Performances of Machine Learning Algorithms for Binary Classification of Network Anomaly Detection System

Mukrimah Nawir, Amiza Amir, Ong Bi Lynn, Naimah Yaakob, R.Badlishah Ahmad
1Embedded, Network, and Advanced Computing Research Cluster (ENAC)
School of Computer and Communication Engineering, Universiti Malaysia Perlis (UniMAP), Pauh, Perlis
2Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin (UniSZA), 22200, Besut, Terengganu

E-mail: mukrimah1992@gmail.com; amizaamir@unimap.edu.my; bilynn@unimap.edu.my; naimahyaakob@unimap.edu.my; badli@unisza.edu.my

Abstract. The rapid growth of technologies might endanger them to various network attacks due to the nature of data which are frequently exchange their data through Internet and large-scale data that need to be handle. Moreover, network anomaly detection using machine learning faced difficulty when dealing the involvement of dataset where the number of labelled network dataset is very few in public and this caused many researchers keep used the most commonly network dataset (KDDCup99) which is not relevant to employ the machine learning (ML) algorithms for a classification. Several issues regarding these available labelled network datasets are discussed in this paper. The aim of this paper to build a network anomaly detection system using machine learning algorithms that are efficient, effective and fast processing. The finding showed that AODE algorithm is performed well in term of accuracy and processing time for binary classification towards UNSW-NB15 dataset.

1. Introduction
Security-based anomalies are an abnormality or malicious behaviours where the network or systems deviate from their realm usual functionality. This wide-ranging issue in network security need defences or protection tools either software or hardware-based tools that ensure the system in a full-secure from compromise by the attackers. For instance, during the communication or transferring data information between users they might compromise by hackers that have intention to stealth the information. Machine Learning (ML) is one of the efficient and modern technique that enable to monitor the patterns in highly system environment that present today [1].

However, network anomaly detection system using machine learning faced difficulty when dealing the involvement of dataset. There are several issues regarding network anomaly detection dataset for evaluation purpose with the present dataset as discussed in paper [2]. Most datasets not fulfilled the requirements for network security. The labelled dataset is very few in public [3] and the conflicts arise when some of this dataset only specific for certain environment (flexibility), less availability, and a lack of ground truth as state in [4].
Hence, in this paper we construct the network anomaly detection systems that capable to manage the large-scale data and frequently exchange their data through Internet by employing the machine learning (ML) algorithms onto it. The challenges and issues of network anomaly detection datasets motivate this paper to investigate the UNSW-NB15 dataset. The creation of this dataset consists a hybrid of real normal and synthetic network attack data as presented in [5]. Another point is that, UNSW-NB15 dataset overcome the problems faced by researchers with the others dataset where they represent the complex patterns that have balanced data distribution of training and testing set. Advantageous of UNSW-NB15 dataset reflect the high efficiency due to the features of instances collect from the payload to the header of information [6].

The paper organized as followed: In Section 2 present the existing work with the current issues of labelled network dataset. Next, the experimental set-up describes details at Section 3. From the conducted experiments, Section 4 discussed the finding of experiments. Last part of this paper to conclude (Section 5).

2. Related Works
Numerous number of network anomaly detection dataset announced by many community such as NSL-KDD (improvement of KDDCup99 dataset), MAWILab [7], the MoMe Cluster, the Cooperative Association for Internet Data Analysis (CAIDA) and RIPE comprehensively critics by many researchers. For instance, CAIDAs datasets not effective to be used for network anomaly detection where their data resources might be removed during the simulation and it is not adaptable to any environment.

Several works had been done for classification to determine the behavior of data in a system. Authors [8] compared the UNSW-NB15 and KDDCUP99 dataset to measure the accuracy and False Alarm Rate (FAR) using five different ML algorithms and found that DT is an efficient algorithm for classification differ from our paper that investigate the accuracy and other measures by binary classes (Normal (1) and Attack (0)).

In addition, the most famous dataset that had been used by many researchers, DARPA datasets (KDDCup1999), are perceived to skewness and biases classification toward the training set that repetitious their data instances [5][9][10]. Hence, these dataset is impractical for network anomaly detection systems. Moreover, since the data is dynamic in the system causes the new attack might be present in the network. These old-fashioned network anomaly detection dataset also in-comprehensive to represent a modern normal behaviours and contemporary synthesized attack as in UNSW-NB15 dataset [11].

