Global Journal of Advance Engineering Technologies and Sciences

EXPERIMENTING VARIABLE LENGTH PATTERNS FOR INTRUSION DETECTION

Mr. Kulkarni Sagar S.[1], Prof. Kahate Sandip A.[2]
[1]Student, Department of Computer, [2]Assistant Prof., Department of Computer,
SP COE, Otur, Pune, India
kulsagar325@gmail.com

ABSTRACT

Due to increasing events of security attacks from last several years, security administrators were using intrusion detection system as a trustable tool to detect security vulnerabilities. There are many researches focusing on host based intrusion detection systems using system call patterns, but all of them suffer from high FPR. Many researchers on system call pattern based intrusion detection uses variable length system call patterns. This paper shows experimental result carried out with variable length patterns and finally concludes the work.

Keywords— Anomaly Detection; Intrusion Detection; Statistical based.

INTRODUCTION

The increasing attempts of security attacks forces security administrator to use intrusion detection system as a trustable tool to detect security vulnerabilities. The intrusion detection system is a specially designed tool to detect security vulnerabilities [1][2]. The detection process can be real time or offline detection. There are two types of intrusion detection system based on location of intrusion detection, HIDS and NIDS. The host based intrusion detection system uses system call patterns, audit logs, log files, CPU different parameters as a source of information. On the other hand network based intrusion detection system uses router table information, network packets, server logs for detecting intrusions. For building an IDS two approaches are used either misuse based or anomaly based. In misuse based system knowledge about specific attacks and system vulnerabilities is given in the form of signature. Misuse based system are simple to implement but writing signature to detect attacks is difficult task. Also such systems can not detect zero-day attacks or even variations of attack whose signature is given to system. The anomaly based system is uses system normal behaviour to detect anomalies. It is very powerful method and has the ability to detect unknown attacks [2][3].

The usefulness of anomaly based IDS attracted attention of many researchers, this is because this type of system does not requires signature for each attack [1][2]. The effectiveness of anomaly based system is depends on source of information, how the information is used and threshold chosen. This paper takes journey that analyzes different HIDS and their limitations, so that useful concepts from these researchers can be taken to build new HIDS that can provide high DR with small FPR.

The rest of paper is organized as follows: Section 2 covers literature review. and section 2 contains concluding remarks.

LITERATURE

One of the, initial anomaly based intrusion detection system was Haystack that uses statistical approach for detecting abnormal behaviour [4]. MIDAS [5] uses rules to detect anomalies and the rule generation process is automatic. In paper [6] Forrest et al proposed a model to categorize self from nonself in computer system. The use of system call patterns for intrusion detection is given in [7][8][9][10][11][12][13][14][15][17] with difference in normal profile generation or way of taking decision of anomaly or model used for detection purpose such as HMM, ANN, SVM, etc. The exception is Syed et al [16] that uses kernel events for detection of anomaly. Their model [16] works by calculating the probability of occurrence in normal and abnormal system call trace. The next section gives different model applied over variable length patterns.

MODELS

The models given in this section are either taken from previous researches or extension to previous researches, such that comparison can be possible for experimental setup. The experiments given in this section lets us to compare effectiveness of different models.

All given models uses Variable Length Patterns (VLP) for generating normal profile. The variable length pattern generation is given in Wespi et al [00].

A. VLPTMR (VLP and Total Mismatch Rate)

This is similar model proposed by Wespi et al [12], in this model variable length system call patterns are extracted and used as normal profile. During detection, mismatches encountered while pattern extracting process is accounted. The threshold over total mismatch rate (TMR) is used for decision making.

B. VLPLMR (VLP and Local Mismatch Rate)

This model is modification over earlier *VLPTMR* model. This model uses variable length patterns for generating normal profile database. During detection, a Locality Frame is maintained of length *l* and mismatches encountered in local region *l* are used for decision making.

This model makes use of local mismatch rate (LMR) for decision making because it has been found in earlier researches [8][9] that, anomaly occurred in burst. Also setting up threshold over for total trace length is difficult task as; length of trace can be different [9].

C. VLPLRI (VLP and Local Rarity Index)

This model uses combination of variable length patterns and Rarity-Index for anomaly detection. Initially, use of Rarity-Index was done by Vardi et al [11], but their model uses fixed length patterns. The VLPRI model initially extracts all the variable length patterns and calculates Rarity-Index of all extracted patterns. This Pattern-Rarity Index database is used as normal profile. During detection, Locality Frame is used for keeping extracted pattern Rarity-Index in local region. If at any given time, the rarity index of local region drops below certain threshold then sequence is flagged as anomalous.

The motivation for this model is taken from research done by Warrender et al [9] and Vardi et al [11], in which they found that rare sequences are anomalous, but none of the research focuses on use of Rarity-Index on variable length patterns. It can be understood that, if patterns encounter in detection phase are rarely used then there is possibility that attack is in progress. This fact is taken from researches on human behaviour, in which it is found that a human normally uses only those commands that are popular in community. To compromise the computer system, attacker executes a commands sequence that is unknown in community or very rarely used, thus making detection possible.

D. VLPLMRRI (VLP- Local Mismatches And Rarity-Index)

This model uses variable length patterns, local mismatch rate and Rarity-Index for anomaly detection. The emerging need of improving DR with reduced FPR brings is motivation behind this model. It tries to provide greater control over DR and FPR. This model uses Pattern-Rarity Index dictionary as normal profile. During detection, two Locality Frames are used one holds Rarity Index of encountered patterns and other holds mismatches encountered during forming given sequence (call it as struggle while forming next sequence). Two thresholds t_1 and t_2 are set on to the locality frames so as to categorize system call traces.

E. SAVLP (Semantic Analysis of VLP)

The semantic relationship is used in this model for detecting anomalies. The semantic analysis of system call patterns is initially done by Creech et al [17], but the definition of word in there model is inappropriate and training time required for dictionary generation is very large. This model is modification over Creech et al [17] model, in which VLP are considered as words and phrases of different length are extracted from training data by combining adjacent patterns. Finally Rarity-Index of all extracted phrases is calculated and maintained in Phrase-Rarity Index dictionary. This Phrase-Rarity Index dictionary is used as semantic information of program system call patterns.

During detection, the patterns from test trace are extracted and stored in locality frame. The average Rarity Index of all seen phrases in locality frame is used for decision making purpose. The hypothesis behind this model is that, if all the possible combinations of VLP are known then it must be possible to build the given test sequence. If locally seen phrases are not present in Phrase-Rarity Index database then rarity index is assumed to be 1. Thus dropping the average Rarity-Index of phrases seen in local region (most popular phrases have Rarity-Index -1).

RESULTS

The evaluation of given models is done using UNM intrusion detection dataset. The processes extracted from UNM dataset are, login, ps, ftp, lpr. The experimental settings for evaluating above systems are kept similar, so that results from different models can

ISSN 2349-0292
Impact Factor (PIF): 2.365

be compared. The experimental settings are given in TABLE 1.

TABLE 1 EXPERIMENTAL SETTING

Process Name	Total Normal Traces Extracted	Number of Normal Training Traces	Number of Test Traces	
			Normal	Attack
Login	9	4	5	12
ftp	7	4	3	5
lpr	5	3	2	1001
ps	24	20	4	26

According to requirements of each experiment, normal profile for each program is generated and detection performed for different threshold levels. Finally for each system ROC curve is plotted to understand relationship between DR and FPR rate.

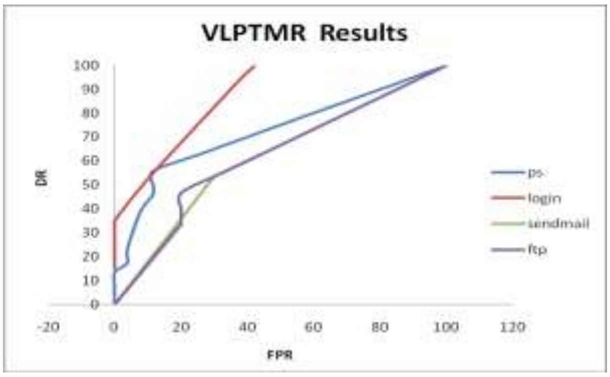

FIGURE 1. VLPTMR MODEL RESULTS

The experimental results of experiment 1 are shown in figure 1. It shows that the best DR is 40%, as threre was no FPR at perticular DR.

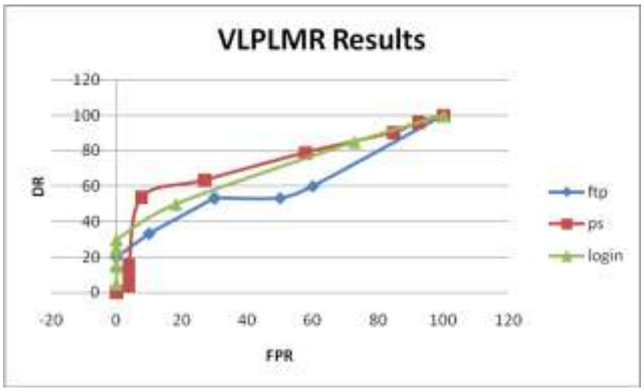

FIGURE 2. VLPTMR MODEL RESULTS

Results of experiment 2 are provided in figure 2, which shows improved results as compared to experiment 1. The reason is that it uses local mismatch rate for detecting anomalies.

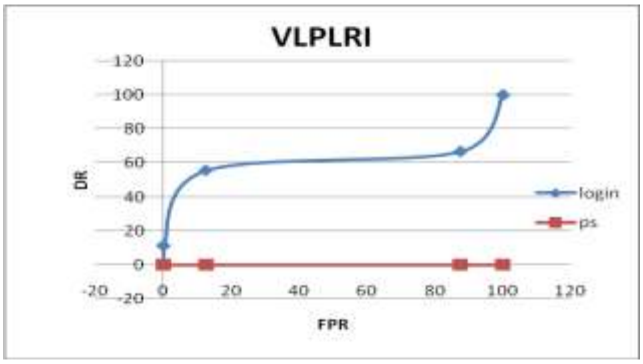

FIGURE 3. VLPLRI MODEL RESULTS

The experiment 3 result are conducted for process ps and login, it shows very poor performance for local region length of 6. The problem with using rarity index is that, Rarity-Index is useful only when there is complete training data or huge training data.

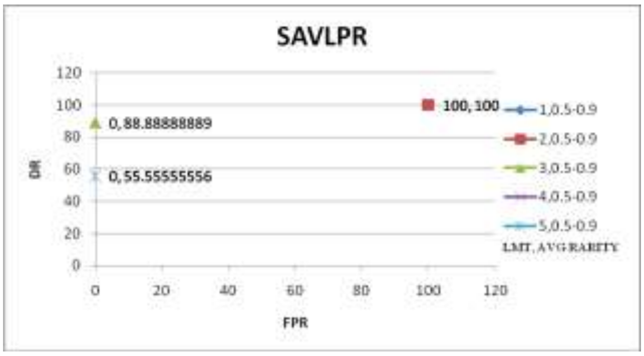

FIGURE 4. SAVLPR MODEL RESULTS

The experiment 4 results for login process are shown in figure 4. It shows that after using semantic analysis concepts with variable length system call patterns, the detection rate reached up to 88% with FPR 0%. The results The results other processes are not tested.

CONCLUSION

It has been found that variable length patterns are useful for reducing the dictionary size required to profile normal activity of a program. But, it is also found while experimenting that, pattern extraction using method proposed by Wespi et al [12] does not cover all system call patterns, this is due to selection of longest pattern while pattern matching. It is also found that, the use of single threshold for intrusion detection

ISSN 2349-0292
Impact Factor (PIF): 2.365

generates FPR, therefore in future systems there should be use of more than one threshold. Finally, the intrusion detection system must make use of approximate pattern matching to reduce FPR, encountered due to incomplete training.

REFERENCES

[1] John McHugh, Alan Christie, and Julia Allen, ”The Role of Intrusion Detection Systems, IEEE SOFTWARE,SEP 2000.

[2] Mehdi Bahrami and Mohammad Bahrami, “An overview to Software Architecture in Intrusion Detection System”, Soft Computing And Software Engineering (JSCSE), 2011.

[3] Herve Debar, “An Introduction to Intrusion-Detection Systems”, IBM Research, 2011.

[4] Debra Anderson, Teresa F. Lunt, Harold Javitz, Ann Tamaru, Alfonso Valdes, "Haystack: an intrusion detection system”, Aerospace Computer Security Applications Conference, Oct. 7481, Dec 1988.

[5] M. M. Sebring, E. Shellhouse, M. E. Hanna, and R. A. Whitehurst, "Expert systems in intrusion detection: A case study", Proceedings of the 11th National Computer Security Conference, Oct. 7481, 1988.

[6] Stephanie Forrest, and Alan Peterson, "Self - Nonself Discrimination in Computer”, Proceeding of 1994 IEEE Symposium on Research in Security and Privacy, 1994.

[7] S. Forrest,S. A. Hofmeyr and A. SoMayaji, ”A sense of self for Unix Processes”, IEEE Symposium, May 1996.

[8] S. Forrest,S.A. Hofmeyr and A. SoMayaji, ”Intrusion Detection Using Sequences of System Calls”, IEEE Symposium, May 1996.

[9] C. Warrender, S. Forrest, and B. Pearlmutter, ”Detecting intrusions using system calls: alternative data models”, Proceedings of the 1999 IEEE Symposium,1999.

[10] A. Somayaji and S. Forrest, "Automated Response Using System-Call Delays.", Proceedings of the 9th USENIX Security Symposium, The USENIX Association, Berkeley, 2000

[11] Wen-Hu Ju and Yehuda Vardi, “Profiling Unix Users And Processes Based On Rarity of Occurrences Statistics with Applications to Computer Intrusion Detection”, Fourth Aerospace Computer Security Applications Conference, October 1988

[12] John Andreas Wespi and Herv Debar, “An intrusion detection system based on the teiresias pattern discovery algorithm”, Proceedings of EICAR, 1998.

[13] Wenke Lee and Salvatore J. Stolfo , "Data Mining Approaches for Intrusion Detection",7th USENIX Security Symposium, Jan 1998.

[14] Xuan Dau Hoang, Jiankun Hu, Peter Bertok, "A Multi-layer Model for Anomaly Intrusion Detection Using Program Sequences of System Calls" ,IEEE, October 1988.

[15] Ye Du, Ruhui Zhang, and Youyan Guo, "A Useful Anomaly Intrusion Detection Method Using Variable-length Patterns and Average Hamming Distance”, Journal of Computers, Aug 2010.

[16] Syed Shariyar Murtaza, Wael Khreich, Abdelwahab Hamou-Lhadj, Mario Couture, "A Host-based Anomaly Detection Approach by Representing System Calls as States of Kernel Modules", IEEE 24th International Symposium on Software Reliability Engineering (ISSRE), 2013.

[17] G. Creech and J. Hu.,”A Semantic Approach to Host-based Intrusion Detection Systems Using Contiguous and Dis-contiguous System Call Patterns”, IEEE Transactions on Computers, 2014.