National College of Ireland

Improving the efficiency of an Intrusion Detection System using Random Forest and K-Means algorithms

MSc Internship
Msc. In CyberSecurity

Shashank Somachand Kattige
Student ID: X18138535

School of Computing
National College of Ireland

Supervisor: Mr. Ross Spelman

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Shashank Somachand Kattige

Student ID: X18138535

Programme: …MSc. In CyberSecurity…… **Year:** ……2019……..

Module: ……Academic Internship……

Supervisor: ……Ross Spelman……

Submission Due Date: …12/08/2019……

Project Title: Improving the efficiency of an Intrusion Detection System using Random Forest and K-Means algorithms

Word Count: ……5524…… **Page Count**……15……..

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ……

Date: ……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Title

Shashank Somachand Kattige
X18138535

Abstract

Intrusion Detection System (IDS) is a system that provides a layer of security to an organization's networks. In today's world, the number of devices that are getting connected and communicating with each other are increasing at an exponential rate. The convenience of connecting to each other has come at a cost of sacrificing the security aspect. Due to that the number of Blackhat hacker are also increasing who gain access to a network illegally. Due to this the number of cyber-attacks is also going up, with different types of techniques applied by them. Today having a firewall on the network is not enough, they cannot stop all the types of attacks coming from the external network. Intrusion Detection System plays an important role in obstructing these attacks at the entry of the network itself. This research paper talks about the new model of the classifier for the Intrusion Detection System. Two familiar classifiers, Random Forest and k-means clustering are used to develop the proposed model. The new technique increases the performance, accuracy and detection rate of the Intrusion Detection System. Every machine learning algorithm have their own advantages and disadvantages. NSL-KDD dataset has been used to train the proposed classifier model. Both Random Forest algorithm and K-means clustering algorithm are quite efficient in classifying the traffic data as normal or malicious when compared to others.

Keywords: Intrusion Detection System, Ensembling Algorithms, and Random forest classifier, K-Means clustering.

1 Introduction

A system which can be either hardware or software monitors a network or system to look for malicious or harmful behavior on a network is called as Intrusion Detection System (IDS). The harmful or malicious activity is then reported to a centralized system or an administrator. Generally, IDS is divided into six types based on how they function, behave and respond. They are active, passive, Network-based, host-based, knowledge-based and behavior-based IDS. A Passive Intrusion Detection System is an IDS that will monitor and log the network traffic. It will raise an alert when it detects any suspicious activities carried out on the network. They can notify a particular entity through many means such as email, text message, message on the monitor of the higher authority or even get pop up notification on the screen. It will also check for the vulnerabilities. But that is all it can do as far as the capability goes. Most of the IDS are passive in today's world. There is much need for a system that can react immediately when there is any incident on the network. This is where Active IDS comes into the picture. An Active Intrusion Detection System is a hardware or software detection system that can, not only monitor the network traffic and report the higher for suspicious activities,

but it can also for take necessary steps to stop or block the attacks. For example, if the active IDS detect any suspicious activity on the network, then it will have the ability to modify the rules on the firewall to stop the network traffic flow coming from the system that is carrying out the harmful activities. It can also have different abilities such as diverting that traffic to a honeypot or can close the processes itself on the suspicious systems. It is also called as Intrusion Prevention System (IPS). - blogs.getcertifiedgetahead.com (2019)

Network-based Intrusion Detection System (NIDS) is a type of IDS which will have a sensor or a network device which will have a Network Interface Card (NIC) in a wide range detection mode with separate Interface to manage the traffic of the network. NIDS detects the network intrusions using data mining system. All the traffic logs are collected at one point in the network and then filtered. The filtration is applied to separate the known traffic, coming from the trusted sources and unknown traffic for further analysis to determine if the traffic is safe or not. The author Raghunath, B (2008) has observed that it will also pick up traffic using network monitoring tools such as packet sniffer. Here you can also implement many rules which can specify what type of packets can be allowed on the network or which protocol can be used and which of the should be blocked. When compared to other types of IDS, NIDS has a much better and wider ability for monitoring traffic. NIDS is typically will be placed at the edge of the network. Preferably at the starting point of the network.

