Communications on Applied Electronics (CAE) – ISSN : 2394-4714
Foundation of Computer Science FCS, New York, USA
Volume 7– No. 32, December 2019 – www.caeaccess.org

Data Deduplication: Its Significant Effect on Network Intrusion Dataset

Aladesote O. Isaiah
Computer Science Department Federal
Polytechnic, Ile Oluji
Ondo State, Nigeria

Adetunji A. Ademola
Department of Statistics
Federal Polytechnic, Ile Oluji
Ondo State, Nigeria

ABSTRACT

This research work adopted future extraction techniques on NSL KDD data set, using deduplication software written in C++ Programming Language, duplicated records of four attack types (DOS, R2L, Robing and U2R) were removed**.** Among the attack types for DOS, Mailbomb with 98.63% has highest percentage reduction rate while Apache2 with 40.30% reduction rate has the least. For R2L, Smpgetattack with 92.70% reduction has the highest while there was no reduction for Ftp_write. With 93.15% reduction, Nmap has the highest reduction rate under Probing attack while Mscan with 60.84% reduction rate has the least while 50% reduction rate for Sqlattack is the highest for U2R attack type. Wilcoxon Sign test is used to test for the significance of the deduplication and results revealed that all the attack types except U2R have significant reduction rate at 5% level.

Keywords

Deduplication, extraction techniques, attack types, Wilcoxon sign test, NSL-KDD

1. INTRODUCTION

The recent and latest development in Information Technology has posed a very serious concern to Network security (Amudha, et al., 2015). The needs to keep track of all existing network security threats, coupled with just emerging new ones seem challenging. The technique adopted to improve network technologies has paved a way for invaders or hackers to devise illegal means of gaining access into a network system [1]. Hence, an active and appropriate intrusion system is needed, to detect attack(s) when noticed [6].

Intrusion Detection System is the art of determining malicious action(s) that attempt to compromise the confidentiality, integrity or availability of a network [9] and also a security approach used to protect computer networks from unauthorised access [8]. The Wilcoxon Sign test is a statistical tool for comparing two paired samples. It is the non-parametric alternative to paired sample t-test when various assumptions underlying utilizing parametric techniques (like the normality assumption) are weak. It tests the null hypothesis that the average signed rank of two dependent samples is zero.

Effective feature extraction from intrusion detection datasets is one of the important research challenges for constructing high performance IDS. Irrelevant and redundant attributes of intrusion detection dataset may lead to complex intrusion detection model as well as reduce detection accuracy [5].

Data deduplication helps to remove duplicate records, thereby leaving a copy of each record in a set of data; this leads to the reduction in the amount of data to be moved into the network [4]. A Novel Approach for Record Deduplication using Hidden Markov Model (HMM) based record deduplication in order to overcome the accuracy level of the existing approach in this research, two real datasets (Bibliographic and Restaurants datasets) gathered from the web were used. The authors analysed and compared Hidden Markov Model and Genetic programming based record deduplication [3]. The result revealed that HMM performed better than Genetic Programming in terms of accuracy, processing and recall.

An iterative Approach to Record Dedupication proposing Particle Swam Optimization (PSO) and Bat Algorithm to web datasets and documents in order to overcome the weaken of Genetic Programming approach which is an existing system to record deduplication [7]. The result revealed that the approaches were to find the best optimization solution for random selection of the input values and removing the duplicate records in the system. The researchers stated that new algorithm should be introduced to find the duplication record, which will reduce the time to find such repeated records without affecting the speed of the process.

2. OBJECTIVES

The objectives of the research are to develop a deduplication system and also determine the significant effect of deduplication on NSL KDD dataset

3. METHODOLOGY

The review of related work was extensively carried out. The result of feature extraction techniques on NSL-KDD dataset was adopted which reduced 41 features to 13 [1]. The records of these 13 attributes were extracted and categorised into each of the network attack types, excluding normal traffic records. A deduplication software was developed with C++ Programming Language to remove the duplicated records from each of the four network attack types. The Wilcoxon Sign test tool was used for comparing the two paired samples. The dataset for comparison is found not to be normally distributed, hence a non- parametric (Wilcoxon Sign test) test of significant difference is used to determine the significant effect of record deduplication on each network attack type.

