Journal of Theoretical and Applied Information Technology
15th August 2018. Vol.96. No 15
© 2005 – ongoing JATIT & LLS
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

MULTI-CLASSIFICATION OF UNSW-NB15 DATASET FOR NETWORK ANOMALY DETECTION SYSTEM

[1]MUKRIMAH NAWIR, [2]AMIZA AMIR, [3]NAIMAH YAAKOB, [4]ONG BI LYNN
[1,2,3,4]Embedded, Networks and Advanced Computing Research Cluster (ENAC),
School of Computer and Communication Engineering (SCCE),
Universiti Malaysia Perlis (UniMAP), 02600, Pauh, Perlis, Malaysia
E-mail: [1]mukrimahnawir@yahoo.com, [2]amizaamir@unimap.edu.my, [3]naimahyaakob@unimap.edu.my,
[4]drlynn@unimap.edu.my

ABSTRACT

Problem to classify more than two classes (called as multi-class) for network anomaly detection system using machine learning techniques are very challenging and become a vital factor when the growth of many network attacks might endanger the performances of network system. A tremendous increase in the various number of network threats compromise the network system motivate the network anomaly detection system to be relevant and necessary to be implement using a powerful tool (machine learning approach) for network security issue. In this work, a model of an Online Average One Dependence Estimator (AODE) algorithm for multi-classification of UNSW-NB15 dataset that high in accuracy with a low false alarm rate (FAR) was built to overcome the issues such as the nature of data (complex data that represent into more than two classes), dynamical data in a network system, and frequent update (for streaming data that need a fast processing). The obtained results from the conducted experiment showed that Online AODE more recently detect the Worms class where the percentage of accuracy for classification is 99.93% with small FAR is only 0.001. Moreover, online AODE is an outperformed based on accuracy compare to online Naïve Bayes (NB) where the classification rate 83.47% and 69.60% respectively for multi-classification the UNSW-NB15 dataset. Since, the given data is a streaming data in a computer network time need to be enumerated to have a fast algorithm for network anomaly detection system before the network system become in a critical condition. Although, the online NB is most fastest for multi-classification yet online AODE give a comparable result based on processing time.

Keywords: *Multi-classification, Network Anomaly Detection System, Averaged One Dependence (AODE), Machine Learning, UNSW-NB15 Dataset*

1. INTRODUCTION

Network Anomaly Detection System (NADS) monitor the behavior network data/patterns in a computer network system. It can be implemented in two learnings either batch or online learning [1]. Anomalies in network data in this case represent network attacks. In simple word, network attacks are the deviation from the normal behavior of network data. It motivates the machine learning tool, it is automatically detecting the worthwhile network by training and testing the data instances in a given network dataset [2][3].

Several works for network anomaly detection had being applied to various environments. The built NADS tremendously specific to the environment that fulfilled their needed. The examples of anomaly detection projects are in aircraft engine measurement, cloud datacenter temperature, telecommunication, and ATM fraud detection. The successful approaches not suitable to all environment. In other word, some approaches well-perform to a real-time anomaly detection. There can be implemented for binary (binary classification) [4]–[8] or multi class (multi-classification) [4], [7]–[13]. It is insufficient to investigate the performance of binary classification that only consider either the network data is normal or anomalous data. Therefore, we performed multi-classification for network anomaly detection toward UNSW-NB15 dataset to determine the recent classes that correctly classified using the built online machine learning algorithm.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

This paper only works and focuses on online network anomaly-based where the machine learning approach is employed to learn the classes of UNSW-NB15 dataset either it is a normal data or various anomalous data. Online learning enable the classifier model to continuously update the features where each data instance arrived included during training phase [8]. It is important to model an online classifier regarding the network data usually dynamic in a network system, where the data keep changes their behavior due to the join or leave the connection. Moreover, the comparison of online and batch classifier based on time taken for multi-classification the UNSW-NB15 dataset is investigated in the last part of this current work.

Since, multi-classification is to classify the network dataset that contains more than two classes there are difficulty and problem to classify the multi-class of given dataset is challenging where the imbalanced data distribution might reduce the classification rate (accuracy) for network anomaly detection [14]. Many researchers only build network anomaly detection for binary class either it is normal or anomalous data. This present work provides a more complex, depth-in and specific investigate the class that recently recognized by a machine learning algorithm.

The contributions of this work as follow:

1. Input data used represent a relevant and complex network data (UNSW-NB15 dataset).
2. Supervised learning of ML algorithm for network anomaly detection (AODE).
3. Conducted a multi-classification, that classify the network data more than two classes which more specific to the subclasses from the given dataset. This is to investigate in which classes being classified accurately.
4. Built and design an online classifier (online AODE) to frequent data and handle the issue nature of network data that dynamic in a computer network system.