The latest related work is present in paper [12] stated that the datasets consists irrelevant and redundant attributes. They believe that by using features selection result a fast processing and high in accuracy of classifier. They employed ML techniques (Random Forest) toward the datasets. From here motivate this present paper to observe the performance of time taken to train and test the model with varying the size of data.

3. Experiment Setting
In our experiment, Ubuntu software version 13.10-0 ubuntu 4.1 is the operating system and the WEKA tool [13] is run on an Intel Xeon (R) CPU E3-1270 v5 @ 3.60GHz x 8, 16GB RAM to employ ML algorithms toward UNSW-NB15 dataset to evaluate based on their classification rate and processing time in order to build a high efficient and fast processing for network anomaly detection system.

According to Fig. 1 the binary classification for network anomaly detection system toward UNSW-NB15 dataset involved four stages (preparation of dataset, training and testing, build classifier model and performance measure metrics). The experiment begins by loading a network dataset that needed for classification purpose. Once the dataset is ready, the data will be undergoing
the training/testing stage. Before testing stage, the classifier model is required as a decision engine. Finally, to analyse the result from the conducted experiments.

![Diagram](image)

**Figure 1.** Binary Classification (UNSW-NB15 dataset) using Machine Learning Techniques

### 3.1. Labelled Network Dataset – UNSW-NB15

Presently, the labelled network dataset that used for network anomaly detection is UNSW-NB15 dataset. In the papers [14][15], describe details on this dataset. But, in this paper we make some modification toward the original UNSW-NB15 dataset where the number of attributes (features) only 43 features instead of 45 features. It is important to select certain features to get the reliability of the experiment. Feature (id) excluded due to it is just an index and a row of numbers. If it is not excluded the experiment is not reliable because they affect the behaviour of the whole dataset.

Since, the UNSW-NB15 dataset provide two labelled features (i.e. *cat-attack* and *label*). We only used label attributes where the data is either normal or attacks data (binary data). Otherwise, this dataset also can be done for a multi-classification by choosing *cat-attack* as label features instead of *label*. These 43 features/attributes according to their type of features (such as basic, flow, time, content, connection, general, and label) used for this present work is as tabulated in Table 1.
Table 1. Features of UNSW-NB15 Dataset

| No. | Name   | Features | No. | Name   | Features |
|-----|--------|----------|-----|--------|----------|
| 1   | dur    | Basic    | 23  | dwin   | Content  |
| 2   | Proto  | Flow     | 24  | tcprrt | Time     |
| 3   | Service| Basic    | 25  | synack | Time     |
| 4   | State  | Basic    | 26  | ackdat | Time     |
| 5   | spkts  | Basic    | 27  | smean  | Content  |
| 6   | dpkts  | Basic    | 28  | dmean  | Content  |
| 7   | sbytes | Basic    | 29  | trans-depth | Content |
| 8   | dbytes | Basic    | 30  | response_body_len | Content |
| 9   | Rate   | Basic    | 31  | ct_srv_src | Connection |
| 10  | sttl   | Basic    | 32  | ct_state_ttl | General |
| 11  | dttl   | Basic    | 33  | ct_Dst_ltm | Connection |
| 12  | sload  | Basic    | 34  | ct_Src_Dport_ltm | Connection |
| 13  | dload  | Basic    | 35  | ct_dst_Sport_ltm | Connection |
| 14  | slloss | Basic    | 36  | ct_dst_Src_ltm | Connection |
| 15  | dloss  | Basic    | 37  | is_ftp_login | General |
| 16  | sinpkt | Time     | 38  | ct ftp_cmd | General |
| 17  | dinpkt | Time     | 39  | ct_flw_http_mthd | General |
| 18  | sjit   | Time     | 40  | ct_src_ltm | Connection |
| 19  | djit   | Time     | 41  | ct_svr_dst | Connection |
| 20  | swin   | Content  | 42  | is_sm_ips_ports | General |
| 21  | stcpb  | Content  | 43  | label | Labelled |
| 22  | dtcpb  | Content  |      |        |          |

3.2. Machine Learning Tasks
Generally, machine learning tasks involved three schemes (Training, Validation and Testing). Firstly, the dataset is train the classifiers for binary classification (either normal or attack data). During training stage, the features/attributes (either numeric or nominal) likewise source and destination network prefixes, time slot, and etc. are extracted and known to representing the patterns of data. Next, we test our training model using tenfold cross validation to start the evaluation of the classification model for network anomaly detection.