4. RESULT AND DISCUSSION

This shows the result obtained when each category of network attacks was run on the data deduplication software

4.1 Data Deduplication on Denial of Service (DoS) Attack type

The results obtained when the Denial of Service (DoS) dataset was run on the software revealed that 737 records of Apache2 were reduced to 440, which is 40.3% reduction. 357 records of Back were reduced to 65, which is 81.79% reduction. 7 records of Land were reduced to 3, resulting in 57.1%. 293 records of Mail bomb were reduced to 4, which amount to 98.63% reduction. 4657 records of Neptune were reduced to 295, which is 93.67% reduction. 41 records of Ping of Death (PoD) were reduced to 14, which equates to 65.85% reduction. 685 records of Processtable were reduced to 367, which is 46.42% reduction. 665 records of smurf were reduced to 10, which is equivalent to 98.50%, 12 records of Teardrop were reduced to 2, which is 83.33% reduction, 2 records of updstorm were reduced to 1, which is 50% reduction while 944 records of warezmaster were reduced to 180 records, which is 80.93% reduction.

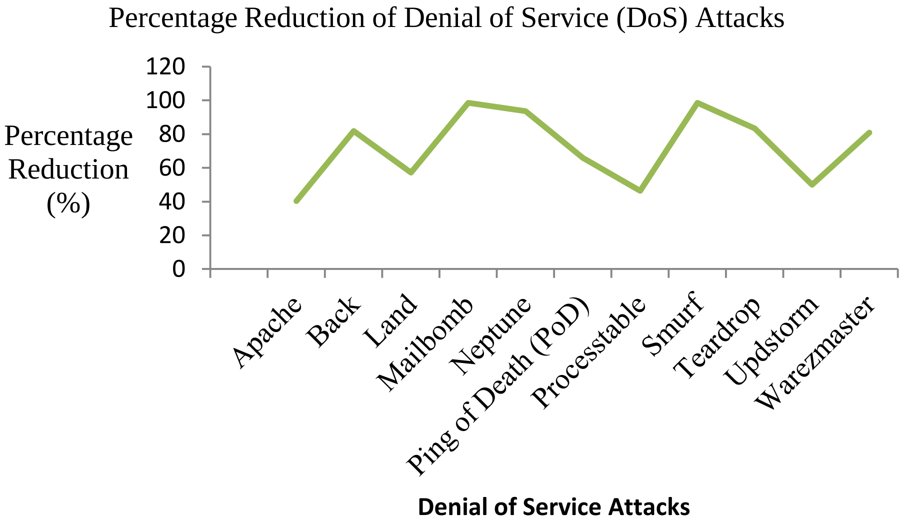

Fig 1: Graphical representation of percentage reduction of Denial of Service (DoS) on deduplication software

4.1.1 *Descriptive Statistics for Denial of Service Attack*

The statistics of Denial of service attack before deduplication and after deduplication was determined. The mean value before deduplication and after deduplication is 763.82 and 125.55 respectively while the standard deviation before deduplication and after deduplication is 1335.381 and 166.999 respectively. It can be deduced that the Mean and standard deviation after deduplication is greatly reduced when compared with the result before deduplication.

Table 1: Descriptive Statistics of Denial of Service Attack

	N	**Mean**	**Standard Deviation**
Before Deduplication	11	763.82	1335.381
After Deduplication	11	125.55	166.999

Where N is the number of attack category on Denial of Service (DoS) attack group.

4.1.2 *Test of Significant effect on Denial of Service (DoS) Attack*

The Wilcoxon signed rank test in Table 2 shows that the observed difference between both measurements is significant. Thus we can reject the null hypothesis that both samples are from the same population, and conclude that Deduplication using DOS has significant effect.