Host-based Intrusion Detection System are the type of IDS will analyze the events on each host rather than monitoring the network. In this the HIDS must be installed on every host system on the network to monitor all the events of hosts. HIDS will investigate the log files and admin files of the host systems and will send those files to be stored at a centralized location. The one advantage of this is, if the system gets corrupted by a malicious software or some other issues. Then you will be able to restore the host system back to its previous normal state using these stored files. HIDS can only alert if there is any malicious activity on the system but it won be able to stop any kind of attacks. According to the author Raghunath, B (2008) the few disadvantages when using HIDS are if the system is compromised then the HIDS can be disabled by a skilled attacker. The other thing to lookout for when deciding whether to go for NIDS or HIDS is the network size of the company size. If the number of host systems are very high, then it is better to go for NIDS rather than HIDS. The reason for it is that it is very difficult to maintain HIDS for those many systems and it will be very difficult to analyze the log files if the intrusion attacks occurs in multiple systems at the same time.

The next one is Signature-based Intrusion Detection System or also known as Knowledge-based IDS will detect for the vulnerabilities or anomalies in the network by continuously comparing the traffic activities with the database containing the information on previous attacks. To be more precise it will check the database of the signatures. Signatures is nothing but the pattern left by the intruder when attacking the network or the systems in it. The database will store all those patterns and will help the Signature-based IDS to recognize the similar patterns on its network. The only disadvantage of this IDS is, the database needs to be up to date at every instance. If not, the IDS will not be able to detect the latest attacks carried out by the attackers. - Omnisecu.com. (2019)

And lastly, the other type of IDS is behavior-based IDS which is also know by another name called Anomaly-based Intrusion Detection System. Here a baseline or normal status is established by observing and recording the activities of a normal host system on the network. If there is any activity that occurs outside these normal patterns, then the IDS will detect it

and will send the alarm notification to the higher authority. This IDS has the ability to change itself by learning new types of intrusion attacks on the network. The flaw of this IDS is that it has high False Alarm Rate (FAR) when compared to other IDS. - Omnisecu.com. (2019)

Data mining has become one of the most popular techniques used in IDS to make improvements. The author Brownlee, J. (2019) has stated that this was introduced to overcome the limitations of the rule-based systems. There are many popular machine learning algorithms that are getting implemented in the IDS to make improvement in them. Algorithms such as Support Vector Machine, Random Forrest, C4.5, Naïve Bayes, Artificial Neural Networks, K-means, Apriori and many more. But these algorithms are further divided into three categories they are supervised, unsupervised and semi-supervised algorithms. Most of the algorithms that are present comes under supervised algorithms. It is a type of machine learning algorithm that first needs to be trained using a dataset so that it can start making predictions and learn what is right and what is wrong. Even though we know the correct answer, this algorithm will make the predictions on the training dataset. To check if its predictions are correct or not, the algorithm will be feed test dataset to check the predictions. This process is stopped once the desired level of the performance is achieved. We have chosen classification technique for the proposed model and Random Forest as one of the machine learning algorithms. Random Forest (RF) is an ensemble which will produce decision tree using feature bagging and arbitrary selection techniques. Other reason for choosing this algorithm is, it has high accuracy when compared to others and will out the result with low amount of errors. Due to this RF has found popularity among the developers of the IDS to resolve many issues present in it using large dataset.

According to Brownlee, J. (2019) the other category of the algorithm is non-supervised machine learning algorithm. In this, the model written in unsupervised algorithm will be given a dataset and then will be instructed to give the output. This output can be either right or wrong, no instructions will be provided to it. This algorithm without using any labelling or classification process will start grouping the data using neural networks. This is further classified into two types, clustering and association. We have chosen k-means clustering algorithm as the other machine learning algorithm to construct our model. The reason for this choice is, this algorithm can adapt to any new type of data very quickly and easily and the other reason is, it accepts large datasets. The other reason to use this algorithm is that it identifies and interprets the hidden structures of an unlabeled data.

This research focuses on observing the normal or abnormal behavior of the network traffic in a dynamic environment. Analyze that traffic and measure the performance of the proposed hybrid model that has been achieved by the combination of Random forest and K-means algorithm. Then compare the performance with the results of the individual algorithms. For that we have used NSL-KDD dataset to train and test the hybrid model. The reason for selecting this dataset is, it contains large amount of data approximately 4,900,00 single connection vectors with 41 attributes describing different types of connections and another label assigned to each of them telling that if that connection is attack type or normal.

