Hybrid Intrusion Detection System using K-means and K-Nearest Neighbors Algorithms

Yi Yi Aung
University of Computer Studies, Mandalay
Mandalay, Myanmar
yiyiaung123@gmail.com

Myat Myat Min
University of Computer Studies, Mandalay
Mandalay, Myanmar
myatiimin@gmail.com

***Abstract*— With the widespread use of technology, companies of all sizes have benefited from the use of resources and technologies on the Internet. On the other hand, real security threats are increasing problems and intrusion detection systems (IDS) can help protect external and internal organizations and provide network security. The intrusion detection system only monitors network traffic and informs the system administrator for unusual activity. It is very similar to a home alarm system that will turn on the alarm when the thief enters the window or door. Various methods such as machine learning, data processing and optimization are also analyzed to support the important developments of IDS and to assist better future problems suggestions. In this paper we use hybrid data mining methods such as k-means and k-nearest neighbors to reduce time complexity of the system with great accuracy. This model is implemented by using KDD'99 data set.**

Keywords— Intrusion Detection System (IDS), K-means, K-nearest neighbors, KDD'99 dataset

I. INTRODUCTION (*HEADING 1*)

Security or security information technology is to protect computers, networks, applications and data from accesses or attacks aimed at exploiting unauthorized. The major areas covered by cyber security are application security, information security, disaster recovery, and network security. Network security includes activities to protect the usability, reliability, integrity and online security. Effective network security targets a variety of threats and stops them from entering or spreading on the network. Network security components include: a)Anti-virus and anti-spyware, b)Firewall, to block unauthorized access to your network, c)Intrusion detection systems, to monitors network traffic and monitor for suspicious activity and report or alert the system or network administrator d) Intrusion prevention systems (IPS), to identify fast-spreading threats, such as zero-day or zero-hour attacks, and e) Virtual Private Networks (VPNs), to provide secure remote access [4].

There are multiple ways of detection that is performed by IDS. In signature-based detection, a schema or signature is compared to previous events to discover current threats. This is useful for finding already known threats, but does not help in finding unknown threats, variants of threats or hidden threats. Another type of detection is anomaly-based detection, which compares the definition or traits of a normal action against characteristics making the event as abnormal. There are three primary components of IDS:

- Network Intrusion Detection System (NIDS): This does analysis for traffic on a whole subnet and will make a match to the traffic passing by to the attacks already known in a library of known attacks.
- Network Node Intrusion Detection System (NNIDS): This is similar to NIDS, but the traffic is only monitored on a single host, not a whole subnet.
- Host Intrusion Detection System (HIDS): This takes a "picture" of an entire system's file set and compares it to a previous picture. If there are significant differences, such as missing files, it alerts the administrator [3].

Some of the important features of a Network Intrusion Detection System are as follows [5],

- It should be fault tolerant and run continuously with minimal human supervision.
- A Network Intrusion Detection System must be able to recover from the crashes, either accidental or caused by malicious activity.
- A Network Intrusion Detection System must be able to detect any modifications forced on the IDS by an attacker.
- It should impose minimal overhead on the system.
- It should be configurable so as to accurately implement the security policies of the system.
- It should be easy to use by the operator.
- It should be capable to detect different types of attacks and must not recognize and legitimate activity as an attack.

Data mining techniques can be used for misuse and anomaly intrusion detection. As there are many number of ID

978-1-5386-5892-5/18/$31.00 ©2018 IEEE
ICIS 2018, June 6-8, 2018, Singapore

techniques using data mining techniques, the unknown technique and system could be thought of as a baseline for future prospect. The paper is implemented by using k-means and k-nearest neighbors data mining algorithms to classify normal and attacks.

II. Literature Survey

This paper convinced about the data mining and intrusion detection system, IDS was a necessary tool in any environment. IDS require a lot of planning and research. Once the research was done correctly there would be lots of benefit. Though the need for computer security in both public and private was obvious Intrusion detection system provided the crucial component of the effective computer system. The work done in this paper was to understand the types in intrusion detection and how to provide security for the system. They used any one for the powerful tool in data mining and then provided security to highest level. In this paper they used one of the powerful tool named WEKA and classified the data. Although they have concentrated more on intrusion detection types it provides the security for computer network system. A total network security or computer security was a paradigm it can be viewed as kind of asymptote [1].