Table 2: Wilcoxon's Test for DOS

	After Deduplication - Before Deduplication	Spearman's rho
Z	-2.934	0.873
P-value	0.003	0.000

4.2 Data Deduplication on Remote to Local (R2L) Attack type

The result obtained when the Remote to Local (R2L) dataset was run on the software shows that 1231 records of guesspasswd were reduced to 178, which is 85.54% reduction. 133 records of Httptunnel were reduced to 20, which is 84.96% reduction. 17 records of Nmaed were reduced to 15, resulting in 11.76% reduction, 14 records of sendmail were reduced to 11, which amount to 21.43% reduction. 178 records of snmpgetattack were reduced to 13, which is 92.70%. Records of ftp_write, imap, worm, xlock and xsnoop remain unchanged before and after deduplication, which is 0% reduction,

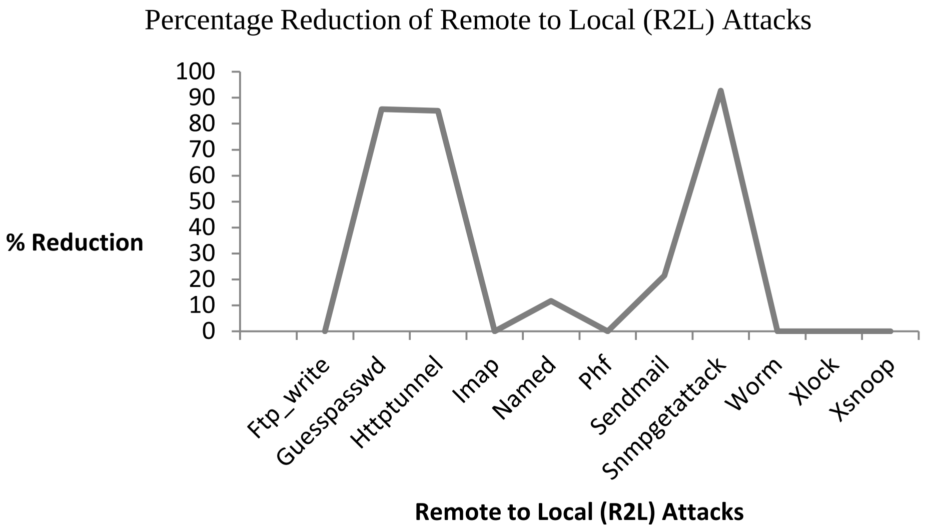

Fig 2: Graphical representation of percentage reduction of Remote to Local (R2L) on deduplication software

4.2.1 Descriptive Statistics for Remote to Local (R2L) attack

The statistics of Remote to Local (R2L) attack before deduplication and after deduplication was determined. The Mean value before deduplication and after deduplication is 144.91 and 23.45 respectively while the standard deviation before deduplication and after deduplication is 365.287 and 51.636 respectively. It can be deduced that the Mean and standard deviation after deduplication is greatly reduced when compared with the result before deduplication.

Table 3: Descriptive Statistics of Remote to Local (R2L) attack

	N	Mean	Standard Deviation
Before Deduplication	11	144.91	365.287
After Deduplication	11	23.45	51.636

Where N is the number of attack category on Denial of Service (DoS) attack group.

4.2.2 Test of Significant effect on Remote to Local (R2L) Attack

The Wilcoxon signed rank test in Table 4 shows that the observed difference between both measurements is significant. Thus we can reject the null hypothesis that both samples are from the same population, and conclude that Deduplication using R2L has significant effect

Table 4: Wilcoxon's Test for Remote to Local (R2L)

	After Deduplication - Before Deduplication	Spearman's rho
Z	-2.023	0.973
P-value	0.043	0.000

4.3 Data Deduplication on Probing Attack type

The result obtained when the Probing dataset was run on the software reveals that 73 records of Nmap were reduced to 5, which is 93.15% reduction. 157 records of Portsweep were reduced to 24, which is 84.71% reduction. 141 records of ipsweep were reduced to 13, which amount to 90.78% reduction. 996 records of Mscan were reduced to 390, which is 60.84%. 319 records of Saint were reduced to 70, which is 78.06% while 375 records of Satan were reduced to 78, which is 89.39%.