2 Related Work

2.1 Intrusion Detection System (IDS)

Intrusion Detection is nothing but a process that manages the security of the network or system and will look for any malicious activities. According to the article Deng, S (2017) in

today's world, with increase in the number of hackers, the techniques used by them are also changing very rapidly to achieve their goal on the remote network. Malicious programs such as Petya or WannaCry Ransomware and many more are born every day. Different strains of the same Malicious software are being developed. To achieve protection against these kinds of actors, an Intrusion Detection System must be built, which can not only detect these types of attacks. But also keep updating itself to learn different techniques to avoid getting attacked by the hackers. According to the article written by Enache, A (2017) to achieve that many researchers are trying to build the system using different approach, but one of the obstacles that must be tackled is dealing with large amount of data that flows through the network traffic. This traffic will contain noise and other junk traffic which has a negative effect on the intrusion detection. They tried to overcome this issue by using swarm intelligence algorithms with feature selection method. But they were not able to get the result what they had desired.

Machine Learning is the study and construction of a meaningful algorithm based on the datasets provided. These algorithms will work by constructing a model that will analyze these data and built a pattern that will make the predictions and decision-making qualities better. This makes Machine Learning algorithms suitable for the Intrusion Detection purposes. This will also help to achieve different solutions to protect itself against different types of attacks. CART (Classification and Regression Trees) uses different machine learning techniques for developing different prediction models based on the dataset provided.

2.2 Machine Learning algorithms

In the research paper written by Choudhury, S. (2015) Intrusion Detection using Data Mining Techniques (IDDM) is one of the type of IDS that detects the malicious attacks on the network by using meta rules, characteristic rules and association rules. In this research paper the author makes the comparative analysis between different types of classifiers such as Bayes classifier, Function classifier, Lazy classifier, Meta classifier, BayesNet, Trees, Random Tree and much more. They have calculated the performance metrics like sensitivity, specificity, precision, accuracy and many more. These metrics were calculated by measuring the four value which are, True Positive (TP), False Negative (FN), False Positive (FP) and True Negative (TN). After calculating all the metrics of all the classifiers, the author has compared all those metrics with each other. After comparison the author has observed that Random forests algorithm is efficient and accurate when it comes to dealing with large amount of data like network traffic. Using this algorithm will help us overcome any imbalances in the network by doing feature selection. This in-turn helps us in detecting new types of attacks. Here it was observed an individual classifier must be different from each other while making improvements to the IDS. If that is not the case the you will not be able to achieve any improvements compared to when individual classifiers are used. All the analysis was carried out on Waikato Environment for Knowledge Analysis (Weka) tool, which is a suite of the machine learning algorithms. In this tool data analysis and predicting modeling was carried out.

Here, Farnaaz, N. (2016) has discussed about the method of using Random Forest algorithm in an IDS. This was the initial research paper that I had gone through to see how it was implemented. According to this paper, the dataset was loaded into the system. Next that a pre-processing technique was applied to set the variables into sorting the types of attacks. After that the main data set was divided into four separate datasets. Then the dataset was partitioned into two parts, training dataset and test dataset. After that the dataset is fed to Random Forest algorithm where it gets trained and tested. After doing that performance

parameters where measured such as accuracy, False Alarm Rate (FAR) and Mathews correlation coefficient (MCC). This in-turn was used to detect four types of attack at that time. These attacks were DOS, probe, I2R and R2L. Here the classification technique was used to achieve this result. They had applied symmetrical uncertainty of attributes which overcame the problem of the information gain. NSL-KDD dataset was used to train the model. Irrelevant features were removed by applying feature selection technique. They have also proved that the accuracy and detection rate of the IDS increases in detecting those four types of attacks. This gave me a good overview idea of how Random Forest algorithm is used to achieve the result.

