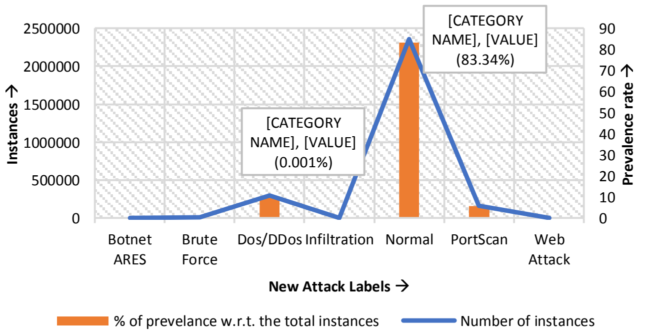

Fig. 2: Graphical representation of majority and minor class prevalence of CICIDS2017 dataset with respect to new attack labels

The same reduction of class imbalance rate can also be visualized in figure 2. The imbalance ratio of the minor class has been improved from 0.00039% to 0.001%.

5. Conclusion

In this research article we have considered a most recent dataset CICIDS2017 for detailed analysis keeping in view its increasing demand in the research community. Various shortcomings of the dataset have been studied and outlined. Solutions to counter such issues has also been provided. We tried to solve such issues through experiment. We also relabel the dataset with the labelling information provided by Canadian Institute of Cybersecurity. Moreover, we have also seen a major issue of class imbalance has been reduced by such class relabeling.

As a future work the dataset can be class wise resampled to generate two or more training and testing samples set separately to be used by research community.

References

[1] Mehmed Kantardzic; Jozef Zurada, "Using Data Mining for Intrusion Detection," in Next Generation of Data-Mining Applications , , IEEE, 2005, pp. doi: 10.1109/9780471696650.ch22.

[2] R. A. Kemmerer and G. Vigna, “Intrusion Detection: A Brief History and Overview,” Computer Society, Vol. 35, No. 4, 2002, doi: 10.1109/MC.2005

[3] Christos Douligeris; Dimitrios N. Serpanos, "Intrusion Detection Versus Intrusion Protection," in Network Security: Current Status and Future Directions, IEEE, 2007, pp. doi: 10.1002/9780470099742.ch7

[4] Seppo J. Ovaska, "Intrusion Detection for Computer Security," in Computationally Intelligent Hybrid Systems: The Fusion of Soft Computing and Hard Computing, IEEE, 2005, pp. doi: 10.1002/9780471683407.ch8

[5] Farooq Anjum; Petros Mouchtaris, "Intrusion Detection Systems," in Security for Wireless Ad Hoc Networks, Wiley, 2007, pp. doi: 10.1002/9780470118474.ch5

[6] M. Tavallaee, E. Bagheri, W. Lu and A. A. Ghorbani, "A de-tailed analysis of the KDD CUP 99 data set," 2009 IEEE Sym-posium on Computational Intelligence for Security and De-fense Applications, Ottawa, ON, 2009, pp. 1-6. doi: 10.1109/CISDA.2009.5356528

[7] Iman Sharafaldin, Arash Habibi Lashkari, and Ali A. Ghorba-ni, “Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization”, 4th International Confer-ence on Information Systems Security and Privacy (ICISSP), Purtogal, January 2018

[8] C. Mera and J. William Branch, "A Survey on Class Imbalance Learning on Automatic Visual Inspection," in IEEE Latin America Transactions, vol. 12, no. 4, pp. 657-667, June 2014. doi: 10.1109/TLA.2014.6868867.

[9] S. Wang, L. L. Minku and X. Yao, "A Systematic Study of Online Class Imbalance Learning with Concept Drift," in IEEE Transactions on Neural Networks and Learning Systems. doi: 10.1109/TNNLS.2017.2771290.

[10] Q. Song, Y. Guo and M. Shepperd, "A Comprehensive Investigation of the Role of Imbalanced Learning for Software De-fect Prediction," in IEEE Transactions on Software Engineer-ing. doi: 10.1109/TSE.2018.2836442.

[11] S. Wang and X. Yao, "Multiclass Imbalance Problems: Analy-sis and Potential Solutions," in IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 42, no. 4, pp. 1119-1130, Aug. 2012. doi: 10.1109/TSMCB.2012.2187280

[12] Benjamin J. Radford, Bartley D. Richardson, Shawn E. Davis. Sequence Aggregation Rules for Anomaly Detection in Com-puter Network Traffic. American Statistical Association 2018 Symposium on Data Science and Statistics, May 2018, pp. 1-13.

[13] R. Vijayanand, D. Devaraj, B. Kannapiran, Intrusion detection system for wireless mesh network using multiple support vec-tor machine classifiers with genetic-algorithm-based feature selection, Computers & Security, Volume 77, 2018, Pages 304-314, ISSN 0167-4048, https://doi.org/10.1016/j.cose.2018.04.010.

[14] L.Nicholas, S.Y. Ooi, Y.H. Pang, S.O. Hwang, S.Tan, "Study of long short-term memory in flow-based network intrusion detection system", Journal of Intelligent & Fuzzy Systems, vol. Pre-press, no. Pre-press, pp. 1-11, 2018, doi: 10.3233/JIFS-169836

[15] A. Gharib, I. Sharafaldin, A. H. Lashkari and A. A. Ghorbani, "An Evaluation Framework for Intrusion Detection Dataset," 2016 International Conference on Information Science and Security (ICISS), IEEE Thailand, 2016, pp. 1-6.

[16] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince and F. Herrera, "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approach-es," in IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 42, no. 4, pp. 463-484, July 2012. doi: 10.1109/TSMCC.2011.2161285

[17] R. Longadge,S.S. Dongre,L. Malik, "Class Imbalance Problem in Data Mining: Review", International Journal of Computer Science and Network (IJCSN), Volume 2, Issue 1, February 2013, ISSN 2277-5420

[18] S.M.A. Elrahman and A. Abraham, A Review of Class Imbal-ance Problem, Journal of Network and Innovative Computing, ISSN 2160-2174, Volume 1 (2013) pp. 332-340

[19] Iman Sharafaldin, Arash Habibi Lashkari, and Ali A. Ghorba-ni, Intrusion Detection Evaluation Dataset (CICIDS2017), Ca-nadian Institute of Cybersecurity, http://www.unb.ca/cic/datasets/ids-2017.html, Accessed on: 13/04/2018

[20] Nallapaneni Manoj Kumar, Pradeep Kumar Mallick, “Blockchain technology for security issues and challenges in IoT”, Elsevier Procedia Computer Science Journal, Volume 132, Pages 1815-1823 , 2018, ISSN:1877-0509, UGC Sl No: 46138 and 48229, DOI: https://doi.org/10.1016/j.procs.2018.05.140.