Intrusion Detection System Based on Frequent Pattern Mining

Khin Moh Moh Aung[#1], Nyein Nyein Oo[*2] , Myo Min Than[#3]

[#]*Department of Information Technology, Yangon Technological University,*

Yangon, Republic of the Union of Myanmar.

[1]khinmohmohaung@gmail.com

[2]nno2005@gmail.com.com

[3]dr.myominthan@gmail.com

Abstract— **Due to the dramatically increment of internet usage, users are facing various attacks day by day. Consequently, the research area for intrusion detection must be fresh with new challenges. Intrusion detection system includes identifying a set of malicious actions that compromise the integrity, confidentiality, and availability of information resources. The major contribution is to apply data mining approach for network intrusion detection system. Among the several features of data mining, association rules mining,FP-growth algorithm, is used to find out the frequent itemsets of incoming packets database. Based on these itemsets, anomaly detection is added. The system will predict whether the incoming data packet is normal or attack. The performance of proposed system is tested by using KDD-99 datasets.**

***Keywords*— intrusion detection, data mining, anomaly, algorithm, KDD-99 datasets**

I. INTRODUCTION

The importance of security of the computer networks is continuing to increase as more business is conducted over the Internet. Computer security can be very complex and may be very confusing to many people. Normally, the significant authentic computer network contains many security systems, such as a traditional internet firewall, susceptibility scanning system and encryption. However, such security mechanisms almost always have inevitable vulnerabilities and they are usually not sufficient to ensure complete security of the infrastructure and to ward off attacks that are continually being adapted to exploit the system's weaknesses often caused by careless design and implementation flaws. Intrusions are actions aimed to compromise the confidentiality, integrity, and/or availability of a computer or computer network. So an intrusion detection system (IDS) is a device or software application that monitors network or system activities for malicious activities or policy violations and report them.Intrusion Detection is an area growing in relevance as more and more confidential data are stored and processed in networked systems.

Specifically, there are two most important intrusion detection techniques depending on the model of the scheme used: Anomaly and Misuse Detection. Misuse detection recognizes confirmation of malicious behavior by matching it adjacent to predefined of attacks, or signatures. Anomaly/Statistical detection defines ordinary activities and endeavors to recognize any deplorable divergence as probably the result of an attack [1].

By applying data mining technology, intrusion detection system can widely verify the data to obtain a comparison between the abnormal pattern and the normal behaviour pattern. Manual analysis is not required for this method. One of the main advantages is that same data mining tool can be applied to different data sources. An important problem in intrusion detection is how effectively the attack patterns and normal data patterns can be separated from a large number of network data and how effectively generates automatic intrusion rules after collected raw network data. To accomplish this, association rule mining, FP-growth algorithm is used in this paper.

II. RELATED WORK

ADAM [2] is a wired Apriori based network intrusion detection system. First, ADAM collects normal, known frequent datasets through mining as training datasets. Secondly, during detection, it runs an on-line algorithm to find last frequent connections, which it compares with known mined training normal datasets and it discards those recent connections which seem to be normal. With suspicious records, it then uses a classifier, previously trained to classify and label suspicious connections as a known type of attack, unknown type of attack or a false alarm. Since the system depends on training data, the efficiency also depends on training data. For mining algorithm, it is clear that Apriori costs much time for several scans of large data. An intrusion detection system which used clustering was proposed in [3]. The authors used Online K-means algorithm (KMO) to analyze network traffic data streams. The main limitation is they also used hard-to-get training data. Another interesting research used anomaly detection to detect web application attacks [4]. The web server access log file is used as data source. The detection method is based on the analysis of the querying parameters relationship of the HTTP requests.

Association rule mining was proposed in [5], where the formal definition of the problem is presented as: Let $L = \{i_1, \ldots, i_n\}$be a set of literals, called items. Let database, D be a set of transaction records, where each transaction T is a set of items such that $T \subseteq L$. Associated with each transaction is a unique identifier, called its transaction id (TID). It says that a transaction T contains X, a set of some items in L, if $X \subseteq L$. An association rule is an implication of the form $X \rightarrow Y$, where$X \subseteq L$, $Y \subseteq L$, and$X \cap Y = \emptyset$;. The rule $X \rightarrow Y$ holds in the transaction set D with confidence c if c% of transactions in D that contain X also contain Y. The rule $X \rightarrow Y$ has support s in the transaction set D if s% of transactions in D contain $X \cup Y$.

