Network Intrusion Detection with Machine Learning Approaches

Md. Ashraful Islam
2017-1-60-109@std.ewubd.edu
Computer Science and Engineering
East West University
Dhaka, Bangladesh

Sabbir Hossion
2017-1-60-112@std.ewubd.edu
Computer Science and Engineering
East West University
Dhaka, Bangladesh

Ashikur Rahman
2017-1-60-081@std.ewubd.edu
Computer Science and Engineering
East West University
Dhaka,Bangladesh

Biplob Sutra Dhar
2017-2-60-121@std.ewubd.edu
Computer Science and Engineering
East West University
Dhaka,Bangladesh

***Abstract*—In these modern days our need for a networking system is almost comparable to daily necessities such as food, water as it is practically impossible to go by without being a part of this vast and complex networking system. But in our struggles we have this system as a blessing and at the same time as a curse because it also opens the possibility for the intruders to intrude in your personal space. Therefore we need to have defense against it. Through our project we tried to contribute to our struggle against intrusion by creating an analysis choosing Random forest, Naïve Bayes and Decision Tree algorithms and its performance on a KDD19 attack Dataset. After normalizing the data we have implemented all these algorithms on this dataset we found that Decision tree (81.534 percent) performs better then Random forest (80.45 percent) and Naïve Bayes (76.13 percent) has less performance then all of them. After all these efforts we realized that our chosen algorithm is not suited for intrusion detection but other experts found Random Forest to be a very effective tool for intrusion detection yet we have found it a little less accurate then Decision tree.**

I. Introduction

In this progressive time of technologies we all are able to connect to each other through a complex system of networks. The number of devices is increasing and so the nature of our network is getting even more complex but the need and popularity of such systems is very much increasing as well. Recent researchers have been doing a lot of experiments Various techniques of data mining and artificial intelligence. The problem of attack identification can be reduced to any data Classification data mining work[1]. As this system takes us to a different future, people have started learning about this system and all its possibilities. The networking system we have created is not without its flaws and a group of people is always there to exploit its vulnerabilities. Intrusion-detection is created to detect attacks against computer systems and networks and in more generalized words information systems. The project we are developing is aimed at solving such issues by using machine learning algorithms. It is very simple to understand that to prevent and defend against such attacks is more than necessary if we truly believe in its impact and necessity for our future. Inspect an intrusion detection system (IDS) Activity in a system for suspicious behavior or Patterns that may indicate system attack and abuse. There There are two main categories of penetration detection Strategy; Extraordinary detection and abuse detection [2].

A. Objective:

The sole objective of our project is to work with the Datasets which are related to various kinds of attack, analyse them by implementing machine learning algorithms to detect and predict them and to mainly analyze which algorithms work better and why..

B. Motivation:

As the number of devices around the globe are increasing drastically, it causes massive dependencies on networking technologies. Most of our industries, company's factories, educational institutes, medical services, E-commerce's, Banks and many other economic zones are highly dependent on network services. Now all these major economic hubs have fallen under these attacks at least once in their service life and so it has been a war between the defenders of cyber-attacks and the attackers and our will to do better for our progressive nation and be a part of its growth has been the driving force of our project.

C. Related Work:

The thought of having a defense against attacks is not particularly a new topic and many others before us have

developed various methods to defend against such attacks. In general, an intrusion detection system (IDS) is a device that tries to create an alert Distinguish between contaminated or normal traffic by observing traffic Of Internet-connected devices[3]. Intrusion detection systems can help network users identify malicious motives no compromises on host and network security[4]. In extraordinary identification, the intrusions of the novel are detected Outlet detection process of random forest algorithm. After creating patterns of network services by random forest The algorithm is determined by the pattern-related outliers Outlet detection algorithm[5]. We evaluate on our approach Knowledge Discovery and Data Mining 1999 (KDD'99) dataset. Experimental results prove that the performance provided by the proposed abuse method is better than the best KDD'99 results; Compared to other unhealthy reported Moving on to identification, acquires our extraordinary identification method High detection rate if false positive rate is low; And The presented hybrid can improve the overall performance of the system the aforesaid IDS's[5]. Unexpected growth of network-based services and Sensitive information on networks, getting network protection More important than ever.

