Performance Analysis of Anomaly based Intrusion Detection System using Modified FP-Growth Algorithm

Khin Moh Moh Aung[#1], Nyein Aye Maung Maung[*2]

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

Yangon, Republic of the Union of Myanmar.

[1]khinmohmohaung@gmail.com

[2]nyeinayemm@gmail.com

Abstract **-Nowadays, research area on network security is facing new challenges with the rapid increment of internet usage and possible attacks day by day. Intrusion detection technology is an effective approach to dealing with the problems of network security. In this paper, performance analysis of the proposed anomaly based intrusion detection system using modified FP-Growth algorithm is carried out based on several parameters. The proposed intrusion detection system is composed of three main parts: preprocessing, normal-attack detection and attack classification. KDD Cup'99 datasets are used as reference input dataset to experimentally analyze the performance of the system. Experimental results show that the proposed intrusion detection system offers preferable accuracy and false alarm rate while reducing the execution time.**

Keywords **- FP-Growth algorithm, intrusion detection, data mining, anomaly, 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. However, the 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. This has created the need for security technology that can monitor systems and identify computer attacks [1]. Intrusion Detection System (IDS) allows organizations to protect their systems from the threats that come with increasing network connectivity and reliance on information system. Intrusion detection systems are software or hardware systems that automate the monitoring of events occurring within a computer system or network [2].

Consequently, the research area for intrusion detection must be fresh with new challenges. Most of the researchers focus on how to identify malicious network behaviors and the characteristics of attack packets, and the way to identify attack patterns based on their analyses. Data mining is a popular technology to create real world applications in many fields, such as educational analysis, market basket analysis and clinical management. In recent year, researchers have explored Data Mining as a new important approach to the IDS.

Mining the large volumes of intrusion detection audit data requires a lot of computational time and resources. Traditional data mining algorithms are overwhelmed by the sheer complexity and bulkiness of the available data. They have become computationally expensive and their execution time largely depends on the size of the data they are dealing with [3].

Although data mining-based intrusion detection system has demonstrated high accuracy, good generalization to novel types of intrusion, it has still challenges dealing with processing time. This paper develops and modifies the association rules mining (FP-growth) for intrusion detection system (IDS) to speed up the detection process and to give high accuracy.

This paper is organized as follows. Section 2 describes related works on intrusion detection system. Section 3 discusses the proposed system and methodology. Section 4 analyses the performance of proposed IDS. Finally section 5 draws the conclusion.

II. RELATED WORKS

In 1980, James P. Anderson presented a paper on "Computer Security Threat: Monitoring and Surveillance". He focused on the collection of records that showed abnormal use of the system, such as use outside of time, abnormal frequency of use, abnormal l patterns of reference to programs or data. He also alerted the problem of the legitimate user that may has access to confidential data; it would be very difficult to record a feasible trail to detect some misuse. His paper was the first based on host intrusion detection and IDS in general [4].

The ADAM (Audit Data Analysis and Mining) system [5] is an anomaly detection system. It uses a module that classifies the suspicious events into false alarms or real attacks. It uses data mining to build a customizable profile of rules of normal behavior and then classifies attacks (by name) or declares false alarms. ADAM is a real-time system. To discover attacks in TCP dump audit trail, ADAM uses a combination of association rules, mining and classification. The system builds a repository of normal frequent item sets that hold during attack-free periods. Then it runs a sliding window online algorithm that finds frequent item sets in the last D connections and compares them with those stored in the normal item set repository. With the rest, ADAM uses a classifier which has previously been trained to classify the suspicious connections as a known type of attack, unknown type or a false alarm. Association rules are used to gather necessary knowledge about the nature of the audit data.

The MADAM ID project at Columbia University has shown how data mining techniques can be used to construct an IDS in a more systematic and automated manner. Specifically, the approach used by MADAM ID is to learn classifiers that distinguish between intrusions and normal activities [6].

In [7], the MINDS project at University of Minnesota uses a suite of data mining techniques to automatically detect attacks against computer networks and systems. Their system uses an anomaly detection technique to assign a score to each connection to determine how anomalous the connection is compared to normal network traffic. Their experiments have shown that anomaly detection algorithms can be successful in detecting numerous novel intrusions that could not be identified using widely popular tools such as SNORT.

WiFi Miner [8] is for dealing with intrusion detection in wireless networks. It is capable of finding frequent and infrequent patterns from pre-processed wireless connection records using infrequent pattern finding Apriori algorithm. This online Apriori-infrequent algorithm improves the join and prune step of the traditional Apriori algorithm with a rule that avoids joining item sets not likely to produce frequent item sets. An anomaly score is then assigned to each packet (record) based on whether the record has more frequent or infrequent patterns. Connection records with positive anomaly scores have more infrequent patterns than frequent patterns and are considered as anomalous packets. The authors described a solution that eliminates the need for hard-to-obtain training data in wireless network environments, and increases intrusion detection rate and reduces false alarms.

In [9], the authors introduced about Data mining technology which can be applied to the network intrusion detection, and can improve the precision of the detection. Basically in this paper, author has presented the study of an example running to contract two algorithms. Presented results have shown that the fuzzy rule mining algorithm is more convenient than Apriori algorithm to mine mass network log database.

Association rule mining plays an important role in the literature of data mining. It poses many challenging issues for the development of efficient and effective methods. Apriori [10] and FP-Growth [11] can be the most representative algorithms for association rule mining, which are commonly used in network intrusion detection systems.

The FP-growth algorithm is the most widely used algorithm for mining frequent item-sets, which is also an algorithm for mining association rules without candidate set. In [12], the authors use FP-tree structure and FP-growth mining method to extract features based on FP-tree without candidate generation. FP-Growth is just accord with the system of real-time and updating data frequently as Network Intrusion Detection System. They employed DARPA intrusion detection evaluation data set to train and test the feasibility of the proposed method. Their experimental results showed that the performance is efficient and satisfactory. With FP-Growth algorithm, the use of divide and conquer strategy to decompose a problem into smaller sub-problems in finding frequent patterns from FP-tree needs large memory when mining large database and its running speed is slow. If the memory cannot hold a large number of conditional FP-tree branches, there will be a performance bottleneck. Mining large data sets may lead to failure [13], [14].

In [15], an efficient intrusion detection system is proposed by using modified FP-Growth algorithm to minimize the processing time while achieving high accuracy of detection rate.

In this paper, performance of the proposed intrusion detection system on processing time, accuracy, precision and false alarm rate are evaluated with different parameters. As well, performance compares with traditional FP-Growth algorithm are also carried out and presented in the evaluation section.

III. Proposed Intrusion Detection system

The proposed system is composed of three main phases: preprocessing, normal-attack detection and attack classification. Fig. 1 shows the detailed system flow diagram of the proposed intrusion detection system.

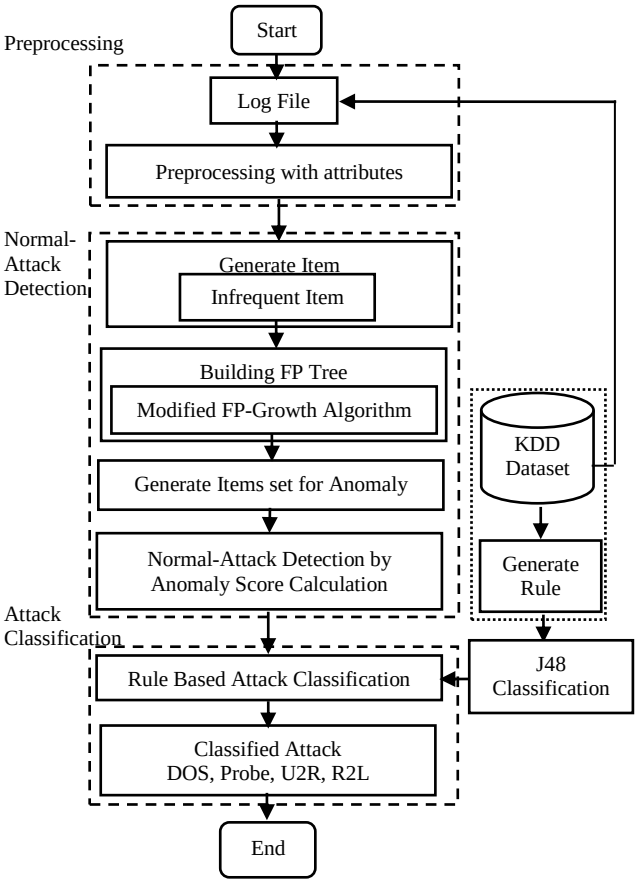

Fig. 1System Flow Diagram

As an incipient preprocessing phase, the proposed system collects the packets log files of the network traffic packets and extracts the utility attributes from these log files. In this system, the KDD Cup'99 datasets are used as reference input dataset. In the first stage of preprocessing phase, the user loads KDD log file into the system and collects transactions from the KDD dataset. In the second stage, every attributes or features of each transaction in the collected dataset are extracted and indexed. These attributes are predefined as system attributes for the proposed anomaly based intrusion detection system. In the third stage, predefined attribute are catalogued while filtering other attributes. Finally, filtered KDD dataset is fetched to the detection phase.

In normal-attack detection phase, modified FP-Growth algorithm is used to find item sets and anomaly score technique is used to determine whether the input transaction is normal or attack. In the first phase, modified FP-Growth algorithm is developed to find out the frequent item sets of incoming transactions database. After scanning the original database, the count of the each item in the database is found. And then, infrequent items database is created by cataloguing from the original database. From the infrequent items database, modified FP-growth tree is constructed and item sets will be got. After that anomaly score phase are executed to detect the normal and attack transaction. The anomaly score assigns +n to every n-item set which is infrequent. The anomaly score assigns -n to every n-item set which is frequent. An item set whose frequency is 1 can be assigned as infrequent. The items set whose frequency is greater than 1 can be assigned as frequent. The positive total score is decided as an attack. Otherwise negative total score is normal packets.

In the classification phase, the proposed system uses rule based approach to classify the detected attack into the

different types of attack. For the proposed system, J48 algorithm is used to classify the types of attack. J48 is an open source Java implementation in the WEKA data mining tool. After importing the trained dataset, WEKA tool generate the IF-THEN Rules and generate dynamic java codes. According of the filtering with the J48 rule algorithm, proposed system will generate the classified attacks such as DoS or U2R or R2L or Probe.

IV. Performance Analysis of Proposed Intrusion Detection System

The proposed system is implemented using Java programming language. KDD Cup'99 datasets are used as reference input dataset.

A. Execution Time

This section describes performance analysis on execution time using traditional FP-Growth algorithm (FP) and modified FP-Growth algorithm (MFP). An intrusion detection system that performs its analysis as quickly as possible enables the security officer or the response engine to promptly react before much damage is done. Thus, it prevents the attacker from subverting the audit source or the intrusion detection system itself. The system's response is the most important step when combating an attack [16].

The experiment is tested by using data transactions from 500 to 5000 by 500 increments. Comparison between the resulting execution time using modified FP-Growth algorithm and traditional FP-Growth algorithm for different number of transactions are illustrated in Table 1 and Fig. 2.

TABLE I
Execution Time Comparison

Trans	FP	MFP
500	3278	2420
1000	7739	6444
1500	14419	12746
2000	19664	17998
2500	29612	27521
3000	33122	31623
3500	35416	33495
4000	42342	39970
4500	45821	43730
5000	46712	43917

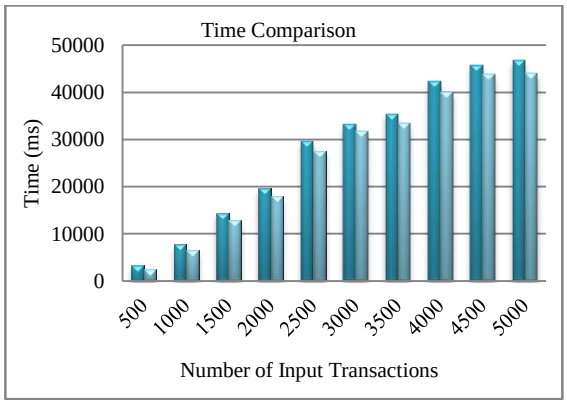

Fig. 2 Execution Time Performance Measures of Modified FP-Growth

According to the comparative results from Table 1 and Fig. 2, modified FP-Growth algorithm performs faster than traditional FP-Growth algorithm in all cases.

B. Analysis with Additional Parameters

An effective Intrusion Detection System (IDS) requires high accuracy and detection rate as well as low false alarm rate. In general, the performance of IDS is evaluated in term of accuracy, detection rate, and false alarm rate. Recently, false alarms rate and accuracy of detection are happened to be the most important issues and challenges in designing effective IDSs. There are many measures available for evaluating the effective of IDS on predictive ability to give a correct classification event to be attack or normal behavior.

Table 2 illustrates for major measures: true negatives (TN), true positives (TP), false positives (FP) and false negatives (FN); which is also known as the confusion matrix. True negatives as well as true positives correspond to a correct operation of the IDS; True negatives (TN) are events which are actually normal and are successfully labelled as normal, true positives (TP) are events which are actually attacks and are successfully labelled as attacks. Respectively, false positives (FP) refer to normal events being classified as attacks; false negatives (FN) are attack events incorrectly classified as normal events. Equations 1 to 3 describe mathematical formulas for evaluating the performances of proposed system based on parameters in the confusion matrix [17].

TABLE III
Confusion Matrix

Actual Class	**Predicted Class**	
	Normal	**Attack**
Normal	True negative (TN)	False positive (FP)
Attack	False negative (FN)	True positive (TP)

$$Accuracy = \frac{TP+TN}{TN+FP+FN+TP} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$False\ Positive\ Rate\ = \frac{FP}{FP+TN} \quad (3)$$

Table 3 shows resulting confusion matrix of various numbers of input data transactions from 500 to 5000 using proposed system.

TABLE IIIII
TP, TN, FP and FN (Confusion Matrix)

500Trans		**1000Trans**		**2000Trans**	
N-N (TN)	438	N-N (TN)	937	N-N (TN)	1870
N-A (FP)	16	N-A (FP)	13	N-A (FP)	30
A-N (FN)	12	A-N (FN)	18	A-N (FN)	33
A-A (TP)	38	A-A (TP)	32	A-A (TP)	67
3000Trans		**4000Trans**		**5000Trans**	
N-N (TN)	2798	N-N (TN)	3737	N-N (TN)	4691
N-A (FP)	52	N-A (FP)	63	N-A (FP)	59
A-N (FN)	31	A-N (FN)	54	A-N (FN)	40
A-A (TP)	119	A-A (TP)	146	A-A (TP)	210

Fig. 3 illustrates the experimental results on accuracy of the proposed system and traditional FP-Growth, which are calculated based on equation (1). According to the results, proposed system shows around 95% to 98% accuracy. Therefore, the proposed system offers stable and good

accuracy as traditional FP-Growth while minimizing the execution time of FP-Growth algorithm.

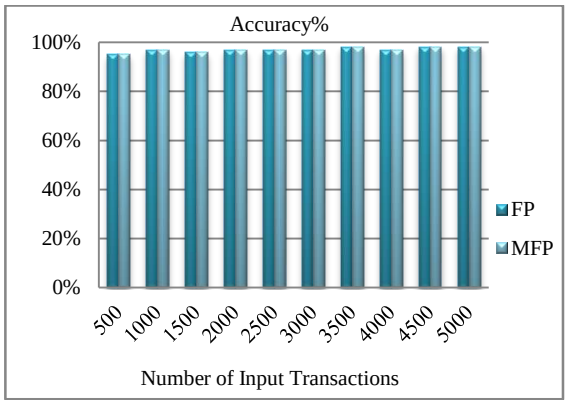

Fig. 3 Accuracy of proposed system

Experimental results on precision of the intrusion detection system using proposed modified FP-Growth and traditional FP-Growth which are estimated using equation (2) are presented in Fig. 4.

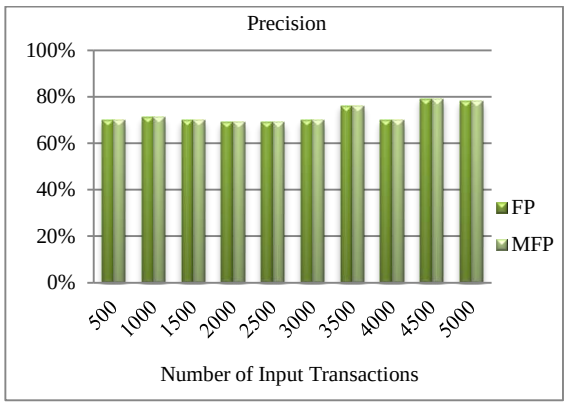

Fig. 4 Precision of proposed system

Precision is how many records are correctly classified by the system. Precisions are also calculated for various numbers of input data transactions (from 500 to 5000). Results indicate that the proposed system gives the precision rate around 69% to 78%.

Additionally, performance of proposed intrusion detection system and traditional FP-Growth algorithm on false alarm rate are also evaluated. False alarm rate (FAR) refers to the proportion that normal data is falsely detected as attack behaviour. Fig. 5 shows false alarm rates for modified FP-Growth algorithm and traditional FP-Growth algorithm by using equation (3). Both systems have very low false alarm rate of under 3%. From 1000 to 5000 transactions, the false alarm rates are 1% and 2% only. So, the system effectively reduces the false alarm rates while increasing the accuracy of the system.