In this paper, comparison made in 23 papers for finding out the situation of intrusion detection now a day. After the comparison among these papers, observation shows most researches focus on anomaly detection, and use the tuples of DARPA 1998 and KDD Cup 1999 mostly. In addition, most researches in intrusion detection use ANN. Because ANN was much more stable and reliable than other models and algorithms. Besides, the second most used model was SVM. In the future, additionally focus on misuse direction. With the help of techniques were using in intrusion detection, compare these models in the paper to identify a "best" model for it. Besides, most researches use the tuples of DARPA 1998 or KDD Cup 1999. Some experiments by using DARPA2000 or other tuples to increase objectivities later on. Probably, establish a new tuples to become the criterion of doing experiments in intrusion detection [2].

The rapid evolution of technology and the increased connectivity among its components imposes new cyber-security challenges such as malware with worm capabilities, release of more shadow brokers tools, etc. The challenge is that the struggle between the relentless attackers and the tireless defenders does not always seem to be evenly balanced. To tackle this growing trend in computer attacks and respond threats, industry professionals and academic are joining forces in order to build Intrusion Detection Systems (IDS) that combine high accuracy with low complexity and time efficiency. The present article gave an overview of existing Intrusion Detection Systems (IDS) along with their main principles. Also this article argued whether data mining and its core feature which is knowledge discover could help in creating Data mining based IDSs that could achieve higher accuracy to novel types of intrusion and demonstrate more robust behavior compared to traditional IDSs [6].

Intrusion detection is an important but complex task for a computer system. Here, various methods for intrusion detection are studied and compared. Crisp data mining methods such as ADAM, Random Forest algorithm are used for intrusion detection but suffer from sharp boundary problem which gives less accurate results. In this proposed method use of fuzzy logic overcame the sharp boundary problem. In this paper, they have proposed a GA-based fuzzy Class Association Rule Mining with Sub-Attribute Utilization and its application to classification, which can deal with discrete and continuous attributes at the same time. In addition, this method was applied them to both misuse detection and anomaly detection and performed experiments with practical data provided by KDD99 Cup [7]

This paper selected ten classification algorithms having the potential in terms of high efficiency, capability to handle large data set, good speed, and requiring minimum or no parameter tuning, which were simulated on KDD'99 dataset to generate classification results for Intrusion Detection System (IDS). Consequently, the ten classifiers were benchmarked depending upon the generic IDS evaluation metrics like accuracy and F-score. Based on this empirical study, the overall best results in terms of accuracy and F-score were produced good results in terms of high detection rate and low false alarm rate were Rules-OneR, J48 and Random Forest [8].

As internet continues to expand its usage with an enormous number of applications, cyber-threats have significantly increased accordingly. Thus, accurate detection of malicious traffic in a timely manner is a critical concern in today's Internet for security. One approach for intrusion detection is to use Machine Learning (ML) techniques. Several methods based on ML algorithms have been introduced over the past years, but they are largely limited in terms of detection accuracy and/or time and space complexity to run. They developed the procedure for incorporating multiple ML algorithms for detecting malicious traffic and showed the proposed techniques outperform non-incorporating techniques that utilizes a single ML algorithm. Specifically, their technique showed over 99% accuracy and detection rate, with partial flow information and attributes. They plan to extend their technique with a greater set of ML algorithms to see the benefits and trade-offs [9].

Since the ready-made data mining algorithms is presented, intrusion detection based on the data mining has developed rapidly. It advances in the ability to handle massive data, but it also has problems like, for instance searching for more effective data mining algorithms, how to improve the correct rate of intrusion detection, how to control the rate of false alarm in anomaly detection and etc These can be the topics for future research, meanwhile they also need lots of work and experiments to develop a system that is more effective and more appropriate. There are many types of approaches in intrusion detection, in which that based on the data mining becomes the hot spot in the present intrusion detection methodology. However, data mining is still in its developing stage, so more thorough study needs to be done. A brief survey of IDS in the data mining field was given in this paper [10].