D. Importance:

Attacks on our networking system have been a key issue of the past decade yet there is not a single sign to get rid of such attacks. Along with technologies various attack methods have been created and used from time to time. It is a fair possibility that will continue till the unforeseeable future. Therefore to keep up with such stubborn problems we also need to take action to prevent it. If we fail to do so most of the networking system will be a playhouse for all kinds of cyber-attacks and intrusions as a result the consequences will be immense as it can be from a regular heist to catastrophic political unrest. So the necessity of our project is more than the words we have to explain its importance as it could be a trigger for huge technological fall out which will leave the damage from government burros to a regular civilian.

II. Methodology:

There are two categories of data collection methods such as secondary and primary data collection methods. For this project, the secondary methods of data collection are used.Since it deals with various journal articles, publications, websites, books, internal records, etc. and this research target was to find the optimal method for network intrusion detection[6]. Qualitative record keeping approaches are followed for the study. This method uses existing reliable documents and data sources as sources of information. This information can be used in new research. It's like going to the library. Any book and other reference material may be used in the study to collect relevant data. This study tried to complement some previous research using validity and reliability concepts. These concepts are used to justify the quality of this research. They indicated how well a technique or method can measure something. Validity is about the accuracy of a measure and reliability is about the consistency of a measure. In this study three machine learning algorithms Decision Tree, Naïve Bayes and Random Forest are used to determine the network anomaly. Whereas the study aims to find the best one according to the performance of those algorithms. Qualitative recording keeping approaches lead us to choose these algorithms for our study.

A.

Decision tree is a diagram used by decision-makers to determine the action process or display statistical probability. It provides a practical and straightforward way for people to understand the potential choices of decision-making and the range of possible outcomes based on a series of problems. Decision trees usually start with a single node and then decompose into additional nodes to show more possibilities (such as choosing the two sides of a coin). The farthest branch on the tree represents the final result.

B.

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. There is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.

C.

Random forests is a supervised learning algorithm. It can be used both for classification and regression. It is also the most flexible and easy to use algorithm. A forest consists of trees. It is said that the more trees it has, the more robust a forest is. Random forests create decision trees on randomly selected data samples, get predictions from each tree and select the best solution by means of voting. It also provides a pretty good indicator of the feature importance.

III. Implementation

At first we took the decision tree algorithm to train it with the dataset. NSL-KDD train dataset first implemented for training phase where 10 fold cross validation is also used as there is the existence of too much data. After completing the training phase we step forward to the test. So we used the NSL-KDD test dataset and took the result for analysis. The following process is also used for Naive Bayes and Random Forest algorithms.

A. Data Collection: Qualitative

Research is a process that aims at real-life investigations Understand social events. It focuses on "why" and so on The "key" is more than just a social event "key" Depends on people's direct experience money-making agents in their daily lives[7].Primarily we have chosen the kdd99 cup dataset as an experimental dataset. As the data file serves a huge amount of uncertain data we need to make it categorical forms. So we use the NSL-KDD dataset.The number of records on the NSL-KDD train and test set is reasonable. This feature makes it affordable to run tests on the entire set without the need to randomly select a small part. As a result, the results of evaluations of different research work will be consistent and comparable[8]..

B. Model Development:

WEKA 3 software software was used to classify the data. This is open source software issued under the GNU General Public License. WEKA is made by a bunch of machine learning algorithms and works for data mining tasks. A computer with core i3-7100 CPU @3.9GHz and 12 GB of RAM was used to give the data to classifiers.

C. Results:

Experimental results are shown in a table based on performance metrics values. It is the easiest way to compare the performance of those algorithms.