III. ANOMALY BASED INTRUSION DETECTION

Anomaly-based Detection or Behaviour-based intrusion detection profiles normal behaviour and attempts to identify anomaly patterns of activities that deviate from the defined profile. This approach is based upon the use of user, system or network profiles of normal behaviour, and searches for significant deviations from these profiles to detect security-related problems. It involves features of a user's current session, system resources, or network traffic which is used to determine whether these parameters exceed a certain threshold set by the specific model. To decide whether the system is running according to normal behaviour, several techniques have been proposed in the resent literature. The major drawback of anomaly detection

is defining its rule set. The efficiency of the system depends on how well it is implemented and tested on all protocols. For detection to occur correctly, the detailed knowledge about the accepted network behaviour need to be developed by the administrators. But once the rules are defined and protocol is built then anomaly detection systems works well [6].

Regarding the underlying algorithm, it defines four different possible approaches, but only two of them have successfully employed in the last decade: algorithms based on statistical models and those based on neural networks. The former is the most widely used more than 50% of existing Anomaly Based Systems. In these systems, the algorithm (during the so-called training phase) first builds a statistical model of the – legitimate, attack-free – network behaviour later (in the detection phase), the input data is compared to the model using a distance function, and when the distance measured exceeds a given threshold, the input is considered anomalous, i.e., it is considered an attack [7].

IV. Intrusion Detection Dataset KDD 99

The KDD Cup 1999 dataset used for benchmarking intrusion detection problems is used in this experiment. In the year of 1998 the Defense Advanced Research Projects Agency (DARPA) intrusion detection estimation created the first standard corpus for evaluating intrusion detection systems. The KDD 99 intrusion detection datasets are based on the 1998 DARPA initiative, which provides designers of intrusion detection systems (IDS) with a benchmark on which to evaluate different methodologies.

The dataset was a collection of simulated raw TCP dump data over a period of nine weeks on a local area network. The training data was processed to about five million connections records from seven weeks of network traffic and two weeks of testing data yielded around two million connection records. The training data is made up of 22 different attacks out of the 39 present in the test data. It contains 41 features and is labeled as either normal or an attack, with exactly one specific attack type. The simulated attacks fall in one of the four categories: Probe, DOS, U2R and R2L. The known attack types are those present in the training dataset while the novel attacks are the additional attacks in the test datasets not available in the training data sets. The training dataset consisted of 494,021 records among which 97,277 (19.69%) were normal, 391,458 (79.24%) DOS, 4,107 (0.83%) Probe, 1,126 (0.23%) R2L and 52 (0.01%) U2R connections [8], [9].

V. Proposed Intrusion Detection System

The proposed system uses association rules mining and anomaly detection approach. Fig.1 shows the overview of system design. As shown in figure, the necessary attributes for each transaction are firstly extracted. 7 out of 41 attributes are extracted. The selected attributes are protocol_type, service, flag, dst_host_srv_count, dst_host_same_srv_rate,dst_host_diff_srv_rate, dst_host_ same_src_ port_rate. The reason why these 7 attributes are chosen is that the numbers are nearly the same and so the efficiency of mining algorithm can be clearly seen. Although the values of different attributes in the same transaction are nearly the same, algorithm can mine the correct rules.

As the second phase, the frequent itemsets are mined from the preprocessed data. In this phase, FP-growth algorithm is used for frequent itemsets mining.

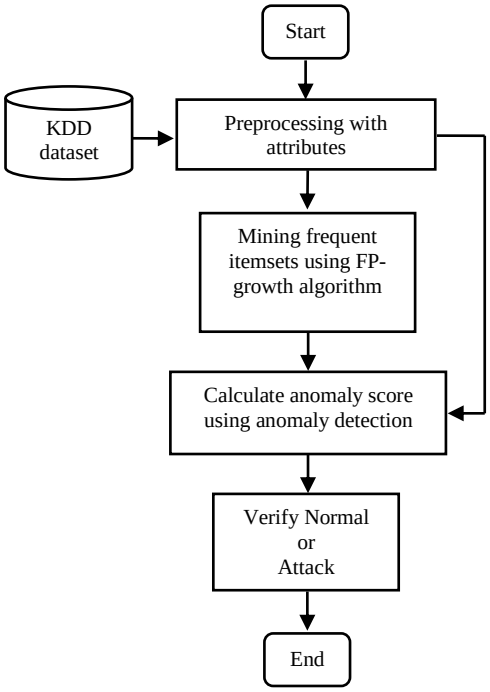

Fig. 1 Proposed system design

FP-growth adopts a divide-and-conquer strategy that compresses the database representing frequent items into a frequent-pattern tree (FP-tree), and proceeds mining of the FP-tree. FP-tree is a good compact tree structure, which contains the complete information of the database in relevance to frequent pattern mining, and its size is usually highly compact and much smaller than its original database. The method is highly compressed so frequent item-sets generation is integrated and do not need to repeatedly scan the item sets.