The remainder topics in this paper is arranged as follows: Section 2, the related works of network anomaly detection system. Subsequent Section 3 is a setup of conducted experiment that explained the chosen dataset and machine learning algorithms. After that, the obtained results presented and discussed in Section 4. We conclude the findings in Section 5.

2. RELATED WORKS

A considerable amount of literature had been published on network anomaly detection system (NADS). These studies motivate by issues such as handling massive number of network data, the network data is dynamical, and the frequently update the data arise regarding monitoring network by detecting any deviation from the normal behaviors/patterns of network data. Authors [8] proposed an online Naïve Bayes classifier for binary classification (2-class problem) as well as multi-classification (23-class problem and 5-class problem) over KDDCUP99 dataset to solve the issue data changing in the computer network system. However, their work not concerning the time taken to complete the multi-classification of UNSW-NB15 dataset which is a vital element when built a network anomaly detection system in a computer network.

Also, they stated that most recent classes permitted the classifier adaptable to the new attack which may advance over time. This scenario called as a bias classification, where the imbalance class with high amount data instances most classified correctly whilst the low amount of data instances tend to be ignore. They found that online NB is time efficient and more accurate than others ML algorithm chosen. Hence, this present work on a multi-classification of UNSW-NB15 to investigate the performance of our chosen ML algorithm (online AODE) used to overcome that issue and compared them. The used of online classifier enable the frequent update data over time for the continuous data in a computer network is necessary, relevant, and acceptable.

Furthermore, Rettig et al. [13] stated that detecting anomalies that dynamically passing over time required a real-time (online classifier) that consider a generality and scalability issue. Their work deficient to distinguish the dynamic and robust anomalous in a network system as in this current work. The geographically location of computer network that connected via Internet motivate a fast streaming data which is online learning is needed. Hence, they evaluate the online anomaly detection over big data streams. In the context of machine learning, there is exclusive constraint faced to develop network anomaly detection within real-time applications [15]. Differ from our work, that consider the accuracy for multi-classification and time efficiency that consist a huge number of network data and dynamical data from the given network dataset (UNSW-NB15).

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

To improve the accuracy of classification an online ensemble classification is built by Bai et. al. using several Bayesian classifiers instead used an online boosting algorithms for fast continuous data with a real-world datasets [16]. Additionally, real-time applications required an automatic detecting the anomalous data in a network caused uncommon system behavior occurs and it is necessary to have fast anomaly detection classifier as in papers [17] that develop dynamic Bayesian networks and authors [18] consider fast algorithm with a high classification rate using a streaming half-space trees for network anomaly detection. Meanwhile, the method used in this current study does not take any assumptions about the network data instances, where the network dataset is labelled (supervised learning) include a multidimensional as well as categorical data.

Overall, this study highlight the need for machine learning approaches for network anomaly detection as had been conducted in the paper [19]. Their work more likely to this current study, by built a real-time network anomaly detection system using machine learning algorithms including NB, Support Vector Machine (SVM), and Decision Tree (DT) in the campus site that consider the scalability, fault-tolerance, and resiliency for monitoring the network traffic. Additionally, this paper presents multi-classification of UNSW-NB15 network dataset in online as well as batch learning to investigate the classification rate (accuracy) and time efficiency of the chosen ML algorithm and compared to the most common used ML algorithm (Naïve Bayes).

Authors performed experimental investigations based on detection accuracy, false positive rate, false negative rate, and time taken using hybrid of two supervised machine learning algorithms for network anomaly detection by observing the behavior of network traffic [5]. Differ from our work, another work conducted a batch AODE algorithm for network anomaly detection by excluding the processing time (that should be enumerated when dealing with a streaming data) for multi-classification [20].

We conducted experiments to handle the issues of large scale data, multi-class, continuous data (dynamic data), and frequent update data to detect which type of classes most detect accurately and fast using our chosen ML algorithm (online AODE algorithm) for network anomaly detection system (NADS). This is the summary of previous works of multi-classification for network anomaly detection system in Table 1.

Table 1: Related Works of Multi-Classification for Network Anomaly Detection System

Ref.	Issues	Dataset	Algorithm
[7]	1. Dynamic data 2. Detection low response time	UNSW_NB15	RepTree
[8]	1.Develop technology 2.Nature of network data 3.Continuous network data	KDDCUP99	Naïve Bayes
[9]	1.Large amount of data 2.Problem of independence features 3.Classification problem	NSL-KDD	AODE
[10]	1.Low detection rate and high false positive rate	NSL-KDD	k-means
[11]	1.Develop a scalable, fault-tolerance and resilient NADS 2.Handle real-time data	Real network data at campus	Naïve Bayes
[12]	1.Detecting various anomalous in computer network	Network data at Kasetsart University	Time series and feature spaces

3. EXPERIMENT SETTING

Whole work run by simulation on the computer was carried out using Ubuntu software version 13.10-0 ubuntu 4.1 operating system, an Intel Xeon ® CPU E3-1270 v5 @ 3.60GHz x d, 16GB RAM. WEKA tool (WEKA 3.8) [19] is used to employ ML algorithms toward UNSW-NB15 dataset as well as an eclipse to write the JAVA language to model an online classifier.