3.3. Build Classifier Model
We employed three types of ML algorithms from Bayesian’s family in WEKA tools. They are Average One Dependence Estimator (AODE), Bayesian Network (BN), and Naive Bayes (NB) to determine the performance in term of classification rate and processing time for classifier model to classify the data instances of UNSW-NB15 dataset. The parameters of these classifiers set to default as in WEKA and using tenfold cross validation to validate the training set before the model been tested.

3.4. Performance Measure
In the context of ML, metrics measure such as accuracy, true positive rate (TP Rate), False Positive Rate (FP Rate), Precision and Recall that representing the performance of binary classification. In addition, this paper presents the varying size of training data with 10 samples (10 000, 20 000, 30 000, 40 000, 50 000, 60 000, 70 000, 80 000, 90 000, and 100 000 data instances) affects the time taken to train and test the classifier model.
4. Result and Discussion

4.1. Performance of ML algorithms for Network Anomaly Detection System

The discussion of the results begins with comparing the AODE algorithms with Bayesian Network and Naive Bayes in term of the correctly classification the data instances (accuracy) and time taken to model the classifier. It is apparent from the chart that (Figure 2) AODE is efficient as well as fast processing for network anomaly detection system compared to another two classifiers with the percentage of accuracy equal to 94.37% and only need to train the model in 4.13s. Meanwhile, BN algorithm that performance almost closed to AODE algorithm with the accuracy is 92.70% and time taken is 4.17s.

![Figure 2. Performance of ML Algorithms for Network Anomaly Detection System](image)

Although, the Naive Bayes algorithm required small amount of time (with 0.79s) to classify the data instances of given dataset (UNSW-NB15) but it is not comparable to AODE and BN algorithms based on classification rate. To measure the performance of this three ML algorithms not enough to consider only their accuracy, the performances of metrics measure (i.e. True Positive Rate (TP Rate), False Positive Rate (FP Rate), Precision (Prec), Recall) of ML algorithms for these binary data (Class 1=Normal and Class 0=Attacks) using UNSW-NB15 dataset need to be investigate as well that tabulated in Table 2.

| ML Algorithm | TP Rate | FP Rate | Prec  | Recall |
|--------------|---------|---------|-------|--------|
|              | Class 1 | Class 0 |       | Class 1 | Class 0 |       | Class 1 | Class 0 |
| AODE         | 0.941   | 0.95    | 0.050 | 0.059  | 0.976   | 0.882 | 0.941   | 0.95   |
| BN           | 0.957   | 0.863   | 0.137 | 0.043  | 0.937   | 0.904 | 0.957   | 0.863  |
| NB           | 0.675   | 0.932   | 0.068 | 0.325  | 0.955   | 0.574 | 0.675   | 0.932  |
4.2. Effect of Time Taken against the Varying Size of Training Data

In our concern, the time taken to build classifier model is enumerating to have an efficient and effective with a fast classification for network anomaly detection system. Figure 3 shows that the effect of time taken by varying the number of training data. The square point is for time taken to build the classifier whereas the cross-point result of time taken to test the model.

The finding showed that NB algorithm is fastest to build the model for network anomaly detection system of binary classification during training or testing scheme where the time taken with increasing the size of training data is in the range of 0.77s-0.94s to train them and to test the model required 1.03s to 2.16s only. As aforementioned that the accuracy of this algorithm is lowest compared the other two ML algorithms. Moreover, the NB algorithm become faster to test the model when the number of training data increases.

Even though, to train the AODE algorithm take a long time compared to NB, they are more fast to test the built model and the correctness to classify the data instances is more robust. Yet, this training time of AODE is linearly with respect to the size of training data and can be learn in incremental way. For instance, with the 50k training data the differences time taken for training and testing for these both algorithm where AODE need double amount of time compared NB algorithm. BN algorithm’s performances, when the large number training data caused the low time taken to train as well as to test their model. This classifier considers slower during training process.