TABLE I. FALSE POSITIVE RATE VALUES OF CLASSIFIERS

	Decision Tree	Naïve Bayes	Random Forest
Normal	0.304	0.367	0.367
Anomaly	0.027	0.323	0.027

TABLE II. PRECISION VALUES OF CLASSIFIERS

	Decision Tree	Naïve Bayes	Random Forest
Normal	0.708	0.657	0.695
Anomaly	0.971	0.924	0.971

TABLE III. RECALL VALUES OF CLASSIFIERS

	Decision Tree	Naïve Bayes	Random Forest
Normal	0.973	0.931	0.973
Anomaly	0.696	0.633	0.677

TABLE IV. F-MEASURE VALUES OF CLASSIFIERS

	Decision Tree	Naïve Bayes	Random Forest
Normal	0.819	0.771	0.811
Anomaly	0.811	0.751	0.798

TABLE V. MEAN ACCURACY VALUES OF CLASSIFIERS AND

	Decision Tree	Naïve Bayes	Random
Average Accuracy (percent)	81.5339	76.1222	80.4
Classification Time(seconds)	0.18 seconds		0.81 se

IV. Conclusion:

Detecting an intrusion has been a well-known topic of our time. There has been a lot of work with this topic and the development of few other platforms has helped greatly and made it friendlier to work with this topic. We took a few Machine learning algorithms such as Random forest, Naïve Bayes and Decision tree and after applying it on our datasets we divided our datasets into training and test dataset and run it through WEKA to receive our results. We found that the Decision tree has better accuracy than Random forest and Naïve Bayes. However Random forest (80.45 percent) comes very close to Decision tree (81.534 percent) while Naïve Bayes has the least accuracy of 76.13 percent. Therefore it makes Naïve Bayes less effective for intrusion detection. It is very concerning that other researchers have found Random Forest the best option for intrusion detection but it is quite clear that depending on the nature of datasets we can always receive different results

V. Challenges:

Working with a particular topic which is fairly new comes with a bunch of issues as our work progresses. From the logistic side to a decision making side there have been a few issues but most importantly finding a proper dataset with ample amount of attributes and features and to rely on its authenticity has been challenging as there are also a lot of faults in it. We had to balance the imbalanced dataset by removing some unnecessary attributes and features. The size of the dataset matters as it takes a lot of time to process the data and even after that we had almost 19-20 percent of incorrectly classified instances.

VI. Limitations:

As we have worked with three different machine learning algorithms we came to realization that whenever a dataset with complex structure more attributes and more features the accuracy of the algorithm suffers as the number of incorrectly classified instances grabs higher percentage which is a clear indication that this three algorithms are not a reasonable approach to detect intrusion because attack dataset carries a lots of complex attributes and a very high number of features.

VII. Future Direction:

Our project works well with medium to low level datasets and it is very cost effective so there is no need to look for expansive ways to detect an intrusion as it can be done with less effort and cost and also it saves time. So we are planning to improve its performance even better to handle complex

datasets.

VIII. References

1) 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, doi: 10.1109/ASONAM.2014.6921657. . Panda, M. and Patra, “Network intrusion deteciton using naive Bayes,” Int. J. Comput. Sci. Netw. Secur., vol. 7, no. 12, pp. 258-263., 2012, [Online]. Available: https://www.researchgate.net/publication/241397131.
2) F. Ertam, I. F. Kilinçer, and O. Yaman, “Intrusion detection in computer networks via machine learning algorithms,” IDAP 2017 - Int. Artif. Intell. Data Process. Symp., 2017, doi: 10.1109/IDAP.2017.8090165.
3) P. Aggarwal and S. K. Sharma, “Analysis of KDD Dataset Attributes - Class wise for Intrusion Detection,” Procedia Comput. Sci., vol. 57, pp. 842–851, 2015, doi: 10.1016/j.procs.2015.07.490.
4) J. Zhang, M. Zulkernine, and A. Haque, “Random-forests-based network intrusion detection systems,” IEEE Trans. Syst. Man Cybern. Part C Appl. Rev., vol. 38, no. 5, pp. 649–659, 2008, doi: 10.1109/TSMCC.2008.923876.
5) M. Denscombe, “The Good Research Guide,” no. 1, pp. 6–8, 2003, doi: 10.16309/j.cnki.issn.1007-1776.2003.03.004.
6) P. Johannesson and E. Perjons, A Design Science Primer CreateSpace. 2012.
7) https://www.unb.ca/cic/datasets/nsl.html