The experiment begins with load the network labelled dataset that need for a classification. After that, the training or learn the data instance is important to evaluate the performance of machine learning classifier. Finally, the built classifier undergoes testing stage. But before that, tenfold cross-validation is used. Then measure the performances such as accuracy, True Positive Rate (TPR), and False Positive Rate (FPR).

Machine learning tasks include three phases which are a training, validation, and testing stages. In addition, the chosen built online classifier (online AODE) compared their performances with

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

the most common used algorithm for network anomaly detection system (online NB).

3.1 Labelled Network Dataset (UNSW-NB15)

Generally, for network anomaly detection system researchers used KDD CUP 1999 dataset. Yet, there are many criticisms on this dataset that had been presented in many papers [21][22][23]. Hence, in this present work the labelled network UNSW-NB15 dataset used instead KDDCUP99 dataset. UNSW-NB15 dataset is represent as a hybrid of the real normal and modern synthesized attack of network data [21]. Since, the dataset labelled we implement ML algorithm for network anomaly detection in supervised learning.

For our work, the UNSW-NB15 dataset contains 257 673 data instances with 44 features (in Table 2). The total classes of this dataset are 10 classes: one is for a normal network data (93 000 instances) and nine classes of anomalous network data (attacks classes). The attacks involved were backdoor (2 329 instances), analysis (2 677 instances), fuzzers (24 246 instances), shellcode (1 511), reconnaissance (13 987 instances), exploits (44 525 instances), DoS (16 353 instances), worms (174 instances), and generic (58 871 instances). The data distribution data of UNSW-NB15 as in Figure 1.

Table 2: Features of UNSW-NB15 Dataset

No.	Features	No.	Features
1	*id*	23	*dtrcpb*
2	*dur*	24	*dwin*
3	*Proto*	25	*tcprtt*
4	*Service*	26	*synack*
5	*State*	27	*ackdat*
6	*spkts*	28	*smean*
7	*dpkts*	29	*dmean*
8	*sbytes*	30	*trans_depth*
9	*dbytes*	31	*response_body_len*
10	*rate*	32	*ct_srv_src*
11	*sttl*	33	*ct_state_ttl*
12	*dttl*	34	*ct_dst_ltm*
13	*sload*	35	*ct_src_dport_ltm*
14	*dload*	36	*ct_dst_sport_ltm*
15	*sloss*	37	*ct_dst_src_ltm*
16	*dloss*	38	*is_ftp_login*
17	*sinpkt*	39	*ct_ftp_cmd*
18	*dinpkt*	40	*ct_flw_http_mthd*
19	*sjit*	41	*ct_src_ltm*
20	*djit*	42	*ct_srv_dst*
21	*swin*	43	*is_sm_ips_ports*
22	*stcpb*	44	***attack_cat***

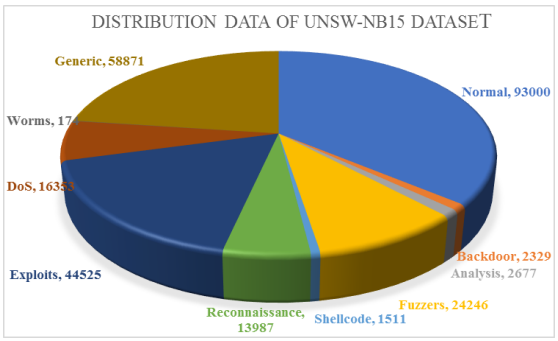

Figure 1: Distribution Data of UNSW-NB15 Dataset

For online AODE the data from given dataset being discretize. This is because online AODE (A1DEUpdateable as in WEKA tool) cannot run the continuous data to learn the knowledge of the data within computer network. Discretization is to transform the network data instance from numeric data into nominal data. It is important because the raw data when the online AODE employed is not ideal form in numeric data.

3.2 Machine Learning Algorithm: Averaged One Dependence Estimator (AODE)

Machine Learning (ML) approach involve three steps: training, validation, and testing. First, the dataset is loaded for multi-classification which consist ten classes. The network data is discretized to modify the network data of given dataset. So, the numeric features convert into nominal type. Second step is the data was learnt by ML algorithm to train the network data. Third, the built model (during training stage) was testing that beforehand undergo tenfold cross validation to start the evaluation of multi-classification for network anomaly detection over the UNSW-NB15 dataset.

Average One Dependence Estimator (AODE) classification is an advance Naïve Bayes algorithm by assume the features is dependence that averaging all the estimation of probabilities of the built classifier in given network dataset [20][24]. The default mode of AODE classifier on Weka is a non-incremental that is probabilities are computed at learning time. We design an online AODE classifier (used via option -I) to keep update the features of continuously data in a computer network one at a time.