III. Data Mining for Intrusion Detection System

The tracking of the user activities on the network and classifying the malicious and normal activities are termed as intrusion detection. And the system used for this purpose is known as intrusion detection system (IDS). An intrusion detection system is a software or hardware or combination that performs an automatic process of monitoring and analyzing of events. In general, an IDS monitors and records events in a computer system, performs an analysis to determine if the events are security incidents, alerts security supervisors of potential threats, and produces event reports.

Data mining studies automatic techniques for learning to make accurate predictions based on past observations. In the intrusion detection case, data mining can be used to build a system that can distinguish intrusions or anomalies from normal network traffic. To build this kind of system, the first step is for the machine learning algorithms to learn the training dataset, which contains both normal traffic and intrusions. This learning phase results in a model that can be used to determine whether the network traffic is normal or an intrusion. There are many possible algorithms that can be used in the intrusion detection problem; their performance is measured using accuracy rate and false positive rate. In order to achieve a higher accuracy and lower false positive rate, many data mining researchers have proposed various ensemble learning approaches. It is well known in the data mining literature that the appropriate combination of a number of weak classifiers can yield a highly accurate global classifier [11].

There are several reasons why data mining approaches plays a role in intrusion detection system. First of all, for the classification of security incidents, a vast amount of data has to be analyzed containing historical data. It is difficult for human beings to find a pattern in such an enormous amount of data. Data mining, however, seems well-suited to overcome this problem and can therefore be used to discover those patterns [13].

IV. Methodology

This section consists of the conversation of the two algorithms of data mining classification approaches. These are K-means and K-nearest neighbors algorithms. The system design of intrusion detection system can be seen in figure 1.

A. K-means Algorithm

K-means clustering is one of the simplest techniques to solve clustering problem. The main goal is to utilize K-means clustering approach is to split and to group data into normal and attack instance. It works on a dataset D which contains n objects and partitions these objects in k clusters. This methods starts with selecting n objects from D. It calculates the cluster center by using the mean value of the objects in each cluster [14].

The K-means clustering algorithm, starting with k arbitrary cluster centers in space, partitions the set of giving objects into k subsets based on a distance metric. The centers of clusters are iteratively updates based on the optimization of an objective function. This method is one of the most popular clustering techniques, which are used widely, since it is easy to be implemented very efficiently with linear time complexity. The principal goal of employing the K-means clustering scheme is to separate the collection of normal and attack data that behave similarly into several partitions which is known as K-th cluster centroids. In other words, K-Means estimates a fixed number of K, the best cluster centroid representing data with similar behavior. In our work, we predefined K=5, representing Cluster 1, Cluster 2, Cluster 3, Cluster 4 and Cluster 5.

B. K-nearest Neighbors Algorithm

The k-nearest neighbors algorithm (KNN) is a non-parametric method used for classification and regression. In both cases, the input consists of the K closest training examples in the feature space. The output depends on whether K-NN is used for classification or regression:

In KNN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its K nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

In KNN regression, the output is the property value for the object. This value is the average of the values of its k nearest neighbors [15].