Next, I wanted to see how k-means algorithm was used in the IDS to improve its performance. For that, I referred the research paper written by Kumar, V. (2013). The author has proved that, using k-means clustering algorithm in IDS was very useful when it came to the job of dealing with large amount of unlabeled dataset. In this, NSL-KDD dataset was used for the analysis. Here, k-means clustering algorithm was able to detect different types of attacks present in the dataset very efficiently. As we know that k-means clustering algorithm is an unsupervised algorithm, the unlabeled data in the dataset will be classified using the clustering technique. The author has conducted this experiment using WEKA tool. After applying the clustering technique, the author divides the clusters into four different groups. The author has also mentioned that the K-Means clustering has three drawbacks which are class dominance problem, force assignment problem and no class problem. But it was also mentioned in the later part that, much more optimal results can be achieved if he was able to use a hybrid model.

Muda, Z. (2011) proposed a hybrid model of the IDS that uses K-Means clustering and Naïve Bayes classification algorithms to achieve higher level of security than the ordinary IDS. Here they have built a model which has two stages. In the first stage K-Means clustering algorithm will divide the data into two groups based on the behavior of the data. If the data is behaving similarly then they are grouped as normal traffic and if dissimilarly then they will be grouped as malicious data. But K-Means clustering algorithm was not able to distinguish between normal and intrusion instances. To overcome these shortcomings K-Means clustering was combined with Naïve Bayes algorithm in the second stage to classify the data further into the correct class of categories. In this stage first the conditional probability was calculated for the clusters created by the K-Means algorithm. Then after calculating the probabilities, the average to all the probabilities is taken at the end to get the overall probability of the model. In this way the optimal performance and detection rate was achieved. The dataset that was used for this experiment was KDD Cup 99. Even though they had achieved what they had proposed, the author has acknowledged that the model has certain certain flaws such as the performance of the Naïve Bayes classifier was very low, and the classification of the attacks was not optimal due to grouping of the similar attack types. [10]

2.3 Feature Selection

Sometimes the classification process gets disrupted due to the multiple classification fields. These fields can consist of sudo correlations that can construct the intrusion detection process. Sometimes certain gets added to other classifications by mistake which can also taint the intrusion detection function of the IDS. To eliminate all these problems, feature selection techniques are applied on the traffic data present in the dataset. This will help the IDS to perform better as the technique will remove any unnecessary features present and will only select the important ones. Doing this the performance of an IDS will be increased. The

features are dependent on the type of the IDS that is being implemented. NSL-KDD contains total of 42 features in it. These features are further divided to make it easy for the IDS to decide which features are required and which are not. Chae, H. (2014). has made an in-depth analysis on how feature selection is carried out using in the IDS on NSL-KDD dataset. Here WEKA 3.7 machine learning tool was used to calculate the attributes such as Attribute Ratio (AR), Class Ratio (CR) and Accuracy was calculated to see which features needs to be selected. In this research also data mining techniques are used for the selection of the features.

2.4 Dataset

Here in this research the data NSL-KDD has been used to train the selected machine learning algorithms. It is a refined version of KDD CUP 99 dataset. According to the author Dhanabal, L. (2015) many researchers who had worked on KDD CUP 99 dataset had come to the conclusion that there we some serious flaws in it. This invalidated the results that they had achieved. The main flaw in this dataset was that all the malicious packets had TTL port numbers 126 or 253 assigned to them, whereas all the benign packets had TTL port with number 127 or 254 assigned. This showed bias toward a particular traffic. Due to that NSL-KDD dataset was introduced to resolve the issues. These issues were resolved by doing the following things, redundant records were removed which enabled the classifiers to produce an un-biased result. The number of train and test data was also increased to get more optimal results. Hence, I chose to go for this dataset.

3 Research Methodology

I have proposed a hybrid model of the Intrusion Detection System that uses both, Random Forest algorithm and K-means clustering algorithm. The purpose of this model is to increase the overall performance and efficiency of the IDS. To achieve that these are the following procedures that has been followed, they are:

