Distributed Online Averaged One Dependence Estimator (DOAODE) Algorithm for Multi-class Classification of Network Anomaly Detection System

Mukrimah Nawir¹, Amiza Amir¹², Naimah Yaakob¹, R.Badlishah Ahmad²¹, Anuar Mat Safar¹, Mohd. Nazri Mohd. Warip¹ and I Zunaidi¹³

¹Embedded, Network, and Advanced Computing (ENAC) Research Cluster, School of Computer and Communication Engineering, Universiti Malaysia Perlis (UniMAP), Pauh Putra, 02600, Perlis
²Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin (UniSZA), 22200, Besut, Terengganu
³Faculty of Technology, University of Sunderland, St Peter's Campus, Sunderland, SR6 0DD, United Kingdom

E-mail: mukrimahnawir@yahoo.com; amizaamir@unimap.edu.my; naimahyaakob@unimap.edu.my; anuarms@unimap.edu.my; nazriwarip@unimap.edu.my;

Abstract. Network monitoring system consists of large data streams, distributed architecture, and multiple computers that are geographically located all over the world caused a difficulty to detect abnormalities in the system. In addition, when handling network traffic, the data in network is fast incoming and requires an online learning where immediately response and predict the pattern of network traffic for classification once there is an event or request occur. Therefore, this paper aims to develop an effective and efficient network anomaly detection system by using distributed online averaged one dependence estimator (DOAODE) classification algorithm for multi-class network data to overcome these issues. The finding of DOAODE algorithm for multi-class classification is high in accuracy with average 83% and fast to train the network traffic recorded less than ten seconds and takes shorter time when the number of nodes increases.

1. Introduction
Network traffic can be classified into binary or multi-class classification. Binary classification is the process to classify that involved only two classes of network traffic either it is normal or attack class [1]. Whereas, multi-class classification is where the network traffic contains several classes which is a normal and several network threats present in a network system. Commonly, the network anomaly detection system is built for binary classification that insufficient to detect the network traffic’s
patterns in a system and motivates this paper to investigate the performance of classifier to detect anomalous data that dealing with multi-class network data [2][3].

The architecture of network system can be either a centralized or distributed computing. For distributed computing, the results of prediction are combining geographically by allocating multiple classifier at every distributed sites and the local data analyses that hold the information and prediction of network traffic. The knowledge from these all local monitoring sites being combined by using distributed aggregation for making a final prediction. Additionally, the network traffic fast incoming into the system and require a fast learning classifier (an online learning) that enable a frequently update the classifier for classification [4][5]. Therefore, this paper develops an effective and efficient distributed online averaged one dependence estimator (DOAODE) algorithm for multi-class classification of network anomaly detection system performed on the UNSW-NB15 dataset.

The structured of this paper as follows: Section 2 is the previous works that develop a network anomaly detection system by using a machine learning algorithm. Subsequent section is the experiment setting including the network dataset, classifier, and distributed online classifier of updating the classifier (Section 3) followed by the results (Section 4) of conducted experiment based on classification rate and training time for network anomaly detection of the proposed method. Last but not least, the conclusion and some recommendation to improve this present work is provided in Section 5.

2. Related Works

In this section, the related works of developing a network anomaly detection system by using machine learning algorithm were summarized and tabulated in Table 1. In [6], the authors developed network anomaly detection by using centralized batch AODE to classify the multi-class of network traffic in NSL-KDD dataset. An improvement made by using ensemble techniques proposed by [7] and produced an accuracy of multi-class classification is 98.73% with 38 numbers of attack types. Although, their work high in accuracy with highest number of attacks compared this current work but they do not consider the time for classification which is important to ensure the performance is efficiently without buffering the network traffic for making a prediction.

Authors by [8] proposed an online Naïve Bayes classifier to solve the issue of profile change over time due to network traffic it inspects and need to retrain with the recent traffic instances. According to their work, online learning able to handle the presence of continuous flow of data where each of new traffic instance arrived were involved in the training phase to updating the network anomaly detection system. Yet, the centralized classifier that they build bring up an issue of single point failure to occur when there are powerful attacks compromised the system cause fatal destruction toward the network system. Therefore, it is important to design a classifier in distributed architecture to overcome this issue.

The closer work to this present work was presented in paper [9], where the authors designed a distributed online Naïve Bayes (NB) algorithm instead of averaged one dependence estimator as a classifier that detect the network traffic’s patterns. Also, the authors used different dataset (NSL-KDD) dataset which is outdated dataset and had been criticism by many researchers. However, results in low classification rate due to the assumption of NB algorithm where the attributes independently to each other. Hence, to solve these issues an appropriate and dependence features assumption machine learning algorithm which is averaged one dependence estimator (AODE) algorithm to increase the accuracy and fast training classification to discriminate normal and several attacks (multi-class network traffic) by using new benchmark UNSW-NB15 network dataset.
Table 1. Related Works of Network Anomaly Detection by using Classification Algorithm

| Ref. | Algorithm | Architecture | Learning | Dataset     | Classification   | Performances                                      |
|------|-----------|--------------|----------|-------------|------------------|---------------------------------------------------|
| [6]  | AODE      | Centralized  | Batch    | NSL-KDD     | Multi-class      | Accuracy multi-class classification (3 classes) = 96.64% |
| [7]  | Ensemble  | Centralized  | Batch    | NSL-KDD     | Multi-class      | Accuracy = 98.73%                                  |
| [8]  | NB        | Centralized  | Online   | KDD-CUP99   | Binary, Multi-class (measured only F1-score, no accuracy result provided) | Accuracy (binary classification) = 93%, Running time = 768.28 seconds |
| [9]  | NB        | Distributed  | Online   | NSL-KDD     | Binary, Multi-class | Binary classification accuracy = 97.05%, time = 6.23 seconds multi-class classification accuracy = 85%, time: 24.87% |
|      | AODE      | Distributed  | Online   | UNSW-NB15   | Multi-class      | Accuracy in the range 83%, Training time < 10 seconds |

3. Experiment Setup

The proposed to develop network anomaly detection system consists of three major steps: online classification, distributing network data, and distributed aggregation as in Figure 1. Presently, this experiment develops a distributed online classifier for network anomaly detection system toward the labelled network dataset named UNSW-NB15 dataset. This dataset consists 44 features with multi-class (ten class; where one is a normal class and the rest are attacks type).

![Figure 1. Proposed Method of NADS](image-url)
3.1. UNSW-NB15 Dataset

The network dataset, UNSW-NB15, is public and available at [10]. It is a newest labelled network dataset that provided two label features either conducted a binary (label feature used is “label”) or multi-class classification (label feature is “attack_cat” used). Also contains realistic normal and several type of attacks that allowing the benchmarking dataset to representing the real network system. In this present work, there only 44 features used instead of 49 features provided by the contributor due to the unnecessarily to be used and to increase the performance of network anomaly detection system. Additionally, there is 2,540,044 data instances of this dataset but for classification purpose the raw data in csv file converted to arff file with the amount about 257,673 data instances. For multi-class classification, there are nine types of different attacks being considered with one normal traffic. They are reconnaissance, shellcode, backdoor, analysis, generic, worms, Dos, fuzzers, and exploits [11] [12].

3.2. Distributed Online Averaged One Dependence Estimator (DOAODE) Algorithm

Averaged One Dependence Estimator (AODE) algorithm [6] [13] [14] is a supervised machine learning algorithm that postulate the features/attributes are dependent one to another feature by averaging all the classifier generated. Moreover, this algorithm can be design in online manner to solve the issue of frequent updated classifier and produced a fast classifier to train the network traffic. In this work, we design the AODE algorithm within a distributed architecture by distribute the network traffic to the present nodes in network. As these nodes has its own classifier (AODE algorithm) for local prediction. The distributed online averaged one dependence estimator (DOAODE) algorithm works as in Figure 2.

For instance, when there are five nodes in a network then each of the nodes hold an AODE classifier so there are five AODE classifiers present in the system for multi-class classification the UNSW-NB15 dataset. Later, one of the nodes that act as a global classifier to aggregate all the results of prediction from local classifiers to make a final prediction of network traffic as either normal or anomalous data by using distributed majority voting. For distributed majority voting, we only consider an odd number of nodes, then when the major or highest voting determine the pattern of network traffic. For example, out of these five nodes three of them predict that the network traffic is a normal traffic and another two nodes are predicted that the patterns are anomalous and the obtained result out of these voting the network traffic is classified as a normal pattern.
DOAODE Algorithm:

Input: instance \( \mathbf{x} = < x_1, \ldots, x_m > \)

Output: the class estimation of AODE

Training phase:
1. Begin
2. Given an instance \( \mathbf{x} = < x_1, \ldots, x_m > \)
3. At initiator node: find the neighbour nodes, NN
4. Calculate probability of \( y \) given \( x \): \( p(y|x) = p(y|x)p(x) \) then
5. \( p(y|x) = \frac{p(y|x)p(x)}{p(x)} \)
6. End

Testing phase:
9. Begin
10. Find the attribute values \( x_i \in \mathbf{x} \) in test sample
11. Calculate \( p(y|x, x_i) = p(x, x_i, y) = p(y, x_i)p(x|y, x_i) \)
12. Calculate probability of \( y \) given \( x \): \( p(y|x) = \frac{p(y|x)p(x|y, x_i)}{p(x)} \)
13. Upon receiving results from nodes NN
14. Output the class values
15. Aggregate the results by using majority voting
16. Update classifier
17. Repeat
18. End

Figure 2. Distributed Online Averaged One Dependence Estimator (DOAODE) Algorithm

4. Results

The proposed distributed online averaged one dependence estimator (DOAODE) algorithm is measured its performances based on classification rate and training time on the network dataset which is UNSW-NB15 dataset that consists 257,673 instances, 10 classes, and 44 features.

4.1. Classification Rate of DOAODE Algorithm for Network Anomaly Detection System

Table 2 is the result of classification rate (accuracy) of distributed online averaged one dependence estimator (DOAODE) algorithm being compared with another different designs which are distributed online Naïve Bayes (DONB), distributed batch averaged one dependence estimator (DBAODE), and distributed batch Naïve Bayes (DBNB) algorithm for network anomaly detection of the multi-class classification. The accuracy of DOAODE and DBAODE algorithms are outperformed and higher than DONB as well as DBNB algorithm where both designs AODE algorithm achieved averagely about 83%.

Meanwhile, the Naïve Bayes algorithm only obtained in average 70% for multi-class classification UNSW-NB15 dataset for network anomaly detection system when the number of nodes is varies in the network system. The trend of accuracies results for DOAODE algorithm inconsistent depending the global node that made final prediction after collecting all the results from the local nodes that present in the network system. Due to distributed classifier cause a difficulty in making decision and when the number of nodes increases the number of conflicting local prediction results. Yet, the distributed classifier avoids the single point failure of the network to be occurred where when
there is a powerful network threat compromise the system cause the whole destruction of network system.

Table 2. Classification Rate of DOAODE Algorithm for NADS

| Classifier/Number of Nodes | Classification Rate (%) |
|---------------------------|--------------------------|
|                           | DOAODE | DONB | DBAODE | DBNB |
| 5                         | 83.87  | 70.89| 83.83 | 70.84 |
| 10                        | 83.91  | 70.84| 83.89 | 70.87 |
| 15                        | 83.70  | 70.83| 83.79 | 70.84 |
| 20                        | 83.67  | 70.73| 83.79 | 70.79 |
| 25                        | 83.51  | 70.74| 83.81 | 70.73 |

4.2. Training Time of DOAODE Algorithm for Network Anomaly Detection System

When dealing with network data in a system the importance measurement is the training time due to the fast data incoming and rapid updating data in a network required an efficient classifier that can determine the patterns of network data for classification purpose. Additionally, most of the machine learning algorithms have low classification time therefore the concern is the training time for a classifier to classify the network traffic where usually expensive and increase dramatically with an increasing of training data. Table 3 shows the experimental results for DOAODE algorithm compared with various designs of classification algorithm including the DONB, DBAODE, and DBNB algorithm for network anomaly detection based on training time.

From the experimental, the proposed method (DOAODE algorithm) shows a reasonable faster classifier to train the network traffic where with the highest number of nodes (25 nodes) produced only 1.78 seconds to build the classifier model for anomaly detection. In addition, the obtained result shows that the decreasing trend of training time with the number of nodes increasing (network size become large). Although, it is slower training classifier compared to DONB still the amount of training time of DOAODE algorithm acceptable as a fast classifier for anomaly detection which is took less than 10 seconds.