FP-growth contains two main steps. In step 1, the database is scanned to discover frequent-1 itemsets. The database is scanned again to build a compact representation of transactions in form of FP-tree. In step 2, frequent itemsets are extracted by using FP-tree. In the third phase, anomaly score for each transaction is calculated based on frequent itemsets. The anomaly score assigns +1 to every n-itemset which is infrequent. The anomaly score assigns -1 to every n-itemset which is frequent. An itemset whose frequency is 1 can be assigned as infrequent. The itemset whose frequency is greater than 1 can be assigned as frequent. The frequent is assigned for greater than 1 in order to get strong accuracy for normal-attack decision. The transaction with 0 or positive total score is decided as an attack. The transaction with negative total score is decided as normal packet. The following is FP-growth algorithm which is used for frequent itemset mining.

Input :FP-tree constructed based on Algorithm ,using DB and a minimum support threshold α.
Output : The complete set of frequent patterns.
Method : Call FP-growth (FP-tree ; null).
Procedure *FP-growth* (Tree; α)
{
(1) if Tree contains a single path P
(2) then for each combination (denoted as β) of the nodes in the path P do
(3) Generate pattern $\beta \cup \alpha$ with support = minimum support of nodes in β;
(4) else for each a_i in the header of Tree do {
(5) generate pattern $\beta = a_i \cup \alpha$ with support =a_i.Support;
(6) constructβ's conditional pattern base and thenβ's conditional FP tree Treeβ;

(7) If Tree β ≠ null
(8) then call FP-growth (Tree β, β)}
}

The anomaly detection can be explained with an example. Assume the Table I is the original database. Table II shows the frequent itemsets by using FP-growth algorithm. And Table III shows the calculation of total anomaly score for each transaction. For TID 3, the total score is positive, so it is decided as an attack. The other transactions have negative total score, and so they are normal packets. If the largest number of items in a transaction is n, anomaly score is calculated to frequent n-1 itemset. In the example, the transactions have three items each, and so the score is calculated to frequent-2 itemset.

TABLE I
ORIGINAL DATABASE

TID	Packet	Assigned Value	Normal/ Attack
1	tcp, smtp, SF	A,B,C	normal
2	tcp, smtp, SF	A,B,C	normal
3	Icmp,ecr-I, SF	D, E, C	smurf
4	tcp, smtp, SF	A,B,C	normal
5	tcp, smtp, SF	A,B,C	normal
6	tcp, smtp, SF	A,B,C	normal

TABLE II
FREQUENT ITEMSETS

A:5	-1
B:5	-1
C:6	-1
D:1	+1
E:1	+1
AB: 5	-1
AC: 5	-1
BC: 5	-1
CD: 1	+1
CE: 1	+1
DE: 1	+1

TABLE III
ANOMALY SCORE CALCULATION

TID	Depth-1	Depth-2	Total Score
A,B,C	-1-1-1=-3	-1-1-1= -3	-6 (*normal*)
A,B,C	-1-1-1=-3	-1-1-1= -3	-6 (*normal*)
D, E, C	+1+1-1=+1	+1+1+1=+3	+4(*attack*)
A,B,C	-1-1-1=-3	-1-1-1=-3	-6 (*normal*)
A,B,C	-1-1-1=-3	-1-1-1=-3	-6 (*normal*)
A,B,C	-1-1-1=-3	-1-1-1=-3	-6 (*normal*)

VI. PERFORMANCE ANALYSIS ON TRANSACTION

To test the performance, the proposed system is implemented with java. The input transactions are taken from KDD 99 database. Five different data sets are tested: 100 transactions including 10 attack packets, 200 transactions including 25 attack packets, 300 transactions including 30 attack packets, 400 transactions including 35 attack packets, 500 transactions including 35 attack packets. Fig. 2 describes the test of normal or attack detection for 100 transactions. In Fig. 3, original attack and detected attack are compared for different packets. Fig. 4 shows the accuracy percentage comparison for five different data sets. As shown in Fig. 4, the accuracy gets higher with the larger data size.