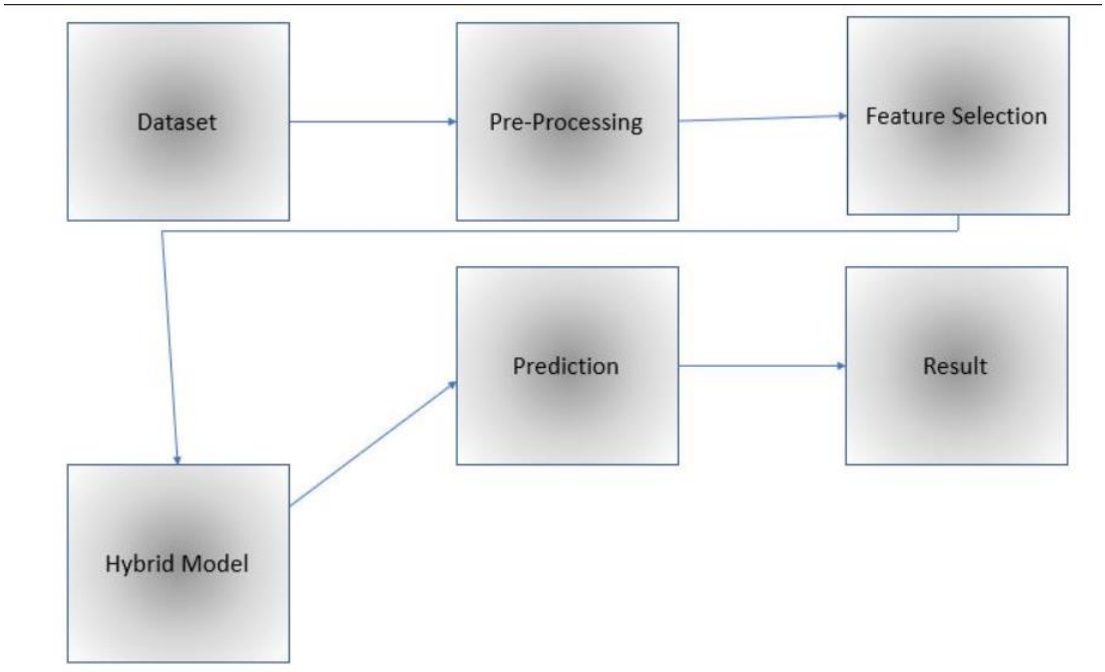


Figure 1. Flow Diagram of the Hybrid Model

3.1 Loading and Extraction of Dataset

The dataset NSL-KDD is a predefined dataset that will consist of the traffic data. This dataset is divided into two parts which are training and testing datasets. It is in a '.csv' format. It has been downloaded from the official website of Canadian Institute for Cybersecurity. The training dataset will be used to train both the machine learning algorithms i.e. Random Forest algorithm and K-Means clustering algorithm. Then the testing dataset will be used to check the performance of the RF algorithm.

3.2 Encoding or pre-processing:

The next process that happens is the encoding of the data that has been loaded by the dataset. Here we are using Label Encoder to convert nominal variables to integer variables scaling between 0 and 1.

3.3 Feature Selection PCA

After the encoding process, the feature selection process is carried out. This is done to further refine the huge data of the network traffic present in NSL-KDD dataset. This is done by using PCA (Principle Component Analysis) technique. The result will be the reduced number of features that is required by this particular model. All other irrelevant features will be dropped.

3.4 Prediction

After all these processes the testing data will be fed to Model again. This will invoke the prediction process after the training of the algorithms.

3.5 Results of the model

And after going through all these processes, the model will calculate the performance values such as accuracy, detection rate and false alarm rate. Which will be show in-terms of tabular and graphical representation form.