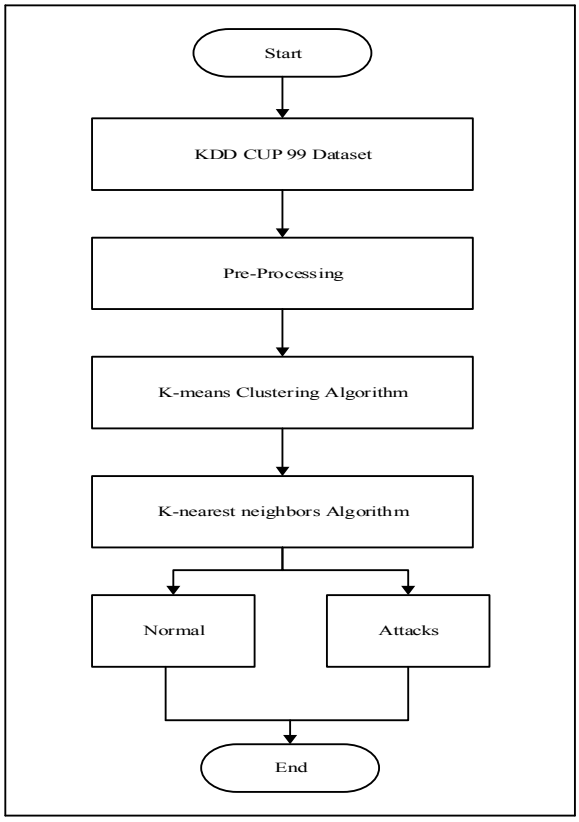

Fig. 1. System flow diagram of intrusion detection system

Identify applicable funding agency here. If none, delete this text box.

V. Experimental Results of Intrusion Detection Model

To facilitate the experiments, we used eclipse java and weka tool to implement the algorithms on a PC with 64-bit window 7 operating system, 4GB RAM and a CPU of Intel core i3-4010U CPU with 1.70GHz. Data come from MIT Lincoln laboratory of KDDCup99 data set. We select 10% data set which contains 494021 Connection records, each record has total of 41 characteristics, 7 symbolic field and 34 numeric fields to experiment because the data sets are very huge. This data set contains four types of intrusions: DoS, Probe, U2R and R2L and also contain normal samples as shown in table 1[3]. Research activities in IDS are still using the KDD Cup 99 dataset for analyzing and exploring new approaches for better IDS.

TABLE I. Class Wise Attacks in KDD Dataset

Class of attack	Attack Name
Normal	Normal
DoS	Neptune, Smurf, Pod, Teardrop, Land, Back
Probe	Ipsweep, nmap, satan, portsweep
R2L	ftp_write, guess_passwd,imap, multihop, phf, spy
U2R	Perl, buffer_overflow, rootkit, loadmodule

Kddcup'99 dataset have two variations of training dataset; one is full training set having 5 million connections and the other is 10% of this training set having 494021 connections. We use 10% data set of KDD CUP 99 to correctly classify the normal and intrusions in the data set. The experimental results of the system are shown from Table II to Table V. The testing results of K-means and k-nearest nekghbors algorithms based on 10 fold cross validation can be seen in table II and table III. And also the results of based on 66-34 percent validation can be seen in table IV and table V.

TABLE II. Testing Results for 10 Fold Cross Validation

dataset	k-means	KNN	Correctly Classified Instances	Correct Instances percentages	Incorrectly Classified Instances	Incorrect Instances percentage
10% P1	Y	Y	108838	99.9936	7	0.0064
10% P2	Y	Y	23491	99.8427	37	0.1573
10% P3	Y	Y	280797	99.9996	1	0.0004
10% P4	Y	Y	78629	99.8375	128	0.1625
10% P5	Y	Y	2054	98.1366	39	1.8634
Total	Y	Y	493809		212	

TABLE III. Testing Results for 10 Fold Cross Validation with Time Complexity

dataset	k-means	KNN	Total instances	Time to build model (sec)
10% P1	Y	Y	108845	0.03
10% P2	Y	Y	23528	0.01
10% P3	Y	Y	280798	0.11
10% P4	Y	Y	78757	0.02
10% P5	Y	Y	2093	0.01
Total	Y	Y	494021	0.18

TABLE IV. Testing Results for 66-34 Percent Validation

dataset	k-means	KNN	Correctly Classified Instances	Correct Instances percentages	Incorrectly Classified Instances	Incorrect Instances percentage
10% P1	Y	Y	37005	99.9946	2	0.0054
10% P2	Y	Y	7981	99.7625	19	0.2375
10% P3	Y	Y	95471	100	0	0
10% P4	Y	Y	26731	99.8282	46	0.1718
10% P5	Y	Y	697	97.8933	15	2.1067
Total	Y	Y	167885		82	