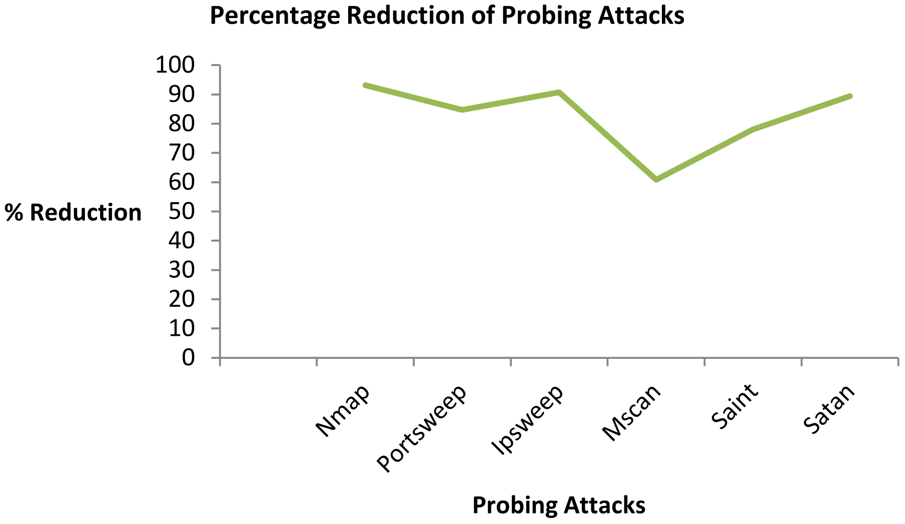

Fig 3: Graphical representation of Percentage Reduction of Probing attacks on deduplication software

4.3.1 Descriptive Statistics for Probing Attack

The statistics of Probing attack before deduplication and after deduplication was determined. The result reveals that the Mean value before deduplication and after deduplication is 403.50 and 96.67 respectively while the standard deviation before deduplication and after deduplication is 376.029 and 146.835 respectively. It can be deduced that the Mean and standard deviation after deduplication is greatly reduced when compared with the result before deduplication.

Table 5: Descriptive Statistics of Probing attack

	N	**Mean**	**Standard Deviation**
Before Deduplication	6	403.50	376.029
After Deduplication	6	96.67	146.825

Where N is the number of attack category on Probing attack group.

4.3.2 Test of Significant Probing Attack

The Wilcoxon signed rank test shows that the observed difference between both measurements is significant. Thus we can reject the null hypothesis that both samples are from the same population, and conclude that Deduplication using PROBING has significant effect.

Table 6: Wilcoxon's Test for Probing

	After Deduplication - Before Deduplication	**Spearman's rho**
Z	-2.201	0.999
P-value	0.028	0.000

4.4 Data Deduplication on User to Root (U2R) Attack

The result obtained when the User to Root (U2R) dataset was run on the software. 13 records of Xterm were reduced to 12, which is 7.69% reduction. 2 records of sglattack were reduced to 1, which is 50% reduction. Records of perl, rootkit, loadmodule, Ps, buffer_overflow remain unchanged before and after deduplication, which is 0% reduction while Eject and Fdformat have no record.