4 Design Specification

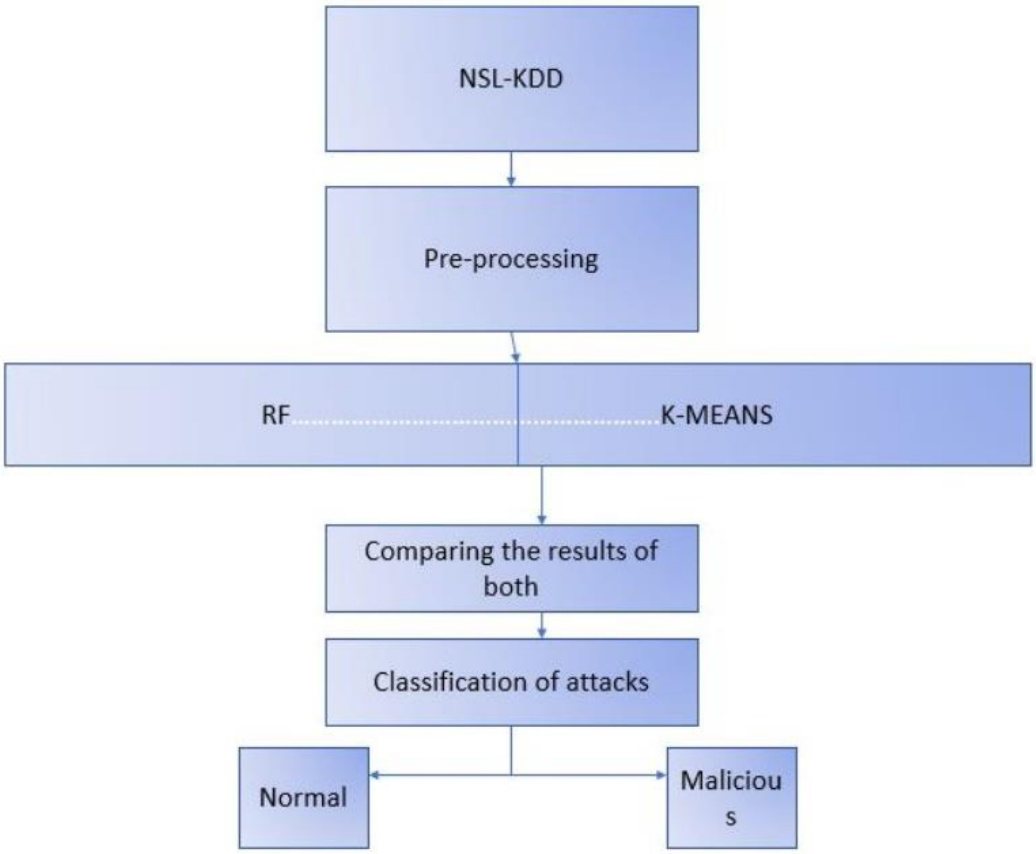


Figure 2. Proposed Framework of the Hybrid Model

As we can see from the above proposed framework, first the dataset is introduced into the system. This dataset is NSL-KDD which is a preloaded dataset. After loading the dataset, the dataset is divided into train dataset and test dataset. After going through the process of training, the dataset will be put under pre-processing procedure where the nominal variables will be converted to numerical value. That is the alphanumerical values will be converted to 1 and 0. After that the dataset is fed to both, Random Forest algorithm and K-Means Clustering algorithm. The output is then compared with each other and then the classification of the attack types will be done. At the end of the process the data will be classified into two categories “Normal” and “Attack”.

5 Implementation

Dataset Loading

In this process, the dataset NSL-KDD which has been taken by us for the procedure will be loaded. To process this dataset, we are using SciKit-learn python framework with modules sklearn, numpy, pandas and other additional libraries. This dataset is divided into two parts, train dataset with the filename “KDDTest+.csv” and test dataset with the filename “KDDTest+.csv”. The train dataset is 80% part of the main NSL-KDD dataset and test dataset is 20%.

Feature Selection and preprocessing

This is a very important procedure that needs to be carried out to increase the efficiency of the machine learning algorithms. This is done to decrease the dimensionality and removing of any irrelevant features that is not required for the IDS. There are 42 features in total present in NSL-KDD database. This dataset is mainly categorised into three types They are Numeric, Nominal and Binary. The nominal and binary features are represented by the numbers and the remaining feature is numerical. These features can be seen in the table below

Types	Features
Nominal Feature	Protocol type, Service Flag
Binary Features	Land, logged in, root shell, su attempted, is host login, is guest login
Numeric Features	Duration, src bytes, dst bytes, wrong fragment, urgent, hot, num failed logins, num comprimised, num root, num _le creations, num shells, num access _les, num outbounds cmds, count, srv count, ser-ror rate, srv serror rate, same srv rate, rer-ror rate, srv rerror rate, same srv rate, di_ srv rate, srv di_ host rate, dst host count, dst host srv count, dst host same srv rate, dst host di_ srv rate, dst host same src port rate dst host srv di_ host rate, dst host serror rate, dst host srv serror rate, dst host rerror rate. Dst host srv rerror rate

This this process is carried out in the file with the name “preeprocess.py”. In this itself it will first read the train and test datasets that are present in the form of .csv file. After the dataset is loaded, encoding process takes place using LabelEncoder module of sklearn.preprocessing library. All the attributes present in the dataset is normalized by assigning them the values between 0 and 1. The data is then grouped into “normal” and “attack”.

Hybrid Model Execution

The hybrid model has been executed in the file “rf_kmeans.py” in which the training and testing dataset are loaded. After which the accuracy, sensitivity and specification values are calculated from the test data and will be sent to the “main.py” file for the printing of the results. Both Random Forest classifier and K-Means clustering are executed. In this the RF classifier first gets trained by taking the data from the KDDTrain+.csv file and tests its performance using the file “KDDTest+.csv” file. The estimator has been set as 1000 which is good enough.

As K-Means clustering algorithm is an unsupervised algorithm, it will only train itself by taking the file “KDDTrain+.csv” file. Doing this the K-Means algorithm will form two groups from the cluster data named as “normal” and “attack”.

After the process of classification and clustering from both the algorithms, the results are compared with each other. The final result will be considered as “normal” if both the result from the algorithms matches as “normal” or else the final output will be considered as “attack”. In this way the proposed hybrid model has been proposed.

Output
The final output results will be calculated and printed out in the form of table and a graphical form. The final outputs of accuracy, detection rate, false alarm rate, sensitivity and specification is compared with the individual outputs of the algorithms and are plotted out.

6 Evaluation

6.1 Performance output table

- Accuracy of the both individual algorithms and our proposed model has been calculated. This is measured by obtaining certain parameters such as True Positive (TP) and True Negative (TN) where the attacks and normal connections are correctly identified. False Positive (FP) and False Negative (FN) where the attacks and normal connections are incorrectly identified. The precision with which the classifier identifies as an attack or normal is considered as accuracy of the IDS. This can be calculated by using the formula:

  $$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

  And, as we can observe from the below table, the value of accuracy is higher for the hybrid model when compared to the accuracy values of the individual algorithms. The value of accuracy is 0.854678 which is higher than the value of accuracy of Random Forest algorithm which is 0.825134 and the accuracy value of K-Means clustering algorithm which is 0.774083.

- The detection rate can be calculated through the following formula:

  $$\text{Detection Rate} = (TP) / (TP + FN)$$

  As we can observe from the table below, the detection rate of the hybrid is slightly below the detection rates of the two individual algorithms.

- Next is the observation of the False Alarm Rate (FAR), to calculate FAR this the formula:

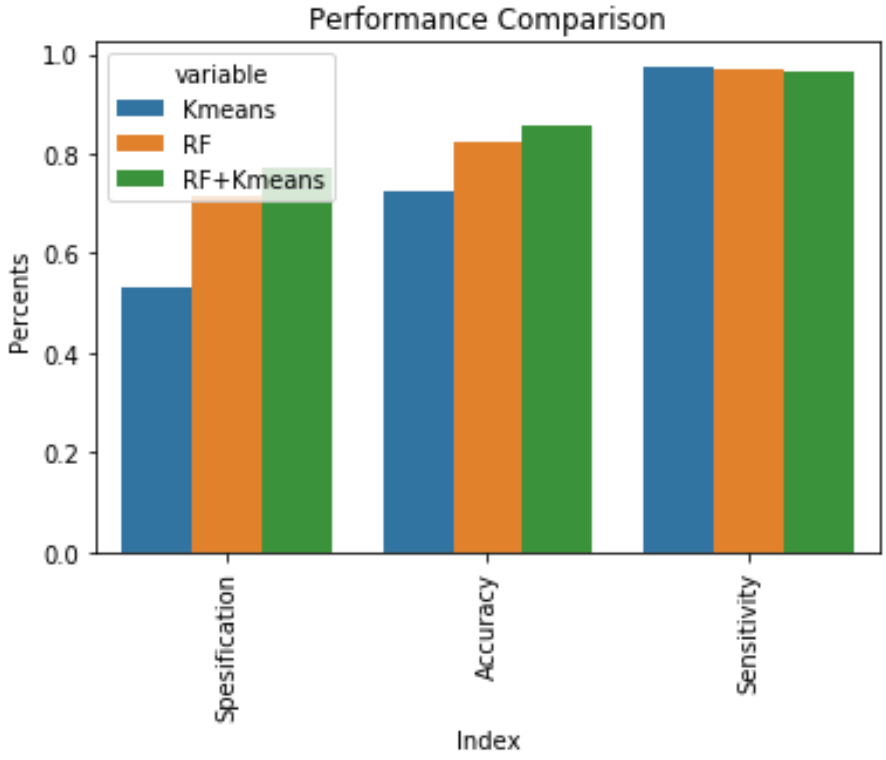
  $$\text{FAR} = (FP) / (FP + TN)$$

  And we can observe that the FAR value of the hybrid model is better than the FAR values of the individual algorithm.