![Figure 3. Effect of Time Taken against the Size of Training Data](image-url)

5. Conclusion

As a conclusion, the UNSW-NB15 dataset is a public network dataset that is relevant for network anomaly detection due to their patterns are complex that represented in modern as well as in contemporary synthesized attacks. From the evaluation made in this paper using ML algorithms (Bayesian group) found that the AODE algorithm is an outperformed, effective, efficient, and fastest classifier for network anomaly detection of binary classification toward the network labelled (UNSW-NB15) dataset.
Acknowledgments
The research reported in this paper is supported by Research Acculturation Grant Scheme (RAGS). The authors would also like to express gratitude to the Malaysian Ministry of Higher Education (MOHE) and University Malaysia Perlis for the facilities provided.

References
[1] B.D.Kumar, K.J. Kumar, Network anomaly detection: A machine learning, CRC Press, 2013.
[2] S. Ali, S. Hadi, T. Mahbod, G.A.Ali, Toward developing a systematic approach to generate benchmark datasets for intrusion detection, computers & security, pp. 357-374, vol. 31, no.3, 2012, Elsevier.
[3] G.Folino, F.S Pisani, P.Sabatino, A Distributed Intrusion Detection Framework Based on Evolved Specialized Ensembles of Classifiers, Applications of Evolutionary Computation: 19th European Conference, EvoApplications 2016, Porto, Portugal, March 30 { April 1, 2016, Proceedings, Part I", Springer International Publishing, pp.315-331, 2016.
[4] R. Koch, M. Golling, G.D. Rodosek, 11 - Towards Comparability of Intrusion Detection Systems: New Data Sets, TNC2014 - TNC2014. (n.d.). Retrieved August 11, 2017, from https://tnc2014.terena.org/core/poster/13.
[5] N. Moustafa, J. Slay, UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). 2015 Military Communications and Information Systems Conference (MiiCIS), pp. 16, 2015. https://doi.org/10.1109/MiCIS.2015.7348942.
[6] N. Moustafa, J. Slay, A hybrid feature selection for network intrusion detection systems: Central points A HYBRID FEATURE SELECTION FOR NETWORK INTRUSION DETECTION SYSTEMS: CENTRAL POINTS AND ASSOCIATION RULES, pp. 513 2015.
[7] F. Romain, B. Pierre, A. Patrice, F. Kensuke, MAWILab: Combining Diverse Anomaly Detectors for Automated Anomaly Labeling and Performance Benchmarking, ACM CoNEXT ’10, p. 12, December,2010.
[8] N. Moustafa, J. Slay, The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set, Information Security Journal: A Global Perspective, vol. 25, pp.18-31, 2016.
[9] F. Gumus, C. Sakar, Z. Erdem, O. Kursun, Online Naive Bayes classification for network intrusion detection, ASONAM 2014 - Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, (Asonam), pp.670-674, 2014.
[10] O.Osanaiye, H.Cai, K.K.R Choo, A.Dehghantanha, Z. Xu, M. Dlodlo, Ensemble-based multi filter feature selection method for DDoS detection in cloud computing, EURASIP Journal on Wireless Communications and Networking, pp.130-139,2016.
[11] D. G. Mogal, S. R. Ghungrad, B. B. Bhusare, A Review on High Ranked Features based NIDS. Ijarccce, 6(3), pp. 349353 2017. https://doi.org/10.17148/IJARCCCE.2017.6380.
[12] J.Tharmin, Z. Shahrzad, Feature selection in UNSW-NB15 and KDDCUP’99 datasets, Industrial Electronics (ISIE), 2017 IEEE 26th International Symposium on, pp.1881-1886, IEEE, 2017.
[13] Weka 3-Data Mining with Open Source Machine Learning in Java, (n.d), Retrieved July 11, 2017, from http://www.cs.waikato.ac.nz/ml/weka/.
[14] The UNSW-NB15 data set description, https://www.unsw.adfa.edu.au/australian-centre-for-cybersecurity/cybersecurity/ADFA-NB15-Datasets/, 2 March, 2016, Accessed date: Sept,2017.
[15] N. Moustafa, J. Slay, The UNSW-NB15 data set description, 2 March 2016. https://www.unsw.adfa.edu.au/australian-centre-for-cyber-security/cybersecurity/ADFA-NB15-Datasets/.