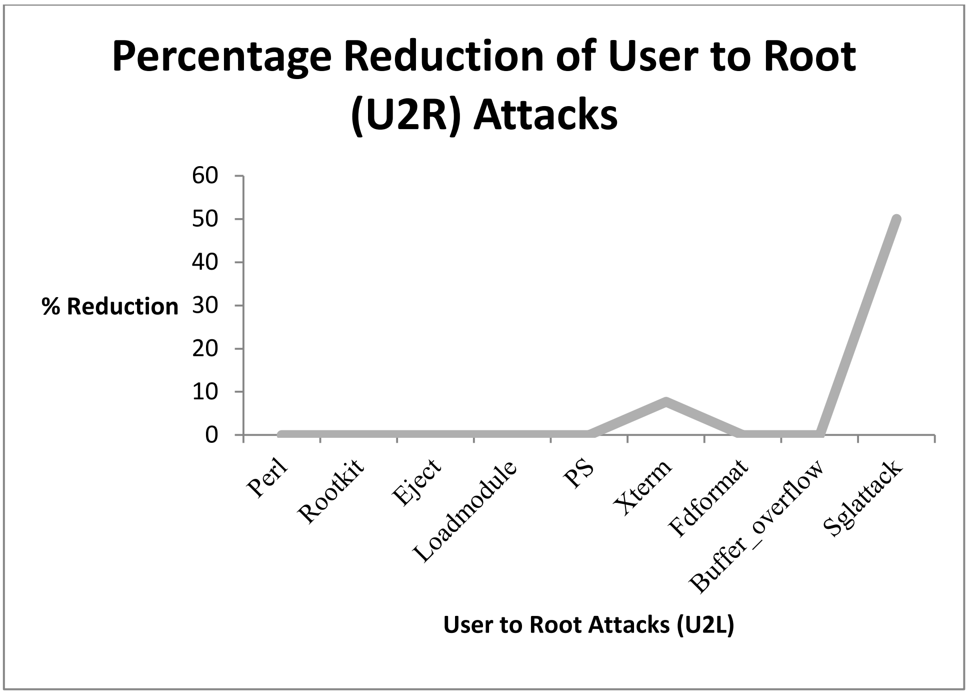

Fig 4: Graphical representation of percentage reduction of User to Root (U2R) on deduplication software

4.4.1 Descriptive Statistics for User to Root (U2R) attack

The statistics of User to Root (U2R) attack before deduplication and after deduplication was determined. The Mean value before deduplication and after deduplication is 39.80 and 9.40 respectively while the standard deviation before deduplication and after deduplication is 102.5769 and 10.024 respectively. It can be deduced that the Mean and standard deviation after deduplication is greatly reduced when compared with the result before deduplication

Table 7: Descriptive Statistics of User to Root (U2R) attack

	N	Mean	Standard Deviation
Before Deduplication	10	39.80	102.576
After Deduplication	10	9.40	10.024

Where N is the number of attack category on User to Root (U2R) attack group.

4.4.2 Test of Significant effect on Denial of Service (DoS) Attack

The Wilcoxon signed rank test in Table 8 shows that the observed difference between both measurements is not significant. Thus we do not reject the null hypothesis that both samples are from the same population, and conclude that Deduplication using U2R has does not have significant effect

Table 8: Wilcoxon's Test for User to Root (U2R)

	After Deduplication - Before Deduplication	Spearman's rho
Z	-1.633	0.988
P-value	0.102	0.000

5. CONCLUSION

Most previous work on intrusion detection system using intrusion datasets only considered extracting significant features or attributes of the datasets, not minding the significant effect of reducing such. This research work made use of the Spearman's Rank Correlation coefficient to determine the significant positive relationship among various techniques used under the four attack types (DOS, R2L, U2R and Probing) of the NSL-KDD dataset. It is seen and concluded that all the attack types except U2R have significant reduction at 5% level.

6. REFERENCES

[1] Aladesote O., Alese, K. & Dahunsi F. 2014. Intrusion Detection System using Hypothesis Testing. Proceedings of the World Congress on Engineering and Computer Science (WCECS) vol. I, 22-24.

[2] Amudha, P., Karthik & Sivakumari 2015. A Hybrid Swarm Intelligence Algorithm for Intrusion Detection Using Significant Feature. The Scientific World Journal, vol. 2015.

[3] Devi, R. & Thigarasu, V. 2014. A Novel Approach for Record Deduplication using Hidden Markov Model (HMM). International Journal of Computer Science and Information Technologies. 5(6), 8070 – 8073.

[4] Dirk M. 2013. Advanced Data Deduplication Technique

and their Application. Dissertation Submitted at the Department of Mathematics & Informatics, Johannes Gutenberg University Mainz.

[5] Farid, Daramont, Harbi, et al., 2009. Adaptive Network Intrusion Detection Learning: Attribute Selection & Classification. International Journal of Computer and Information Engineering 3(12), 2009.

[6] Jaiganesh, V., Sumathi, D. & Mangayarkarasi, S. 2013. An Analysis of Intrusion Detection System using Back Propagation Neural Network. IEEE Computer Society Publication 2013.

[7] Jiang, Y., Lin, C., Meng, W. et al, 2014. Rule-based deduplication of article records from bibliographic databases. Database. Vol. 2014.

[8] Prajowal, M. 2014. A Practical Approach to Anomaly based Intrusion Detection System by Outlier Mining in Network Traffic. A Thesis Presented to the Masdar Institute of Science and Technology in Partial Fulfilment of the Requirements for the Degree of Master of Science in Computing and Information Science.

[9] Shona D. & Senthilkumar, 2016. An Ensemble Data Preprocessing Approach for Intrusion Detection System using Variant Firefly & BK-NN Techniques. International Journal of Applied Engineering Research, 11(6), 4161 – 4166.