TABLE V. Testing Results for 66-34 Percent Validation with Time Complexity

dataset	k-means	KNN	Total instances	Time to build model (sec)
10% P1	Y	Y	37007	0.03
10% P2	Y	Y	8000	0.01
10% P3	Y	Y	95471	0.13
10% P4	Y	Y	26777	0.03
10% P5	Y	Y	712	0
Total	Y	Y	167967	0.2

In analysis of 10 fold cross validation, the correctly classified instance records of decision k-means and k-nearest neighbors based approach is 493809 records. And the incorrectly classified instance records of that approach is 212 records. The time needed to train the model for the approach is 0.18 seconds. And also, in 66-34 percent validation, the correctly classified instance record of k-means and k-nearest neighbors based approach is 167885 records. And the incorrectly classified instance records of that approach is 82 records. The time needed to train this model is 0.2 seconds.

VI. Conclusion and Futrue Work

This paper proposes a hybrid intrusion detection framework. This framework use two data mining techniques (i.e. K-means and k-nearest neighbors) to detect normal and attacks. Experimental results show that the accuracy of k-nearest neighbors algorithm based on K-means is good in classification of normal and attacks. And also the model training time of K-means and k-nearest neighbors algorithm based intrusion detection system is more suitable time in big data amount of today intrusion database.

REFERENCES

[1] Dr. D. Aruna Kumari, N. Tejeswani, G. Sravani and R. P Krishna, "Intrusion Detection Using Data Mining Technique (Classification), International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 6(2), 1750-1754, 2015.

[2] K. K. Tiwari, S. Tiwari and S. Yadav, "Intrusion Detection Using Data Mining Techniques", International Journal of Advanced Computer Technology (IJACT), ISSN: 2319-7900, 2006.

[3] https://www.techopedia.com/definition/3988/intrusion-detection-system-ids

[4] https://economictimes.indiatimes.com/definition/cyber-security

[5] D. R. Patil and T. M. Pattewar, "A Comparative Performance Evaluation of Machine Learning-Based NIDS on Benchmark Datasets", International Journal of Research in Advent Technology, Vol. 2, E-ISSN: 2321-9637, April 2014.

[6] Z. Dewa and L. A. Maglaras, "Data Mining and Intrusion Detection Systems", International Journal of Advanced Computer Science and Applications, Vol 7, No 1, 2016.

[7] Sandeep D.and M. S. Chaudhari, "Review on Data Mining Techniques for Intrusion Detection System.

[8] P. Aggarwal and S. K. Sharma, "An Emprrical Comparison of Classifiers to Analyze Intrudion Detection", Fifth International Conference on Advanced Computing & Communication Technologies, IEEE, 2015.

[9] U. Albalawi, S. C. Suh and J. Kim, "Incorporating Multiple Supervised Learning Algorithms for Effective Intrusion Detection", International Journal of Computer, Electrical, Automation, Control and Information Engineering, Vol 8 No 2, 2014.

[10] L. S. Parihar and A. Tiwari, "Survey on Intrusion Detection Using Data Mining Methods", www.ijsart.com, ISSN(Online): 2395-1052, Vol. 2, Issue 1, January 2016.

[11] I. Syarif, Ed Zaluska, A. P. Bennett and G. Wills, "Application of Bagging, Boosting and Stacking to Intrusion Detection", 2011.

[12] KDD Cup'99 Intrusion Data Sets, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[13] R. Patel, A. Thakkar and A. Ganatra, "A Survey and Comparative Analysis of Data Mining Techniques for Network Intrusion Detection Systems", International Journal of Soft Computing and Engineering (IJSCE), Vol 2, Issue 1, ISSN: 2231-2307, March 2012.

[14] B. V. C, A. Shaji, Prof. P Jayakumar and Nimmi M. K, "A Study on Intrusion Detection and Protection Techniques", International Conference on Emerging Trends in Engineering & Management (ICETEM), IOSR Journal of Computer Engineering (IOSR-JCE), e-ISSN: 2278-0661, p-ISSN: 2278-8727, pp 07-10, 2016.

[15] https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm