Proc. of the Intl. Conf. on Advances in Computer Science and Electronics Engineering — CSEE 2013

ISBN: 978-981-07-5461-7

Intrusion Detection Systems by Applying An Improvement in Apriori Algorithm

Samaneh S. Alavifar, Akbar Farhoodi Nejad, Ahmad Faraahi

***Abstract*— For many years, researchers have used the discovery of intrusion technology to supply the security of networks and computer systems. Increasing Cyber threats has caused the need to have a flexible approach that can offer the efficiency of this type of attack and to deal with more complex and unknown attacks. On this basis, the systems of intrusion discovery based on data mining have been presented that derives the unknown patterns and Cyber attacks from the data network. What is presented in this paper is an improvement on Apriori algorithm to realize the intrusion pattern in less time in the network and take proper reaction against it. (Abstract)**

***Keywords*— intrusion detection, data mining, Apriori algorithm, distributed systems (key words)**

I. Introduction

Information technology development and implementation of System Information around the world, have been caused security concerns become an important issue in the scientific community, private and public sectors. Since the network admitted as encompassing basic information about the state of the art of the security of data and lack of permeability through the network, Information security as a special case, has been studied so far. In recent years, many techniques have been used by researchers for the design of intrusion detection systems.

These methods and techniques can detect anomalies through on statistical methods [4, 5], Neural Networks [6, 7] and detection based on data mining [3, 8, 9, and 10].

II. THEORETICAL BACKGROUND

A. *Intrusion detection systems*

Intrusion Detection System is responsible for identification any illegal logging, abuse or harm by both internal and external users.[15]

In general, the intrusion is referred to illegal acts, which endangers accuracy, privacy and accessibility. [1]

Method used in the Intrusion detection systems is divided into two categories [11] anomaly detection and misuse detection:

- Method to identify abnormal behavior (anomaly detection) which creates a pattern of normal behavior and any abnormalities that may not be consistent with the pattern may be indicative of an intrusion.
- Signature based (misuse detection): in this technique, pre-designed patterns (made by signature) are maintained as the rule [12, 13].

B. *Data mining*

Data mining is a set of techniques by which they can discover hidden knowledge in data, ie dynamic model, through data obtained. So far, many algorithms have been introduced at the International Conference on Data Mining (ICDM) attempted to find algorithms that are well-known and powerful algorithms in which C4.5, k-Means, SVM, Apriori - , EM, Page Rank, AdaBoost, kNN, Naive Bayes, and CART [14]. Data mining algorithms were selected as the top ones. Apriori algorithm is one such algorithm based on the improved algorithm, this paper is a kind of developed Apriori algorithm.

III. PROPOSED ALGORITHM

The detection rate of abnormal model besides accurately detection in IDS is very important that with increased dimensions and network traffic, most of what we see so far, are highlighted in the papers. The object of this paper is to design the smart detection system based on data mining that divide computational load between two or more systems may result in preventing bottlenecks in the system and increase the speed of the intrusion detection system. Two separate systems have been simulated as virtual machines on a system. An overview of the proposed system is shown in Figure1.

As figure illustrates, data mining operations by two separate systems is implemented as parallel, in the intrusion detection system. Proposed structure consists of three main parts:

1. Data collection and preprocessing module network
2. Balanced distribution module data between two systems to obtain rules
3. Intrusion detection analysis module

Samaneh S. Alavifar,M.s
Computer Engineering and Information Technology Payame Noor University
Islamic Republic of Iran

Akbar Farhoodi Nejad, Phd
Computer Engineering and Information Technology Payame Noor University
Islamic Republic of Iran

Ahmad Faraahi, Phd
Computer Engineering and Information Technology Payame Noor University
Islamic Republic of Iran

Proc. of the Intl. Conf. on Advances in Computer Science and Electronics Engineering — CSEE 2013

ISBN: 978-981-07-5461-7

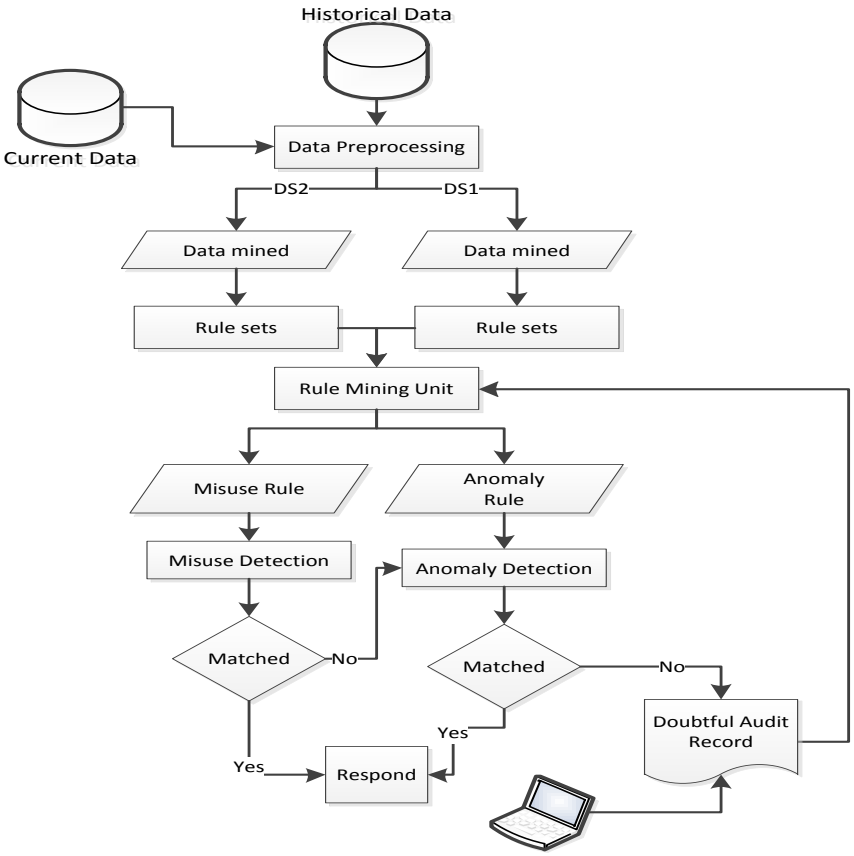

Figure 1- Proposed Structure in IDS

In the first module, first historical data of network with the current data of users are collected and conduct preprocessing of data as they can be acceptable input data of detection system.

Then, in the second module, we determine that the data of database is made of what item sets (1 - Itemset)? After extracting them, we get their number and the amount to be divided by 2. Then we send the first half of item set to the first system and the second half to the second system. The first System, takes all compounds containing an item sets (1 - Itemset) from main Database and the second system also takes its elements in the same manner, and then calculate its composition, and remove them from the database, after this stage, each systems separately extract its rules. Finally, what remains of the database items are common between the two systems must calculate their number and their rules must be determined.

IV. RESULT AND DISCUSSION

The simulation of this system was done in Visual C #. Net 2008 and system specifications Intel Core 2 Duo 2.Ghz. (T6400) and 2Gb DDR2 Ram and 320Gb SATA HDD, the Number of each item in the database has been created by the randomized algorithm to test the value of 100,000 up to a million variables. The value of Minsupport and Minimum confidence can be changed dynamically in the simulation program. We have assumed in all states Minimum support = 1% and Minimum confidence = 2%.

A. *Optimum State*

The best case is when the data is split between two distributed systems have not any association. In this case, after data division between the two systems, no data remains in the main database to be processed.

The table below shows the production data and their combination:

Proc. of the Intl. Conf. on Advances in Computer Science and Electronics Engineering — CSEE 2013

ISBN: 978-981-07-5461-7

Table 1- Data Set For Optimum State

Item	The Number Of Items									
ab	7888	1885	7556	27024	48734	94861	4352	108750	55409	136397
abc	3920	16620	38479	40840	51670	65425	86830	35718	38048	60687
ac	13157	29584	7278	28542	47895	10957	88839	106996	35155	88948
bc	8819	30471	45340	18977	53179	37385	41294	74945	10995	20675
a	11357	295	21047	39018	50159	966	91406	1916	135556	71247
b	1157	22382	28140	38006	11910	12354	91517	10205	93460	79579
d	10972	18726	19007	46204	50998	84338	34045	32369	136569	45815
f	8739	12262	15948	23076	60479	81539	47656	47680	22713	95790
ef	13414	3102	33796	32540	32544	15899	37946	88020	52805	81362
df	3052	16547	39128	46255	18788	78639	25223	38110	138161	116020
def	7678	33795	42008	33456	24408	96257	102678	98704	106044	58456
de	9847	14331	2273	26062	49236	21380	48214	156587	75085	145024
Sum	100000	200000	300000	400000	500000	600000	700000	800000	900000	1000000

We run both algorithms of basic Apriori, and proposed algorithm with data in Table 1, and the results from run has been gained:

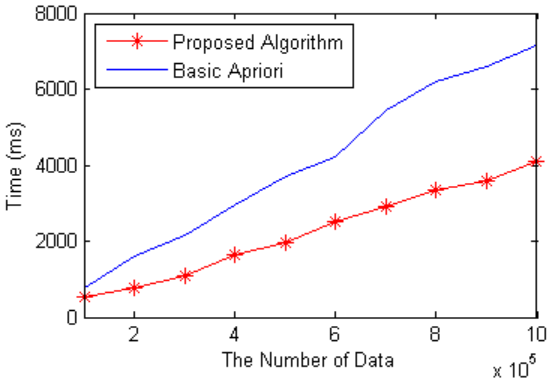

Figure 2- Optimum State

As Figure 2 illustrates, the more the number of data increases, due to division of tasks and parallelism, Increased savings we will have in the time of running algorithm.

B. The Middle State

Middle state is the situation in which the data are divided into three groups and one-third of data is given to the first distributed system and one-third to the second distributed system and the rest are given to the common system.

The table 2 shows the production data and their combination.

Table 2 - Data Set For Middel state

Item	The Number Of Items									
ab	13573	32733	13210	67950	36307	77704	32124	23212	19344	85993
a	12371	14056	1097	2510	15819	6277	95513	52639	3567	56819
abc	4717	23929	26283	30207	41666	48268	17926	65958	78338	132427
bc	959	2844	44385	56762	16424	63707	101393	50229	115655	62625
ae	4139	15060	33963	20011	17289	79909	94365	10020	44039	98518
bd	10849	6024	39859	32794	56797	80248	49907	79784	51996	61707
aef	6528	25535	15678	23663	41234	53781	18213	124686	35807	60794
bcf	5386	17135	33232	57371	34405	38512	19217	82194	107012	142966
d	10438	15327	41780	14913	77195	34440	121480	99064	68460	122346
de	14470	12349	10696	9587	64747	56760	75002	75344	126911	74537
df	7258	21552	18284	74950	27835	11872	25957	63421	125533	36199
def	9312	13456	21533	9282	70282	48522	48903	73449	123338	65069
Sum	100,000	200,000	300,000	400,000	500,000	600,000	700,000	800,000	900,000	1,000,000

ISBN: 978-981-07-5461-7

Figure 3 comparing the proposed algorithm with a basic Apriori algorithm in the Middle state

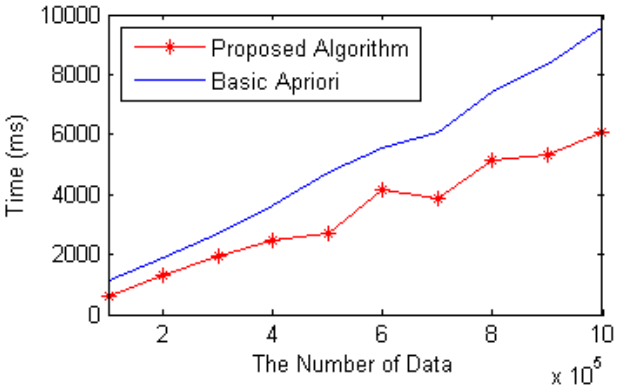

Figure 3- Middle state

C. *The Worst Case*

The worst case of algorithm is when all data are shared though two distributed systems and the data couldn't be divided between two distributed systems; hence the systems are not able to produce its own rule sets. As a results its recommended that some particular conditions set for randomly selection 10 items, if 8 of them had the shared conditions, the algorithm will turn to Basic Apriori. Entire data are common in the systems:

Table 3- Data Set For Worst Case

Data Items	The Number Of Items									
af	12289	4014	19488	4025	44635	37337	71494	109215	65781	65819
cd	432	17352	23321	35000	54374	52098	17631	107686	101096	32435
ae	9662	11228	8313	30042	54780	71394	75579	49037	106289	72837
bd	14700	13712	23969	38442	64518	1350	9568	46853	138141	25606
ce	13206	22796	5884	49409	68576	79208	73752	30689	86208	179059
bf	3149	23885	12254	52040	38224	96	11073	88464	24927	29969
ad	8060	12453	53122	50459	10144	16208	60958	12778	21295	142830
cf	7430	3742	48371	46142	8927	50078	78160	88354	102654	94650
abd	12558	5444	7827	31733	39198	73130	84937	84864	55395	44959
ade	4552	33310	18309	26250	73770	54799	41714	95672	14021	155349
bdf	6207	33657	35306	32142	32218	94259	68269	15836	149913	67905
ace	7755	18407	43836	4316	10636	70043	106865	70552	34280	88582
Sum	100,000	200,000	300,000	400,000	500,000	600,000	700,000	800,000	900,000	1,000,000

The results obtained from the execution of data set in Basic Apriori and proposed algorithm can be compared in figure 4:

Proc. of the Intl. Conf. on Advances in Computer Science and Electronics Engineering — CSEE 2013

ISBN: 978-981-07-5461-7

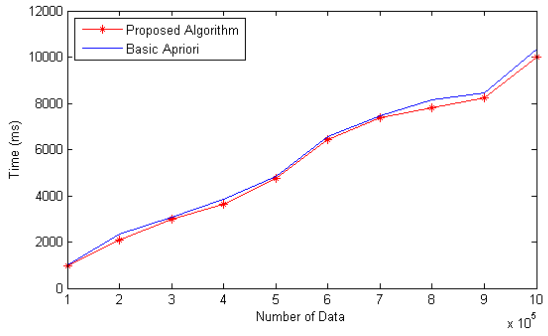

Figure 4- worst case

I. CONCLUSION

The present study was designed to determine the effect of distributing execution of intrusion detection system with data mining. In this investigation, the aim was to assess more speed to finding intrusion and increase saving time of reaction to the cyber threats because with the rapid development of Internet and network technologies, security issues also wants to highlight.

References

[1] Heady, R., Luger, G., Maccabe, A., and Servilla, M. The architecture of a network level intrusion detection system. Technical report, Computer Science Department, University of New Mexico, August 1990.

[2] Ma Xiaochun. The Research and Application of Data Mining in Network Intrusion Detection System [D]. Xi an: Northwestern Polytechnical University, 2005

[3] M. N. Mohammad, N. Sulaiman, and O. A. Muhsin, "A novel intrusion detection system by using intelligent data mining in weka environment," *Procedia Computer Science*, vol. 3, no. 0, pp. 1237-1242, 2011.

[4] S.-H. Kim, and J. R. Wilson, "A discussion on 'Detection of intrusions in information systems by sequential change-point methods' by Tartakovsky, Rozovskii, Blažek, and Kim," *Statistical Methodology*, vol. 3, no. 3, pp. 315-319, 2006.

[5] Anderson J. P., et al., "Detecting Unusual Program Behavior Using the Statistical Components of NIDES", Computer Science Laboratory SRI-CSL-95-06, 1995.

[6] E. Corchado, and Á. Herrero, "Neural visualization of network traffic data for intrusion detection," *Applied Soft Computing*, vol. 11, no. 2, pp. 2042-2056, 2011.

[7] G. Wang, J. Hao, J. Ma *et al.*, "A new approach to intrusion detection using Artificial Neural Networks and fuzzy clustering," *Expert Systems with Applications*, vol. 37, no. 9, pp. 6225-6232, 2010.

[8] J. J. Davis, and A. J. Clark, "Data preprocessing for anomaly based network intrusion detection: A review," *Computers & Security*, vol. 30, no. 6–7, pp. 353-375, 2011.

[9] S.-Y. Wu, and E. Yen, "Data mining-based intrusion detectors," *Expert Systems with Applications*, vol. 36, no. 3, Part 1, pp. 5605-5612, 2009.

[10] N. H. Park, S. H. Oh, and W. S. Lee, "Anomaly intrusion detection by clustering transactional audit streams in a host computer," *Information Sciences*, vol. 180, no. 12, pp. 2375-2389, 2010.

[11] Taylor C., Foss J. A., "NATE: Network Analysis of Anomalous Traffic Events, A Low-cost Approach", Proceedings of New Security Paradigms Workshop, New Mexico USA, pp. 89-96, 2002.

[12] Roesch M. Snort-ligh tweight intrusion detection for networks. In:Proceedings of the 13th USENIX Conference on System Administration. Seattle, Washington; 1999. pp. 229 e 238.

[13] Vallentin M, Sommer R, Lee J, Leres C, Paxson V, Tierney B. The nids cluster: Scalable, stateful network intrusio n detection on commodity hardwa re. Lecture Notes in Computer Science 2007;4637:1 07e 26.

[14] X. Wu, V. Kumar, J. Ross Quinlan *et al.*, "Top 10 algorithms in data mining," *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1-37, 2008/01/01, 2008.

[15] P. K. a. A. Ghorbani, "Research on Intrusion Detection and Response: A Survey," *International Journal of Network Security*, vol. 1, no. 2, 2005.

About Authors:

Samaneh S. Alavifar is a student of MSc in Computer engineering from Payame Noor University in Tehran, IRAN.
Her research interests are in the areas of computer networks, data mining, Network Security. She is also Member of Board of Director BStech Lim Co. Engineering Consultant in risk and Emergency and working on Passive defense in cyber and network security.

Akbar Farhoodi Nejad received the MSc and Ph.D. degrees in computer engineering from University of New South Wales, Sydney, Australia,
His research interests are in the areas of neural network and data mining.