The advantages of AODE algorithm for network anomaly detection of multi-classification over UNSW-NB15 dataset are suitable for large dataset, high in accuracy, can classify the data instances that consists more than two classes, enable to predict the class probabilities for each class present in a given

Journal of Theoretical and Applied Information Technology
15th August 2018. Vol.96. No 15
© 2005 – ongoing JATIT & LLS
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

dataset, and low variance. Due to these advantages, it is well-suited and well-performed for network anomaly detection [9] .

4. RESULT AND DISCUSSION

Multi-Classification of UNSW-NB15 dataset for network anomaly detection system measure in several parameters which are true positive rate (TPR) rate, false positive rate (FPR) and accuracy of chosen built ML algorithm (online AODE) that been employed regarding to the advantages of AODE algorithm. To measures these parameters, confusion matrix is necessary and tabulated as in Table 4 (confusion matrix for online AODE) and Table 5 (confusion matrix for online NB). The built network anomaly detection should be high in true positive rate with low false positive rate.

Additionally, the comparison between two machine learning algorithms for network anomaly detection over UNSW-NB15 dataset discussed detail based on their accuracy. Table 3 is the comparison of performance measure in term of time taken for multi-classification the UNSW-NB15 dataset using online AODE, online NB, batch AODE and batch NB algorithm.

The chosen dataset in our work consist a complex, real normal, and modern synthesized attack of network data that can be recognized well by chosen algorithm (online AODE) for multi-classification that differ from the work in [13] that faced a difficult to recognize a complex and multi-dimensional data in a given dataset that continuously updated the data. Therefore, the built classification algorithms cannot be employed over the UNSW-NB15 dataset.

Subsequently, the results for every measure presented and discussed as well as the confusion matrix of AODE and NB algorithms (as in Table 4 and Table 5 respectively) for multi-classification against UNSW-NB15 dataset). Followed by the finding of experiment that compared the performance of online AODE and online NB for network anomaly detection. These performance measures similarly presented as in paper [5].

4.1 True Positive Rate (TPR)

To measure the true positive rate the number of correctly predicted class over the number of actual class as in equation (1). If the TPR value of ML algorithm close to 1 means that the ML algorithm is well performed for network anomaly detection system. Otherwise, the ML algorithm worst to recognize the type of classes present in a computer network system.

AODE algorithm give best TPR toward the Generic attack class (TPR=0.9850) compare to others class. The worst TPR is the classification of Backdoor attack data with only TPRate equal to 0.1670. In other word, online AODE not recently recognize Backdoor attack in a UNSW-NB15 dataset. The trend of bar chart produced shows a variation of TPR values for different classes of UNSW-NB15 dataset. This is influenced by the factor of imbalanced data in a given dataset.

For instance, the highest TPR of online AODE algorithm towards UNSW-NB15 dataset is a Generic type where its TP's amount 57968 data instances over the total amount data for Generic classes (TP+FN) is 58871 number of data instances produces a TPR equal to 0.985.

Our design comparable in term of TPR value when compared to work of [9] that consider small number of classes. Even we used more classes than their work the TPR difference is insignificantly (only differed 0.02, when their work recorded the highest TPR value by classified the U2R and R2L attacks equal to 0.987).

$$TPR = TP / (TP+FN) \quad (1)$$

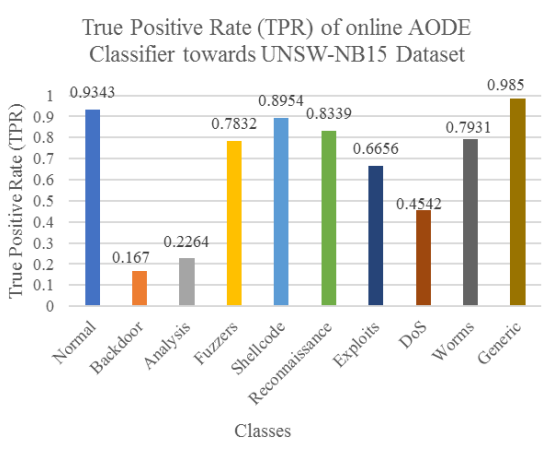

Figure 2: True Positive Rate of online AODE Classifier towards UNSW-NB15 Dataset

4.2 False Positive Rate (FPR)

FPR or False Alarm Rate (FAR) is the ratio of number of normal class that misclassified (classified as attacks) over the total number of normal classes. In simple word, a case where an anomaly detected in the computer network

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

system, but it is not in a real world. Hence, the value should be close to zero unlikely to the aforementioned of the TPR value for network anomaly detection. Equation 2 is the formula to calculate the FPR. The lowest the value of FPR the best performance.