Method	Accuracy	Detection Rate	FAR
RF	0.825134	0.971473	0.714408
Kmeans	0.724083	0.975798	0.533624
RF+Kmeans	0.854678	0.966838	0.769812

6.2 Performance Output 2:

From the below graph we can observe that the accuracy and specification values are much higher for the proposed hybrid model when compared to the outputs of the other two individual algorithms. Where as we can see that the value of the sensitivity is slightly lower for the hybrid model when compared to the individual algorithms.



6.3 Discussion

Here we have proposed a hybrid model of Random Forest algorithm and K-Means algorithm, where the model trained and tested itself under the dataset NSL-KDD. We have also trained the individual algorithms under the same dataset and then compared their performance results. We have observed that the above proposed hybrid model of Random Forest algorithm and K-Means algorithm performs better when it comes to accuracy, specification and False Alarm Rates, the hybrid model performs better than the individual algorithms. Whereas, Random Forest algorithm performs better individually when it comes to sensitivity and

detection rate. That is one of the drawbacks of the hybrid model. Looking from the results we can say that some improvements needs to be made in the future to overcome the before mentioned problem.

7 Conclusion and Future Work

The research was on the topic of increasing the efficiency and performance of an Intrusion Detection System by implementing the hybrid model of Random Forest algorithm and K-Means algorithm. From the observation of the result that we can conclude that the proposed hybrid model will have improve the efficiency and performance to a certain extent with few drawbacks. For the follow up research, we can attempt to make more improvements by making the hybrid model of with other algorithms and using multiple datasets to train and test the machine learning algorithms.

References

Get Certified Get Ahead. (2019). *Active VS Passive IDS Responses | Get Certified Get Ahead*. [online] Available at: https://blogs.getcertifiedgetahead.com/active-vs-passive-ids-responses/ [Accessed 12 Aug. 2019].

Raghunath, B. and Mahadeo, S. (2008). Network Intrusion Detection System (NIDS). *2008 First International Conference on Emerging Trends in Engineering and Technology*.

Omnisecu.com. (2019). *Types of Intrusion Detection Systems (IDS)*. [online] Available at: http://www.omnisecu.com/security/infrastructure-and-email-security/types-of-intrusion-detection-systems.php [Accessed 12 Aug. 2019].

Brownlee, J. (2019). *Supervised and Unsupervised Machine Learning Algorithms*. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/ [Accessed 12 Aug. 2019].

Deng, S., Zhou, A., Yue, D., Hu, B. and Zhu, L. (2017). Distributed intrusion detection based on hybrid gene expression programming and cloud computing in a cyber physical power system. *IET Control Theory & Applications*, 11(11), pp.1822-1829.

Enache, A., Sgarciu, V. and Togan, M. (2017). Comparative Study on Feature Selection Methods Rooted in Swarm Intelligence for Intrusion Detection. *2017 21st International Conference on Control Systems and Computer Science (CSCS)*.

CHOUDHURY, S. and Bhowal, A. (2015). Comparative analysis of machine learning algorithms along with classifiers for network intrusion detection. *2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM)*.

Farnaaz, N. and Jabbar, M. (2016). Random Forest Modeling for Network Intrusion Detection System. *Procedia Computer Science*, 89, pp.213-217.

Kumar, V. (2013). K-Means Clustering Approach to Analyze NSL-KDD Intrusion Detection Dataset. (2231-2307).

Muda, Z., Yassin, W., Sulaiman, M. and Udzir, N. (2011). Intrusion detection based on K-Means clustering and Naïve Bayes classification. *2011 7th International Conference on Information Technology in Asia*.

Chae, H. (2014). Feature Selection for Intrusion Detection using NSL-KDD.

Dhanabal, L. (2015). A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms.