ISBN 978-93-84468-20-0
Proceedings of 2015 International Conference on Future Computational Technologies (ICFCT'2015)
Singapore, March 29-30, 2015

Association Rule Pattern Mining Approaches Network Anomaly Detection

Khin Moh Moh Aung[1] and Nyein Nyein Oo[2]

[1,2]Department of Information Technology, Yangon Technological University, Yangon, Myanmar

Abstract**:** *The research area for intrusion detection is becoming growth with new challenges of attack day by day. Intrusion detection system includes identifying a set of malicious actions that compromise the integrity, confidentiality, and availability of information resources. The major objective of this paper is to apply association rule pattern mining approaches for network intrusion detection system. In this paper, traditional FP-growth algorithm, one of the association algorithms is modified and used to mine itemsets from large database. The required statistics from large databases are gathered into a smaller data structure (FP-tree). The itemsets generated from FP-tree are used as profiles to check anomaly detection in the proposed system.*

Keywords: *data mining, intrusion, anomaly, frequent itemset, algorithm*

1. Introduction

Network security technology has become crucial in protecting government and industry computing infrastructure. Modern intrusion detection applications facing complex problems. Intrusion detection is an area growing in relevance as more and more sensitive data are stored and processed in networked systems. A comprehensive Intrusion Detection System (IDS) requires a significant amount of human expertise and time for development. Data mining-based IDSs require less expert knowledge yet provide good performance [1]. Intrusions are actions aimed to compromise the confidentiality, integrity, and/or availability of a computer or computer network. Intrusion detection is the process of monitoring and analyzing the events occurring in a computer system in order to detect signs of security problems.

There are two types of intrusion detection techniques depending on the model of the scheme used: Anomaly and Misuse Detection. Anomaly technique is based on the detection of traffic anomalies. The deviation of the monitored traffic from the normal profile is measured. Various different implementations of this technique have been proposed, based on the metrics used for measuring traffic profile deviation. Misuse technique looks for patterns and signatures of already known attacks in the network traffic. A constantly updated database is usually used to store the signatures of known attacks. The way this technique deals with intrusion detection resembles the way that anti-virus software operates [2].

Association rules data mining technology can be widely used in intrusion detection system 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. In the proposed system, modified FP-growth algorithm is presented to apply in network intrusion detection system.

2. Related Work

The term data mining is frequently used to designate the process of extracting useful information from large databases. Data mining and machine learning technology has been extensively applied in network intrusion detection and prevention systems by discovering user behavior patterns from the network traffic data. Association rules and sequence rules are the main technique of data mining for intrusion detection. ADAM [3] 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. In [4], the authors approach to adopt a novel 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 (NIDS).

The other researcher used Histogram based detector to identify anomalies and then applied Association rule mining to extract anomalies. Apriori and FP Growth algorithm was used to generate the set of rule applied on metadata. In that paper, the author compared the results of Apriori algorithm and FP Growth algorithm and showed how FP Growth algorithm achieves better results in reducing the time and space complexity. Implementation by FP Growth Algorithm was extended as future work [5].

3. 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].

4. Data Mining Association Rule

Data mining generally refers to the process of extracting or mining knowledge from a large amount of data. This process first understands the existing data and then predicts the new data. In general, data mining tasks are categorized into two categories: predictive and descriptive. The general properties of the data in the database are characterized by descriptive mining. Inference on the current data in order to make predictions is performed by predictive mining. Specifically, two data mining approaches have been proposed [8] and used for anomaly detection: association rules and frequency episodes. Association rule algorithms find correlations between

features or attributes used to describe a data set. Association rules mining started as a technique for finding interesting rules from transactional databases. Association rule mining was proposed in [9], 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). The 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→Y, where $X \subseteq L$, $Y \subseteq L$, and X∩Y=Ø;. The rule X→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→Y has support *s* in the transaction set D if *s*% of transactions in D contain X *U* Y.

Given a set of items I={$I_1,I_2,\ldots,I_m$} and a database of transactions D={$t_1,t_2, \ldots, t_n$} where t_i={$I_{i1},I_{i2}, \ldots, I_{ik}$} and I_{ij} I, the Association Rule Problem is to identify all association rules X→Y with a minimum support and confidence. The support of the rule is the percentage of transactions that contains both X and Y in all transactions and is calculated as |X Y| / |D|. The support of the rule measures the significance of the correlation between itemsets. The confidence is the percentage of transactions that contain Y in the transactions that contain X. The confidence of a rule measures the degree of correlation between the itemsets and is calculated as |X Y | / |X|. The support is a measure of the frequency of a rule and the confidence is a measure of the strength of the relation between the sets of items [8].

5. Frequent Pattern Mining

Frequent pattern mining was first proposed in [9] for market basket analysis in the form of association rule mining. It analyses customer buying habits by finding associations between the different items that customers place in their "shopping baskets".

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. The required statistics from large databases are gathered into a smaller data structure (FP-tree), which is generated with just two database scans. Network Intrusion Detection System adopts FP-growth algorithm in anomaly detection. The following is traditional FP algorithm which is used for frequent itemsets 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 β U α with support = minimum support of nodes in β;
(4) else for each a_i in the header of Tree do {
(5) generate pattern β = a_i U α with support = a_i . Support;
(6) construct β's conditional pattern base and
then β's conditional FP tree Tree β;
(7) If Tree β ≠ null
(8) then call FP-growth (Tree β, β)}
}

The following steps are proceeded to construct the FP-tree.

1. Scan the transaction database DB once. Collect the set of frequent items *F* and their supports. Sort *F* in support descending order as *L*, the list of frequent items and discard items that are below the threshold value (α).

2. Create the root of an FP-tree, *T*, and label it as "null". For each transaction *Trans* in DB do the following processes:
   Select and sort the frequent items in *Trans* according to the order of *L*. Let the sorted frequent item list in *Trans* be [p | P], where p is the first element and P is the remaining list.

   If T has child node (N) such that N.item-name=p.item-name, then increment N's count by 1; else create a new node N, and let its count be 1, its parent link be linked to T, and its node-link be linked to the nodes with the same item-name via the node-link structure. If P is nonempty, call insert tree (P;N) recursively.

TABLE I: Database Generated by Scanning of Traditional FP Algorithm

TID	Items
1	A B C
2	A B C
3	D E C
4	A B C
5	A B C
6	A B C

Original database

Items	1 Item's Count
A	5
B	5
C	6
D	1
E	1

Frequent 1 items

TID	Frequent Items (order)
1	C A B
2	C A B
3	C D E
4	C A B
5	C A B
6	C A B

Frequent items database

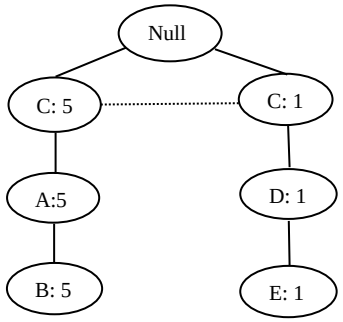

Fig. 1: Example of FP-growth tree

In Table 1, Original database have transactions and the items. After scanning the original database, the count of the each item is found. And then, items are ordered by descending in each transaction. From the frequent items database, FP-growth tree is constructed as in Fig.1 and itemsets will be got.

6. Proposed FP Algorithm and System Design

In the traditional FP-growth algorithm, the frequent patterns are found and infrequent patterns are discarded. Mining infrequent patterns is a challenging endeavor because there are enormous numbers of such patterns that can be derived from a given data set. More specifically, the key issues in mining infrequent patterns are how to identify interesting infrequent patterns, and how to efficiently discover them in large data sets. Since mining infrequent item sets is also very important in many applications such as anomaly detection system. The following Table 3 is modified FP-Growth algorithm for intrusion detection system.

TABLE II: Modified FP-Growth Algorithm

Input: Datasets and a minimum support threshold α
Output: The complete set of infrequent patterns.
F1 = find-frequent-1 itemsets (D)
new D = create-Infrequent-Itemsets-Database (D,F1)
R = root of Tree
for each level $L_i \geq 1$
Obtain all subsets of L_i & add them as leaf nodes
Increment the count of each leaf node depending on the non-existence item in new D
end for
Traverse the Tree to generate Frequent Itemsets with count
Procedure find-frequent-1 itemsets (D)
scan D
F1 = all single items with frequency counts
return F1
end Procedure
Procedure create-Infrequent-Itemsets-Database (D, F1)
scan D
for each transaction T_i
D2. T_i = add T_i.item if its support ≤ minimum threshold
return D2
end Procedure

TABLE III: Database Generated by Scanning of Proposed FP Algorithm

TID	Items
1	A B C
2	A B C
3	D E C
4	A B C
5	A B C
6	A B C

Original database

Items	1 Item's Count
A	5
B	5
C	6
D	1
E	1

Frequent 1 items

TID	Infrequent Items
1	D E
2	D E
3	A B
4	D E
5	D E
6	D E

Infrequent items database

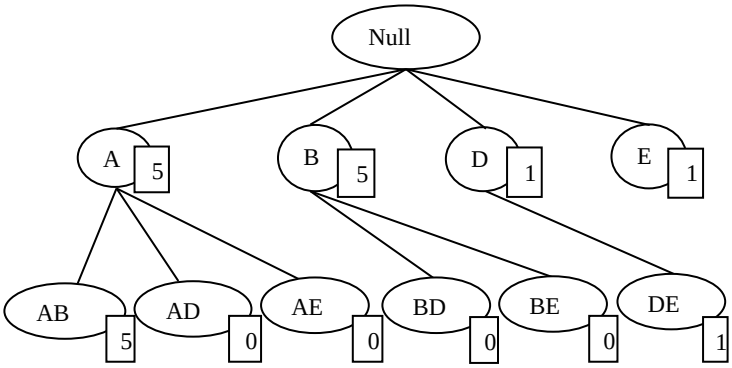

Fig. 2: Example proposed FP-growth tree

In Table 3, Original database have transactions and the items. After scanning the original database, the count of the each item is found. And then, infrequent items are taken in each transaction. From the infrequent items database, proposed FP-growth tree is constructed in Fig. 2 and itemsets will be got.

TABLE IV: Comparison of Proposed FP Database and Traditional FP Database

TID	Infrequent Items
1	D E
2	D E
3	A B
4	D E
5	D E
6	D E

TID	Frequent Items (Order)
1	C A B
2	C A B
3	C D E
4	C A B
5	C A B
6	C A B

By comparing trees and new database in Table 4, the proposed FP algorithm is more compact and it can mine easily from large database because of space and time reduction. The modified algorithm is proposed to find the infrequent patterns for network anomaly detection in this paper.

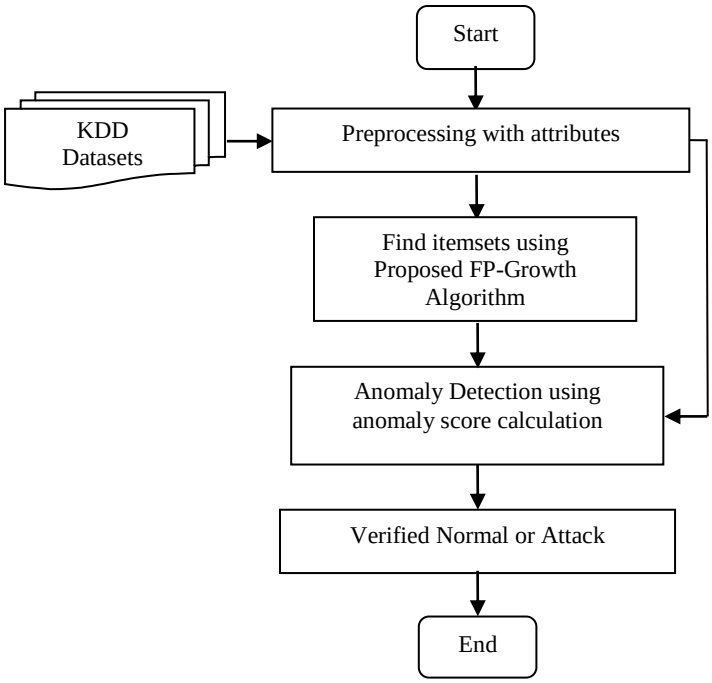

Fig. 3: Overview of Proposed System Design

Fig. 3 shows the overview of proposed system design. As shown in figure, attributes are preprocessing from KDD datasets (example standard datasets). Proposed FP-growth algorithm is used to find out the itemsets. Based on these itemsets, anomaly detection is applied. The system will predict normal or attack for each packet according the total score calculation.

7. Conclusion

Intrusion detection technology is an effective approach to the problems of network security. In this system, data mining-based network intrusion detection system using proposed FP-growth is applied. Proposed FP-tree is a good compact tree structure, which contains the complete information of the database in relevance to pattern mining, and its size is usually highly compact and much smaller than its original database. The proposed

algorithm works efficiently to find the infrequent items. Moreover, large amount of data can be handled for network traffic in Network Intrusion Detection System.

8. Acknowledgements

The author would like to express special thanks to ICFCT-2015 Organizing Committees and all who help and support for this paper.

9. References

[1] E. K. Reddy, V. N. Reddy and P. G. Rajulu, "A Study of Intrusion Detection in Data Mining," in *Proc. 2011 WCE*., 2011, July 6 - 8, 2011, London, U.K.

[2] S. P. Parekh, B. S. Madan and R. M. Tugnayat, "Approach For Intrusion Detection System Using Data Mining," *Journal of Data Mining and Knowledge Discovery*, vol. 3, Issue 2, 2012, pp.-83-87.

[3] D. Barbara, J. Couto, S. Jajodia and N. Wu, "ADAM: A Testbed for Exploring the Use of Data Mining in Intrusion Detection," in *Proc*. *SIGMOD Record*, vol. 30. No. 4, December 2001, pp. 15-24

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

[5] G. Joshi, "Anomaly Extraction Using Association Rule Mining," in *Proc. 2014 IJERA*, vol. 4, Issue I, January 2014, pp. 88-92.

[6] V. Jyothsna and V. V. Rama Prasad, "A Review of Anomaly based Intrusion Detection Systems," in *Proc*. *2011 IJCA*, vol. 28, No.7, September 2011, pp. 26-35.

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

[8] K. Nalavade .and B.B. Meshram, "Finding Frequent Itemsets using Apriori Algorithm to Detect Intrusions in Large Dataset," in *Proc*. *2014 IJCAIT*, vol. 6, Issue I, June-July 2014, pp. 84-92.

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