International Journal of Electrical and Computer Engineering (IJECE)
Vol. 3, No. 5, October 2013, pp. 708~712
ISSN: 2088-8708

Intelligent Detection of Intrusion into Databases Using Extended Classifier System

Navid Moshtaghi Yazdani, Masoud Shariat Panahi, Ehsan Sadeghi Poor
Department of Mechatronic, University of Tehran

Article Info

Article history:

Received Apr 17, 2013
Revised Jul 18, 2013
Accepted Aug 12, 2013

Keyword:

Data mining
Extended classifier systems
Intrusion detection system
Learning agent
Security of databases

ABSTRACT

With increasing tendency of users to distributed computer systems in comparison with concentrat-ed systems, intrusion into such systems has emerged as a serious challenge. Since techniques of intrusion into systems are being intelligent, it seems necessary to use intelligent methods to en-counter them. Success of the intrusion systems depends on the strategy employed in these sys-tems for attack detection. Application of eXtended Classifier Systems (XCS) is proposed in this paper for detection of intrusions to databases. The extended classifier systems which are known as one of the most successful types of learning agents create a set of stochastic rules and com-plete them based on the methods inspired from human learning process. Thereby, they can grad-ually get a comprehensive understanding of the environment under study which enables them to predict the correct answer at an acceptable accuracy once encountered with new issues. The method suggested in this paper an improved version of extended classifier systems is "trained" using a set of existing examples in order to identify and avoid attempts to intrude computer sys-tems during phases of application and encountering these attempts. The proposed method has been tested on several problems to demonstrate its performance while its results indicate a 91% detection of various known intrusions to the databases.

Corresponding Author:

Navid Moshtaghi Yazdani
Department of Mechatronic
University of Tehran,
Amir abad Road, Tehran, Iran.
Email: navid.moshtaghi@u.ac.ir

1. INTRODUCTION

Huge volume of valuable data in databases and extensive demand to access these data have led to increased number of attempts to intrude the databases. This has rendered secu-rity of the databases throughout the world to a priority. Communication of different net-works increases the possibility of intrusion into these networks, while it connects more people to them. Therefore, intrusion detection systems are created to identify risks, which are now being used in computer networks and databases extensively in order to protect information against suspicious attacks. Security of computers and databases are of signifi-cant importance for all people, thus great attempts are currently being done to provide in-trusion detection and methods to encounter. Expectations of intrusion detection systems are increased by recent technological advancements and some intrusion detection systems utilize learning methods to model the intrusion. The database systems of organizations are accounted for the main technology of data management to access and store their infor-mation. Thus, security of data managed by these systems would be critically important. The security mechanisms which are associated with encryption, access control and firewall are defeated against attacks in protection of sensitive information. In this regard, em-powerment of the intrusion detection systems is emphasized to keep the information safe against unauthorized users. The intrusion detection is a protective system which identifies occurrence of malfunctions. Anomalous traffic and effort to intrude the

network can be detected and logged in these systems by using the information extracted during data collection, port scans, taking control of computers and finally hacking them. Generally speak-ing, intrusion can be defined as a set of unlawful activities of an individual which endan-gers privacy or accessibility of resources of someone else. The intrusion can be divided into internal and external categories. The internal intruders are authorized users of the or-ganization with specific extent of access to resources of the system. They thus have some information about the configuration of system and database and also program interface. On the other hand, the external intrusions are committed by intruders who are not author-ized ad are unfamiliar with the organization and its security issues. The internal intruders are principally accounted for more critical threats for the database system.

2. CLASSIFICATION OF VARIOUS INTRUSIONS TO COMPUTER NETWORKS

Network attacks can be categorized into two following groups considering how they are performed: Service interruptionattacks in which a great volume of service requests are sent to the server to interrupt its service function; and access attacks to the network which enable unauthorized access to resources of the network. Most current intrusion detection systems operate based on the data generated by auditing mechanisms of the operating system or network. One of the most famous network-based intrusion detection systems is *Snort* which has two characteristics of package surveillance and package record. In addition to these two main characteristics, the network-based intrusion detection system support prompt transmission of alerts and can even be utilized as a system to prevent intrusion of online IPs. Limitations and drawbacks of the network-based intrusion systems are summarized below:

- Inability to detect staged intrusions
  Since many intrusions are implemented by sending and receiving information packages in several stages [1], and because the main structure of the staged intrusion detection system decides only based on an information package received by its detection motor, it is unable to identify and detect these attacks and intrusions.
- Inability to collect all packages
  Network-based intrusion detection systems need to examine all packages forreaction toward intrusion detection.Therefore, these systems are unable to perform properly against heavy traffic of the network and reduce performance of the network.
- Inability to detect anomaly during package fragmentation
  Network-based intrusion detection systems have the ability to identify each package individually regardless of the network topology. One major drawback of these systems is their inability to store status of existing nodes in the network. Thus, they are unable to identify and detect anomaly of the network and new attacksthe rules of which have not been generated yet.
- Increased traffic
  Network-based intrusion systems need to control all packages in order to react towards attack identification. They would be unable to perform properly when there is heavy network traffic, so performance of the network will be reduced.
- Inability to prevent service provision
  Language of rule for network-based intrusion detection systems are defined according to each transmission package of the network. When network traffic is high this system would be unable to detect some attacks including prevention of service provision. Moreover, these systems operate with systemic files and commands with the mapping between files in the operating system and interfaces/pages in the database being inaccurate. As a result, it cannot properly demonstrate user's behavior. Furthermore, evaluation of the user's behavior within operating system is not adequate to detect intrusion for the database because definition and structure of the data is not shown in the log files and it is required to prepare them in the database. Therefore, the main reasons for need to an intrusion detection system special for database management can be introduced as below:
  1. An activity which seems harmful for the database is not necessarily harmful for the network and the operating system. So the intrusion detection systems of network and operating system are not much robust and effective against the attacks to database.
  2. Extensive application of databases and database managements.

3. LITERATURE REVIEW

A great deal of researches has been conducted on intrusion detection systems during the two decades passed from introduction of these systems for the first time. However, most of these studies have led to create systems which are just operated at the level of network and operating system since they are unable to detect intrusions at the level of databases as discussed before. Some of these researches will be addressed here.

Wenhui et al. designed a two-layered mechanism for intrusion detection in a web-based database service [2]. Behavior of each data source is modeled from web servers and database systems using their raw log files. Pre-alerts received from the first layer are transmitted to the second layer which is responsible to explore and find correlations between them. The main idea utilized here is that the anomaly of each data source cannot be enough evidence for intrusion.

Hashemi et al. improved this work from two aspects [3]. First, they developed the concept of harmful transaction. Every transaction made on data item without authorized access or change is referred to as harmful. Second, variations on each data item enhance the detection rate by extraction of temporal patterns.

Hu has used dependency of the data items at transaction level to detect suspicious transactions [4]. Afterwards, dependency rules are developed from the dependencies iterated for several times. The transactions which do not obey each of the dependency rules are considered as suspicious.

The concept of data item dependency was further improved by Srivastava et al. [5] who have also studied data sensitivity. Some data items are more sensitive than others against suspicious manipulation.They introduced a weighted data mining algorithm to find dependency among the sensitive data. A transaction is called suspicious when it refuses to follow these rules.Nevertheless, the main problem in data mining algorithm is the dependency of identification properties with adequate values of support and confidence. Those properties which have been less referred to might never appear in the correlation rules. Application of weighted data mining algorithm will mitigate this effect to some extent, though will not resolve the problem completely.

DEMIDS is another intrusion detection system. DEMIDS is a abuse detection system specially designed for relational databases. It is known as one of the first and valuablecontributions in this field. It has been assumed in this system that the users commit no anomaly during learning. Sufficient data must be collected in order to consider all possible states. Then the database extracts users' profile using the log files. The user's profile shows normal behavior of him/her and describes his/her access to the database. Each security policy-based profile contains a set of important features which are selected by the security administrator. The main idea of using such a system is that the working spacewhich is comprised of a set of characteristics and properties forms the user's access patterns. The working place is obtained by extraction of frequent item sets. These are in fact a set of features with given values which aregenerated based on data structures, integration limitations in system catalogues and user's behavior saved in the log files.

Prompt database intrusion detection was first introduced by Lee et al. [6]. This model observes behavior of the database at the level of sensitive transactions. Data are updated periodically in the prompt databases and an unexpected data update will be alerted.

4. PROPOSED METHOD

A system based on machine learning is proposed in this research to identify attacks to da-tabases. Machine learning covers a wide variety of learning algorithms either supervised or unsupervised, the aim of which in the field of data mining is to avoid exhaustive search of the data and to replace such time-consuming searches with intelligent methods that fa-cilitate classification or modeling behaviors of the data trough finding the existing patterns among the data. Several techniques have been presented during the last two decades in the field of data mining which have utilized supervised, unsupervised and/or reinforcement algorithms for pattern detection and allocation. Classifier systems are one of the most suc-cessful of them. Generally speaking, the classifier systems involve a set of rules with "if-then" format in which each rule provides the target problem with a potential solution that is updated by genetic algorithm. During this gradual evolution, the system learns behavior of its environment and then gives proper answers to the queries asked by the user. The first classifier system entitled "Learning Classifier System (LCS)" was proposed by Holland in 1976. In this system, the value of each rule was evaluated with an index called "strength". The strength of a rule was increased in proportion with its correct answers to the training examples within reinforcement learning regulations. Meanwhile, an evolutionary search algorithm (usually genetic algorithm) was appointed responsible for generation of new rules and elimination of inefficient rules in specific periods of time. At the end of training, this set of rules was relatively able to provide acceptable solutions encountering new ques-tions. At the same time, the successful performance of LCS depends on selection of ap-propriate values for

control parameters of the system which were mainly dependent on de-signer's experience. Some other types of the classifier systems have been suggested since introduction of LCS, eXtended Classifier System (XCS) is one of them. The ability of these systems was considerably limited before development of XCS occurred as early as 1995. Nevertheless, they were gradually mutated into more intelligent and accurate agents such that XCS and its improved versions are currently believed to solve complicated prob-lems without further adjustment of the parameters. By development of eXtended Classifier Systems Real valued (XCSR), some intrinsic drawbacks of the binary classifier systems such as their inability to introduce specific ranges of variables were solved to a large extent. These systems are now one of the most successful learning agents for solving data mining problems in partially-observable environments. In spite of the capabilities of XCS, this system has also some limitations. For instance, re-alistically determining chance of each rule for survival and participation in the process of generating new rules requires using a great deal of training data.Thus, XCS training pro-cess is often time-consuming and expensive. A new method is proposed in this paper to enhance speed and rate of convergence for XCS training process which will be discussed in the next section.

5. IMPROVED XCS ALGORITHM

The finite set of training data is used in the proposed method to modify features of the rules (i.e. "prediction", "prediction error" and "fitness"). This is done by using the following equations:

Updating prediction and prediction error:

If $exp_i < 1/\beta$ then $Pi = Pi + (R-Pi) / exp_i$, $\varepsilon i = \varepsilon i + (|R-Pi| - \varepsilon i) / exp_i$

If $exp_i \geq 1/\beta$ then $Pi = Pi + \beta (R-Pi)$, $\varepsilon i = \varepsilon i + \beta (|R-Pi| - \varepsilon i)$

Updating fitness:

If $\varepsilon i < \varepsilon 0$ then $ki=1$

If $\varepsilon i \geq \varepsilon 0$ then $ki = \beta (\varepsilon i/\varepsilon 0)^{-\gamma}$

$Fi = fi + \beta [(ki/\sum kj) - fi]$

Where, β is rate of learning and γ is power of rule accuracy. ε, exp and P denote prediction error, rule experience and rule prediction, while R, k and f represent premium received from the environment, rule accuracy and rule fitness, respectively. Index i shows rule number in the set of rules.

In the next stage and to expand variety in the data set, several pairs are selected as parents using"stochastic selection with remainder"from strings which display the conditional part of the existing data. The conditional part of the new data is created via intermediate crossover method which is applied on these parent strings. In this method, the value of each conditional variable is calculated from the equations below:

$a_i = \alpha(a_i^F) + (1-\alpha)(a_i^M)$

Where, a_i is value of the conditional variable, a_i^F and a_i^M denote values of the i^{th} conditional variable in the first and second parents (father and mother), respectively. $\alpha$is participation coefficient of the parents which is determined adaptively. Performance part of the new data is also generated using a nonlinear mapping from the space of conditional variables to the space of performances which is created by the existing data.

Diversification of the existing data is continued as long asthe termination condition of learning is satisfied using the completed data (e.g. when correct answers of the system to the test data reaches a pre-defined threshold percent) [7].

6. APPLICATION OF IMPROVED XCS FOR DETECTING INTRUSION TO DATABASE

The first and most important action in detecting intrusion to the database is determination of the existing significant characteristics in the packages sent to the network. For this pur-pose, source address, destination address, source point, destination port, type of package and number of data bytes must be extracted from the packages. Moreover, connection rec-ords must be formed according to all the packages sent to the network. Each connection record contains some information including time of starting connection, duration of con-nection, number of packages sent from each side, status of connection termination and re-quested service of connection. The proposed method is expected to reduce these limita-tions. In this method each rule is defined as "then-if", where the conditional part includes the following items:

1. Is the data updated periodically or not?
2. The transaction follows dependency rules created.
3. Are the changes recalled according to temporal pattern or not?
4. If the user is internal, is the rules of internal user the same as the one applied? And is the level of organization observed in this transaction or not?

5. The similarity between new transactions and normal profile, where the normal pro-file is created by a sequence of correlated transactions.
6. The transaction obeys important policies which are applied. The "then" part of the condition also involves identification of intrusion to the database.

The detection stage of this research is implemented offline, which means that the existing information in the log files are for detection. Standard DBMS log files store basic infor-mation from surveys and executed transactions in the database. In other words, logs of the database which are stored in a file are examined and the file is then analyzed. These logs are comprised of some 13,360 records. Since one needs data for both training and test in order to train XCS, some 12,000 entries of the abovementioned data are considered for training with the other 1360 entries being considered as the test examples. The results ob-tained from application of the trained XCS on the test data sets are listed in the table below.

Table 1. Detection of attack to database using improved XCS

Number of test examples	Number of correct answers	Percentage of correct answers	Number of incorrect answers	Percentage of incorrect answers
1360	1243	91.39%	117	8.602%

7. CONCLUSION

Taking into consideration the ever-increasing application of the distributed computer sys-tems and the necessity to use intelligent methods for intrusion detection and prevention, a new method was proposed to identify attacks to the databases. This method was based on using an extended classifier system (XCS) which after being trained using a set of existing examples and reinforcement learning techniques will be able to identify attempts conduct-ed to intrude the databases and provide preventive incentives against them. Performance of the suggested method was assessed by application of it on a sample problem and its results were presented. Comparison of the obtained results is indicative of a 91% detection rate for various types of known attacks to the databases.

REFERENCES

[1] H Sadoughi Yazdi. "Determining Behavior of Classifiers byDecision Stereotype based on Markov's Hidden model". *Journal of Electrical Engineering and Com-puter Engineering of Iran*. 2008; 6(2).
[2] S Wenhui, D Tan. "*A Novel Intrusion Detection System Model for Securing Web-based*". Proceedings of the *25th* International Computer Software and Applications Conference on Invigorating Software Development, ISBN: 3-1615-1312-1; 241: 2331.
[3] S Hashemi, Y Yang, D Zabihzadeh, and M Kangavari. "Detecting intrusion transactions in databases using data item dependencies and anomaly analysis". *Expert Systems*. 25(5), 2338.
[4] Y Hu, B Panda. "*A Data Mining Approach for Database Intrusion Detection*". Proceedings of the ACM Symposium on Applied Computing. 111-116 2334.
[5] A Srivastava, S Sural, AK Majumdar. "*Database Intrusion Detection using Weighted Sequence Mining*". ISSN: 1116233X: 8-11, 2336
[6] V Lee, J Stankovic and S Son. "*Intrusion detection in real-time databases via time signatures*". In Proceedings of the IEEE Real-Time Technology and Applica-tions Symposium (RTAS), 2333.
[7] M Shariat Panahi, N Moshtaghi Yazdani. An Improved XCSR Classifier System for Data Mining with Limited Training Samples. *Global Journal of Science, Engineering and Technology*, (ISSN: 2322-2441). 2012; 2: 52-57.