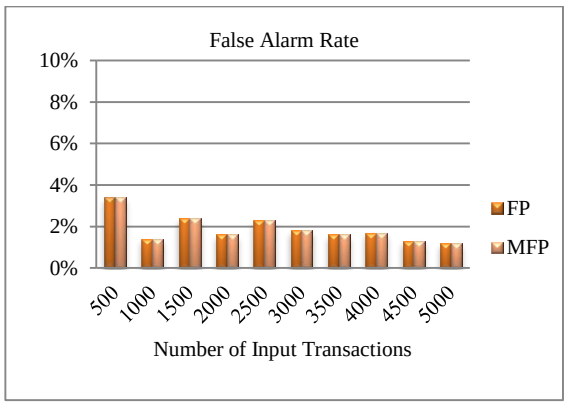

Fig. 5 False Alarm of proposed system

V. Conclusions

In this paper, performance of the proposed data mining-based network intrusion detection system using modified FP-growth is analysed based on execution time, accuracy, precision and false alarm rate. According to the experimental results, the proposed system significantly reduces the execution time of the FP-Growth algorithm. Moreover, the proposed system achieves accuracy of 95% to 98% and it also offers precision rate from 69% to 78% for different size of transactions. As well, it shows very low false alarm rate of around 3%. So the proposed system gives good accuracy and reduces the false alarm rate while reducing the processing time.

Acknowledgment

The author would like to express special thanks to the Rector of Yangon Technological University and Head of Department of Information Technology for their providing and kindly permission to submit this paper. The author is deeply grateful to the supervisor for the invaluable guidance, supervision and suggestion. The author also thanks to ICSE 2015 Steering Committee and all who support and help for preparing this paper.

References

[1] Aleksandar Lazarevic, Vipin Kumar andJaideep Srivastava,"Intrusion detection: A survey," *Computer Science Department, University of Minnesota*.

[2] Seymour Bosworth and M.E. Kabay, *Computer Security Handbook*, 4th ed., Canada: John Wiley & Sons, 2001.

[3] M. Siddiqui, "High Performance Data Mining Techniques for Intrusion Detection," M.Sc. thesis, School of Computer Science, College of Engineering & Computer Science, University of Central Florida, Orlando, Florida, Spring Term 2004.

[4] James P. Anderson, "Computer security threat: Monitoring and surveillance,"James P. Anderson Co. Box 42 Fort Washington, Pa. 19034 215646-4706, Rev April 15 1980.

[5] D. Barbara, J. Couto, S. Jadodia, and N.Wu, "Adam:Detecting intrusion by data mining," in *Proc. IEEE*, 2001, paper, p 11-12.

[6] Wenke Lee and Salvatore J. Stolfoand Kui W. Mok, "Mining audit data to build intrusion detection models," *AAAI*, pp 1-7, 1998.

[7] AleksandarLazarevic, LeventErtoz, Vipin Kumar, AyselOzgur and Jaideep Srivastava, "A comparative study of anomaly detection schemes in network intrusion detection," Computer Science Department, University of Minnesota, USA.

[8] A. Rahman, C.I. Ezeife and A.K. Aggarwal, "WiFi Miner: An Online Apriori-Infrequent Based WirelessIntrusion Detection System," University of Windsor, Windsor, Ontario N9B 3P4.

[9] Ye Changguo, Wei Nianzhong and Wang Tailei, "The research on the application of association rules mining algorithm innetwork intrusion detection," in *Proc. IEEE*, 2009, paper, p 849-852.

[10] R. Agrawal, T. Imielinski and A. Swami, "Mining Association Rules between Sets of Items in Large Database," in *Proc.1993ACM SIGMODConf.*,1993, pp. 1-10.

[11] Jia Wei Han, Jiai Pen, Yiwen Yin and Runying Mao, *Mining Frequent Patterns without CandidateGeneration: A Frequent-Pattern TreeApproach*,Kluwer Academic Publishers, Manufactured in The Netherland, 2004.

[12] T.Peng and W. Zuo, "Data Mining for Network Intrusion Detection System in Real Time," in *Proc. IJCSNS, 2006*, vol. 6, No.2B, February 2006, pp. 173-177.

[13] Xianhong Zhang, "Research of data mining algorithm based on the intrusion prevention system," Applied Mechanics and Materials, vol. 644 650, pp 1787-1790, 2014.

[14] Li Yin Huan, "Design of Intrusion Detection Model Based on Data Mining Technology", in *Proc. ICICEE*, 2012, paper, p. 571-574.

[15] Khin Moh Moh Aung and Nyein Nyein Oo, "Association Rule Pattern Mining Approaches Network Anomaly Detection," in *Proc.ICFCT*,2015,paper,p.216.

[16] Ali A. Ghorbani, Wei Lu and MahbodTavallaee, *Network Intrusion Detection and Prevention: Concept and Techniques*, London, Springer: 2010.

[17] A. Mokarian, A. Faraahi and A. G.Delavar, "False Positives Reduction Techniques in Intrusion Detection Systems-A Review,"*IJCSNS*, vol.13 No.10, pp 128-134, October 2013.