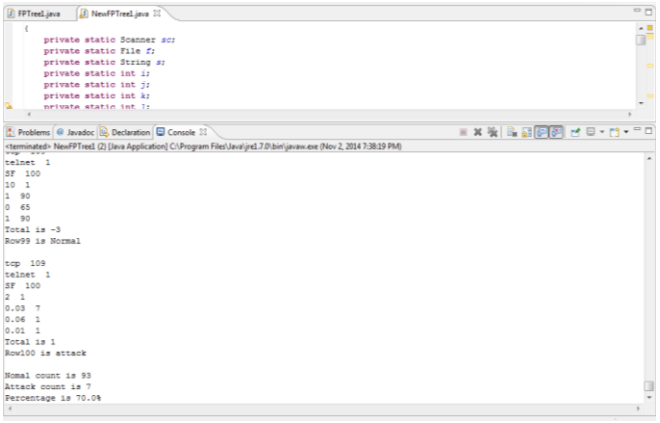

Fig. 2 Example test of 100 transactions

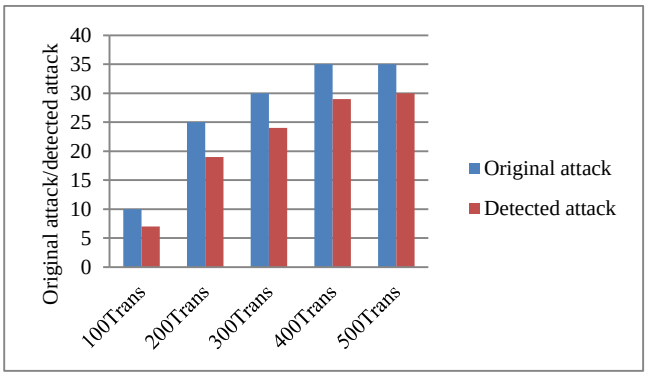

Fig. 3 Comparison of original and detected attack for different transactions

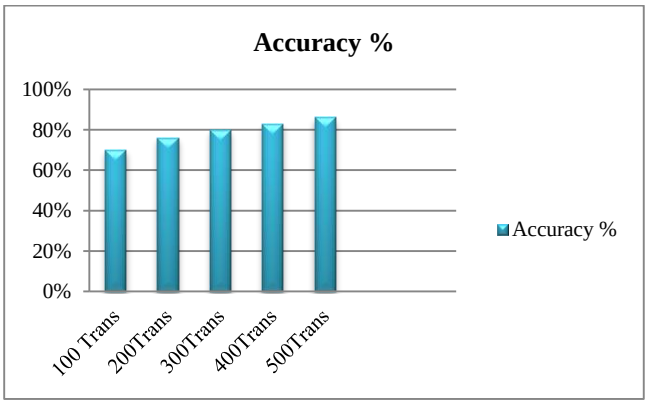

Fig. 4 Performance of accuracy

VII. CONCLUSION

Intrusion detection technology is an effective approach to the problems of network security. In this paper, data mining-based network intrusion detection system using FP-growth is applied. FP-tree is a good compact tree structure, which contains the complete information of the database in relevance to frequent pattern mining, and its size is usually highly compact and much smaller than its original database. From the experimental result, the accuracy for normal-attack decision is better for larger data set. In the future work, the new FP algorithm will be constructed to get more accuracy and more efficiency by using this

proposed system. Furthermore, the system will be able to detect what type of attack is.

ACKNOWLEDGMENT

The authors would like to express special thanks to all who support and encourage preparing and submitting this paper.

REFERENCES

[1] Rashmi Singh and DiwakarSinghJ. Breckling, "A review of network intrusion detection system based on KDD dataset," *Int J EnggTechsci*, vol 5(1), pp. 10-15, 2014.

[2] D. Barbara, J. Couto, S. Jadodia, and N.Wu, "Adam: A testbed for exploring the use of data mining in intrusion detection," ACM SIGMOD RECORD: Special Selection on Data Mining for Intrusion Detection and Threat Analysis, 30(4), 2001.

[3] S. Zhong, T. Khoshgoftaar, and S. Nath, "A clustering approach to wireless network intrusion detection," in *Proceedings of the 17th IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, 2005, p.190–196.

[4] Jingli Zhou, Jifeng Yu, LiqinXiong, "Efficient association rule mining for web application anomaly detection," *International Conference on Electrical and Computer Engineering Advances in Biomedical Engineering*, vol.11, pp. 370-375, 2012.

[5] AhmedurRahman, C.I. Ezeife and A.K. Aggarwal, "Wifi miner: An online apriori-infrequent based wireless intrusion detection system," School of Computer Science, University of Windsor.

[6] V. Jyothsna and V.V.Rama Prasad, "A review of anomaly based intrusion detection systems," *International Journal of Computer Application(0975-8887)*, vol 28-No.7, pp.26-35, 2011.

[7] DamianoBolzoni, SandroEtalle, "Approaches in anomaly – based intrusion detection system," University of Twente, The Netherlands.

[8] Mr. KamleshLahre, Mr. TarundharDiwan, Suresh Kumar Kashyap and PoojaAgrawal, "Analyze different approaches for IDS using KDD 99 data set," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol 1-Issue:8, pp.645-651, August, 2013.

[9] MahbodTavallaee, EbrahimBagheri, Wei Lu, and Ali A. Ghorbani, "A detail analysis of a KDD cup 99 data set," in *Proceeding of the IEEE Symposium on Computational Intelligence in Security and Defense Applications (CISDA)*, 2009.