An online AODE algorithm with small FPR value is in detecting the Worms with only 0.001 value and worst-case for classify the Exploits and DoS class ($FPR_{Exploits}$ = 0.053 and FPR_{DoS} = 0.052) as in bar chart shown in Figure 3. Most of the classes present the lowest FPR value in the range 0.001 to 0.005 which is best result for multi-classification.

Whilst, the FPR obtained in the previous work [5] that employed an ensemble machine learning algorithm very low (from 0.09% to 11.50%) due the capability of ML algorithm and it only implement the NADS for binary classification. The drawback is cannot determine the more specific recognition toward the classes present.

$$FPR = FP / (FP+TN) \quad (2)$$

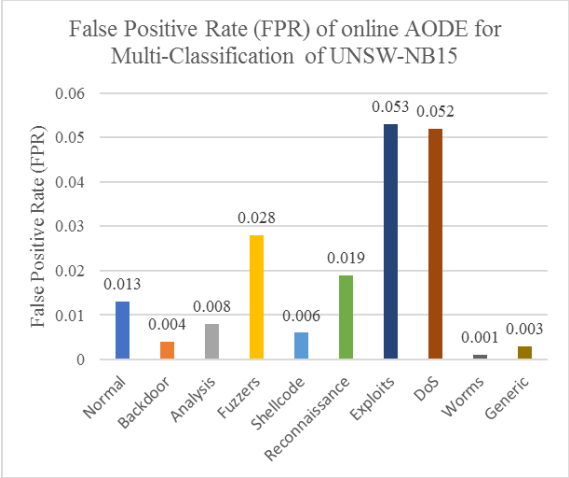

Figure 3: False Positive Rate (FPR) of Online AODE for Multi-Classification of UNSW-NB15 Dataset

4.3 Accuracy

Once, the value of TPR and FPR are determined. The accuracy can be calculated as in equation 3 for every classes present in a given network labelled dataset. The TP values can be obtained in a confusion matrix as in Table 4 and Table in the grey color for every classes of UNSW-NB15 dataset. Whilst, TN values for certain class is the sum of all columns and rows excluding that class's column and row. For network anomaly detection system, accuracy is an important element and need to be consider proving that the work is relevant and overcome any issues regarding the multi-classification process.

For instance (from Table 4), the calculation to measure the accuracy for normal class that consists 93000 amount of data instances in UNSW-NB15 dataset. The value of TP equal to 86893 data instances and TN's value equal to 162470 data instances. Hence, the accuracy of online AODE to classify the normal class is the division of summation TP and TN values with total number of instances that consists 257673 data instances where the result equal to 0.9677 (96.77%).

From the graph in Figure 4, the Worms class is the most recent detect by using online AODE (accuracy = 99.93%) and worst to detect Exploits class (accuracy = 89.81%). The imbalanced distribution class caused the varies of accuracy for network anomaly detection. This called biased classification. The less number of data might produce a low classification rate, and this also will affect the effectiveness of network anomaly detection system for ML algorithm to do a classification that depending to all classes.

The findings that resulted in [9] due to the classification built the NADS with the highest accuracy for multi-classification using AODE algorithm produced 97.19% to classified the DoS attack class when they conducted their experiment towards four classes (U2R, R2L, Probes, DoS) which is lowest when compared to our works, highest accuracy to classify the Worms attack with 99.93%.

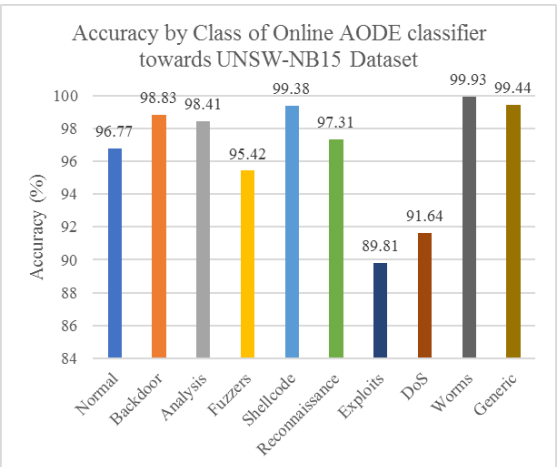

Figure 4: Accuracy by Class of Online AODE Classifier towards UNSW-NB15 Dataset

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Accuracy by class = $(TP_1+TP_2+\ldots+TP_n) + (TN_1+TN_2+\ldots+TN_n) / (N)$ (3)

where TP is number of attack class correctly classified, n number of classes (n=1, 2, 3, …,10), TN is number of normal class classified correctly, N (257 673 instances) is total number of data instances of given dataset

Though, many classes involved in this present work (ten classes) the accuracy by class for each class in a given dataset consider higher from their works to detect the normal as well as anomalous network data. From this ten classes, 3 of them (Fuzzers=95.42%, Exploits=89.81%, and DoS=91.64%) are lower than previous work (below 97.19%). Therefore, online AODE algorithm is an outperformed for multi-classification with large amount of data instances and large number of features.

4.4 Comparison online AODE and online NB

There are most commonly found that NB classifier for network anomaly detection in various domains give an outperformed result [8], [11], [4]. But, in this present work with a given dataset (UNSW-NB15) showed that online AODE gives the best performances (percentage of accuracy, TP Rate, and FP Rate) result compared to online NB as in Figure 5. It can be said, machine learning algorithm is influenced factor where different domains might employed the different machine learning algorithms. This is because the nature data in different domains might varies.

In [8] measured the performances based on F1-score instead of accuracy for multi-classification which is very important when develop a network anomaly detection system. It also found that multi-classification more accurate prediction of anomalous data with only few instance, whereas in our work well performed for network anomaly detection of multi-classification the 257 673 data instances of UNSW-NB15 dataset. Therefore, their work does not show the effectiveness and efficiency of their NADS.

The features of UNSW-NB15 dependent to one another feature caused high accuracy for multi-classification with 83.47% by using online AODE algorithm due to the assumption of AODE where the feature is dependence by averaging all the estimation probabilities of the classifier that overcome the independence assumption of NB classifier (accuracy of online NB algorithm only equal to 69.60%) in computer network system for network anomaly detection.

Unsupervised learning, k-means algorithm, difficult to identifying the dynamic network data. As in [10] required to determine and set a suitable number of cluster to have a high classification rate and low false positive rate. Yet, their work conducted the experiment using 22 classes given 81.61% accurately classified the data instances. It is lower than this present work that employed online AODE algorithm with the difference of percentage of accuracy is 1.86%. From this, we can be said that online classifier (online AODE) is suitable for handling a dynamic network data instead of batch classifier that time-costly.

Although, the accuracy for overall of machine learning algorithm to determine the data either normal or abnormal drops for multi-classification of UNSW-NB15 dataset due to the imbalance data in given dataset (bias classification) but still the result is acceptable and consider high than online NB classifier. Another measure metrics that need to be enumerate are TPR and FPR. Both give best result for online AODE compare to online NB where the TPR value 77.84% and 6.57% FPR value for online AODE. Meanwhile, the online NB a little bit low TPR value produced (70.32%) and high FPR value very high with 31.67% which mean that it is worst-case for network anomaly detection the given dataset.

Differ from the work in [8] that state the imbalanced classes causes the most amount of data more accurate to be classified compared to the less amount of data which tend to be ignored for multi-classification (bias classification). Our work, online AODE algorithm, shows that almost all the present classes in the network system capable to be recognized and be learnt accurately. This can be proved when we employed online AODE algorithm for multi-classification the percentage of accuracy is more than 90% except for Exploits type which is equal to 89.81%. Yet, still consider high because it is almost to 90%.

To deal with the issues modern attacks dynamically in a computer network system. Another multi-classification of UNSW-NB15 dataset conducted that almost similar to our work by the authors in paper [7]. But, they used RepTree algorithm with the accuracy is 79.20% a little lower compared to our chosen algorithm AODE algorithm with the percentage of accuracy equal to 83.47%. This bring the significantly differences which is 4.27%.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Furthermore, the experiment also evaluates the performance of these two algorithms for two learnings methods (batch as well as online learning) based on the processing time required for network anomaly detection. Even though, the time taken of online AODE algorithm for network anomaly detection more than online NB where the difference for both algorithm is only about 1.18 seconds. Since, theoretically that time complexity of AODE algorithm is $O(cd^2)$ and NB is only $O(cd)$ [25], [26]. Where c is the number of class and d is a dimensionality of features. Hence, online AODE algorithm for network anomaly detection over the UNSW-NB15 dataset is comparable to the others algorithm.

The work by [7] faced a difficulty to network anomaly detection that required a frequent update, that may affect the consumption time for multi-classification. When the network data consists many classes (more than two classes), consequently the time taken to complete the classification and detecting the anomalous as well normal data is longer. Therefore, they conducted the network anomaly detection toward same dataset as in this work (UNSW-NB15 dataset) using different ML algorithms.

The finding from their work takes 2.69 and 0.37 seconds to training and testing using the RepTree algorithm for multi-classification [7]. To complete the multi-classification cost 3.06 seconds and take longer than this present work. Online AODE algorithm in this current work only take 1.27 seconds and it is time efficient for multi-classification over UNSW-NB15 dataset. Our finding proved that by built online algorithm speed up the classification.

Moreover, the built online AODE improve as well as fastest by taking a short time to classify the classes of network labelled UNSW-NB15 dataset. Batch AODE algorithm takes 7.69 seconds whereas online AODE take less than 2 seconds for multi-classification. Similarly, the NB algorithm proved that by using online learning the shorter time needed for classification. This result of processing time tabulated as in Table 3.

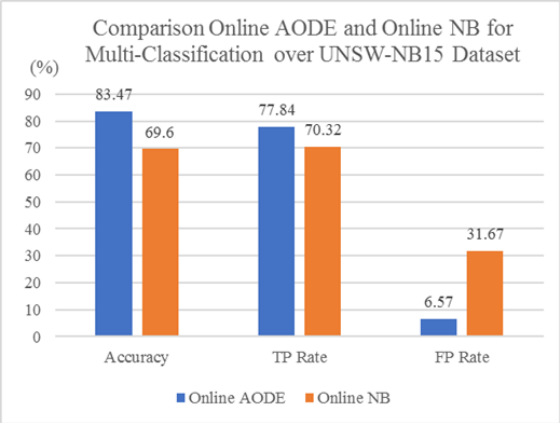

Figure 5: Comparison Online AODE and Online NB for Multi-Classification over UNSW-NB15 Dataset

Table 3: Time Taken for Multi-Classification of UNSW-NB15 Dataset

ML algorithm	Time Taken (seconds)
Online NB	0.09
Online AODE	**1.27**
Batch NB	1.20
Batch AODE	7.69

5. CONCLUSION AND FUTURE WORK

This project was undertaken to multi-classification of UNSW-NB15 dataset for network anomaly detection system using machine learning and evaluate the performance of online implementation of AODE algorithm. Online AODE algorithm is to keep update the features of UNSW-NB15 dataset where allowing the built model to adapt to the recent network attacks present in a computer network that passing over time. The following conclusions can be made: 1. Online AODE is high in accuracy for Worms attack class of multi-classification with 99.93% compare to the other classes. 2. Furthermore, online AODE is the high classification rate for multi-classification over UNSW-NB15 dataset compare to another Bayesian family (Naïve Bayes algorithm) where online AODE gives 83.47% meanwhile online NB produce 69.60% classification rate. In term of processing time for both algorithms, the online AODE comparable to online NB with only small different where it is only taken 1.27 seconds. 3. Also, the finding proved that online classifier is a fast algorithm compare to design a batch classifier for network anomaly detection. Online AODE algorithm outperformed for multi-classification of UNSW-NB15 for network anomaly detection. The future work regarding this work by build a

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

distributed online classifier using machine learning algorithm for network anomaly detection system that concerns scalability, centralization, and geographically location issues of the network data.

ACKNOWLEDGMENTS

The research reported in this paper is supported by Research Acculturation Grant Scheme (RAGS), Number Grant: 9018-00080. The authors would also like to express gratitude to the Malaysian Ministry of Higher Education (MOHE) and University of Malaysia Perlis (UniMAP) for the sponsor, financial support, and facilities provided.

REFRENCES:

[1] N. A. Hussein, "Design of a Network-Based Anomaly Detection System Using VFDT Algorithm," no. May, 2014.

[2] S. Ben-David and S. Shalev-Shwartz, *Understanding Machine Learning: From Theory to Algorithms*. 2014.

[3] D. Bhattacharyya and J. Kalita, *Network anomaly detection: A machine learning perspective*. 2013.

[4] M. Idhammad, K. Afdel, and M. Belouch, "Distributed Intrusion Detection System for Cloud Environments based on Data Mining techniques," *Procedia Comput. Sci.*, vol. 127, pp. 35–41, 2018.

[5] M. N. Chowdhury, K. Ferens, and M. Ferens, "Network Intrusion Detection Using Machine Learning," pp. 30–35, 2010.

[6] M. Idhammad, K. Afdel, and M. Belouch, "DoS Detection Method based on Artificial Neural Networks," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 4, 2017.

[7] M. Belouch, "A Two-Stage Classifier Approach using RepTree Algorithm for Network Intrusion Detection," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 6, pp. 389–394, 2017.

[8] F. Gumus, C. O. Sakar, Z. Erdem, and O. Kursun, "Online Naive Bayes classification for network intrusion detection," *ASONAM 2014 - Proc. 2014 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min.*, no. Asonam, pp. 670–674, 2014.

[9] A. Sultana and M. A. Jabbar, "Intelligent network intrusion detection system using data mining techniques," in *2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*, 2016, pp. 329–333.

[10] S. Duque and M. N. Bin Omar, "Using Data Mining Algorithms for Developing a Model for Intrusion Detection System (IDS)," *Procedia Comput. Sci.*, vol. 61, pp. 46–51, 2015.

[11] S. Zhao, M. Chandrashekar, Y. Lee, and D. Medhi, "Real-time network anomaly detection system using machine learning," *2015 11th Int. Conf. Des. Reliab. Commun. Networks*, no. October 2016, pp. 267–270, 2015.