Table 3. Training Time of DOAODE Algorithm for NADS

| Classifier/Number of Nodes | Training Time (Seconds) |
|---------------------------|--------------------------|
|                           | DOAODE | DONB | DBAODE | DBNB |
| 5                         | 8.76   | 0.89 | 25.25  | 1.00 |
| 10                        | 4.96   | 0.46 | 28.39  | 0.49 |
| 15                        | 3.73   | 0.31 | 56.54  | 0.32 |
| 20                        | 2.19   | 0.25 | 63.11  | 0.24 |
| 25                        | 1.78   | 0.19 | 74.23  | 0.20 |

5. Conclusion

In summary, this paper proposed a distributed online implementation of averaged one dependence estimator (DOAODE) classifier for network anomaly detection system. Based on the independence assumption of Naïve Bayes classifier motivate this paper to choose an AODE algorithm that is assuming the attributes’ dependency by averaging all the estimation of probabilities of the present models in network traffic accurately gives an outperformed, high classification accuracy, and fast training time for anomaly detection the network dataset (UNSW-NB15) where the obtained of
accuracy for multi-class classification is in the range 83% and took less than 10 seconds when the number of nodes are varies from 5 to 25 nodes.

6. References

[1] Nawir M, Amir A, Lynn OB, Yaakob N, Ahmad RB. Performances of Machine Learning Algorithms for Binary Classification of Network Anomaly Detection System. In Journal of Physics: Conference Series 2018 May (Vol. 1018, No. 1, p. 012015). IOP Publishing.

[2] Nawir M, Amir A, Yaakob N, Lynn OB. Multi-Classification of UNSW-NB15 Dataset for Network Anomaly Detection System. Journal of Theoretical & Applied Information Technology. 2018 Aug 15;96(15).

[3] Rettig L, Khayati M, Cudrée-Mauroux P, Piórkowski M. Online anomaly detection over big data streams. In2015 IEEE International Conference on Big Data (Big Data) 2015 Oct 1 (pp. 1113-1122). IEEE.

[4] Ahmad S, Purdy S. Real-time anomaly detection for streaming analytics. arXiv preprint arXiv:1607.02480. 2016 Jul 8.

[5] Limthong K. Real-time computer network anomaly detection using machine learning techniques. Journal of Advances in Computer Networks. 2013 Mar;1(1).

[6] Sultana A, Jabbar MA. Intelligent Network Intrusion Detection System using Data Mining techniques. In Applied and Theoretical Computing and Communication Technology (iCATcT), 2016 2nd International Conference on 2016 Jul 21 (pp. 329-333). IEEE.

[7] Vasudha K. Intrusion Detection System using Decision Tree-based Attribute Weighted AODE. International Journal of Advances Research in Computer and Communication Engineering, 2014 December (pp. 451-467).

[8] Gumus F, Sakar CO, Erdem Z, Kursun O. Online Naïve Bayes classification for network intrusion detection. InProceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2014 Aug 17 (pp. 670-674). IEEE Press.

[9] Idhammad M, Afdel K, Belouch M. Distributed Intrusion Detection System for Cloud Environments based on Data Mining techniques. Procedia Computer Science. 2018 Dec 31;127:35-41.

[10] Weka 3 – Mining Big Data with Open Source Machine Learning Software in Java. Retrieved 2016-11-21, from http://www.cs.waikato.ac.nz/ml/weka/bigdata.html

[11] Moustafa N, Slay J. UNSW-NB15: A Comprehensive Data set for Network Intrusion Detection Systems (UNSW-NB15 Network Data set). In Military Communications and Information Systems Conference (MilCIS), 2015 2015 Nov 10 (pp. 1-6). IEEE.

[12] Moustafa N, Slay J. 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. 2016 Apr 4;25(1-3):18-31.
[13] Webb GI, Boughton J, Wang Z. Averaged one-dependence estimators: preliminary results. *In Proceedings of the Australasian Data Mining Workshop 2002 Dec 2 (Vol. 2002). Language Thought, and Reality: Selected Writings of Benjamin Lee Worf*. MIT Press.

[14] Webb GI, Boughton JR, Wang Z. Not so naive Bayes: aggregating one-dependence estimators. *Machine learning*. 2005 Jan 1;58(1):5-24.

**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 Higher of Education (MOHE) and University Malaysia Perlis (UniMAP) for the facilities provided.