[12] K. Limthong, "Real-Time Computer Network Anomaly Detection Using Machine Learning Techniques," *J. Adv. Comput. Networks*, no. March, pp. 1–5, 2013.

[13] L. Rettig, M. Khayati, P. Cudré-Mauroux, and M. Piórkowski, "Online Anomaly Detection over Big Data Streams."

[14] H. Volden, "Anomaly detection using machine learning techniques," 2016.

[15] S. Ahmad and S. Purdy, "Real-Time Anomaly Detection for Streaming Analytics," Jul. 2016.

[16] Q. Bai, H. Lam, and S. Sclaroff, "A Bayesian Framework for Online Classifier Ensemble," *J. Mach. Learn. Res.*, vol. 32, no. 2005, pp. 1–25, 2014.

[17] D. J. Hill, B. S. Minsker, and E. Amir, "Real-time Bayesian Anomaly Detection for Environmental Sensor Data."

[18] S. C. Tan, K. M. Ting, and T. F. Liu, "Fast anomaly detection for streaming data," *Proc. Twenty-Second Int. Jt. Conf. Artif. Intell. - Vol. Vol. Two*, pp. 1511–1516, 2011.

[19] "Weka 3 - Data Mining with Open Source Machine Learning Software in Java." [Online]. Available: https://www.cs.waikato.ac.nz/ml/weka/downloading.html. [Accessed: 10-Apr-2018].

[20] L. M. L. de Campos, R. C. L. de Oliveira, and M. Roisenberg, "Network Intrusion Detection System Using Data Mining," *{...} Appl. Neural {...}*, no. July, pp. 104–113, 2012.

[21] N. Moustafa and J. Slay, "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," *2015 Mil. Commun. Inf. Syst. Conf.*, pp. 1–6, 2015.

[22] N. Moustafa and J. Slay, "A hybrid feature selection for network intrusion detection systems: Central points A HYBRID

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

FEATURE SELECTION FOR NETWORK INTRUSION DETECTION SYSTEMS: CENTRAL POINTS AND ASSOCIATION RULES," pp. 5–13, 2015.

[23] R. Fontugne, P. Borgnat, P. Abry, and K. Fukuda, "MAWILab," *Proc. 6th Int. Conf. - Co-NEXT '10*, p. 1, 2010.

[24] Z. A. Baig, A. S. Shaheen, and R. Abdelaal, "An AODE-based intrusion detection system for computer networks," *World Congr. Internet Secur. WorldCIS-2011*, pp. 28–35, 2011.

[25] G. I. Webb, J. R. Boughton, and Z. Wang, "Not so naive Bayes: Aggregating one-dependence estimators," *Mach. Learn.*, vol. 58, no. 1, pp. 5–24, 2005.

[26] G. I. Webb, J. Boughton, and Z. Wang, "Averaged One-Dependence Estimators : Preliminary Results," no. 1, 2002.

Journal of Theoretical and Applied Information Technology
15th August 2018. Vol.96. No 15
© 2005 – ongoing JATIT & LLS
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 4: Confusion Matrix for Online AODE Algorithm

Class	Normal	Backdoor	Analysis	Fuzzers	Shellcode	Reconnaissance	Exploits	DoS	Worms	Generic
Normal	86 893	2	886	4 330	276	347	172	65	8	21
Backdoor	0	389	187	267	25	100	676	670	2	13
Analysis	15	218	606	250	0	36	827	718	0	7
Fuzzers	1 949	316	232	18 989	414	233	1 149	909	3	52
Shellcode	19	0	0	48	1 353	77	3	9	0	2
Reconnaissance	12	30	12	21	49	11 664	978	1 208	1	12
Exploits	155	327	531	1 149	406	3 155	29 635	8 728	120	319
DoS	40	176	170	399	205	625	7 161	7 428	11	138
Worms	1	0	0	2	3	9	17	1	138	3
Generic	12	14	12	74	58	28	386	296	5	57 986

Table 5: Confusion Matrix for Online NB Algorithm

	Normal	Backdoor	Analysis	Fuzzers	Shellcode	Reconnaissance	Exploits	DoS	Worms	Generic
Normal	63 551	4 017	200	16 537	215	381	3 535	2 971	222	1 371
Backdoor	3	1 007	0	59	16	14	195	992	43	0
Analysis	1	859	145	4	0	0	464	1 195	9	0
Fuzzers	13	2 647	7	15 267	286	122	339	4520	252	793
Shellcode	0	8	0	209	1 055	91	6	130	4	8
Reconnaissance	6	522	4	113	83	9999	237	1669	1353	1
Exploits	104	3 984	2	1 327	620	662	21116	12690	3985	12
DoS	41	2 987	17	236	239	88	2461	9797	483	4
Worms	0	5	0	1	6	1	6	10	145	0
Generic	4	68	0	97	98	6	750	433	151	57264