Intrusion Detection System Based on User Behavior Using Data Mining Techniques

Hind Alsharif, Omar Batarfi
Computer Science Department
Faculty of Computer and Information Technology
King Abdul Aziz University,Jeddah, Saudi Arabia

Abstract **- Intrusion Detection System (IDS) in computer technology is a little bit different than physical intrusion detection system, which detects any physical changes in the protected premises. In computer technology, IDS is used to examine all traffics and activities either in computer unit or network. These IDS could be old technology systems, which we refer in this paper as traditional Intrusion Detection System (tIDS), or it could be an intelligent system based on AI, machine learning, data mining and other intelligent techniques. In tIDS, which based on errors detection, the system works according to its database. This database is usually predefined by security experts. IDS is used to classify suspicious behaviors as intrusion acts or regular activities. Experts update the database manually [1]. Thus, it is hard to keep track of every single update and hard to analyze an event as a suspicious act with acceptable efficiency and satisfaction. So the need for automated tools became immanent to support security experts. Such a support could be achieved using data mining techniques as one of the possible ways to automate the system. This may handle the problem with high degree of accuracy.**

This paper demonstrates the advantage of using data mining techniques in IDS. The system depends on users' behaviors in order to extract features and then generate rules. The generated rules will be used as a pattern recognition tool. These rules enable the system to classify any irregular activity as an intrusion act.. In this research we hypothesize that, depending on time of a day and location of the activity in the database we could classify a suspicious behavior as an intrusion acts. The experimental results show high level of accuracy, efficiency, robustness where the system can handle errors, scalability where we can use the system in large number of users as well as reliability as the system shows 0% error rate of this technique

Keywords **- Data mining, intrusion detection system, pattern recognition.**

I. INTRODUCTION

At first there is a must to define what is IDS and what are its limitations. Intrusions are considered as "any set of actions that attempt to compromise the integrity, confidentiality, or availability of a resource" [2]. The solution is defiantly to find a way to detect this intruder, IDS is one of these ways. IDS is designed to detect intrusions' behaviors. It collects data from certain nodes in a network or a system. This data is used in order to discover whether the network or the system's behavior violates the security policies or not [3]. Most IDSs are based on hand-crafted signatures that are developed by manual encoding of expert knowledge. These systems compare the activities of the system being monitored with pre-defined signatures. The major problem with this approach is that IDS fails to detect new attacks or attacks without pre-defined signatures [4].

Recently, there has been an increased interest in data mining based approaches to build detection models for IDSs [3]. The data mining technology has a great advantage on the way of extracting the characteristics and rules from the data. This might help in the intrusion detection [2]. Data mining is " exploration and analysis, by automatic or semi-automatic means of large quantities of data in order to discover meaningful patterns" [5]. Data mining is used to process amounts of data stored in a system repositories. There are various useful data mining algorithms that can be used in IDS, such as correlation analysis, sequence analysis, classification or clustering... etc. Correlation analysis algorithm can be used to infer relationships of attributes in a network connection record. Sequence analysis algorithm, however, can discover the timing relationship of network connection records. Correlation analysis and sequence analysis of the data mining algorithm are used to obtain the regular pattern. These two algorithms are used in anomaly intrusion detection. Classification is a data mining algorithm that may create rules from trained data, which can identify normal behavior and intruders [3]. Clustering on the other hand is a non-supervised learning to cluster records according to their attributes and properties.

The proposed method works in a way where it analyzes the users' behaviors. These behaviors are represented as average login times (the time when the user login to the system) and durations (the time a user may spend) e.g. when does a particular user usually login into particular attributes in the database? And how long does he\ she usually spend manipulating these attributes? The behaviors classification method is achieved based on the reports for each user. The reports can be monthly or yearly as the organizations needs. They consist of the user ID number and the average of the timings for each department. We had to write a code to generate formatted reports by reading the activities from the source and then writing them to a file. After achieving this classification, we can extract rules based on the features for each class. However we have three classes of employees: administration, finance, and human resources (HR). Each one of these classes will have their own rule known as pattern to be used in pattern recognition approach to identify any abnormal behavior that may break those rules using several data mining techniques. Users activities are going to be preprocessed so it can be ready to go through features extraction, rule generation, classification, and then pattern recognition. Experimental results show that the proposed system accurately identifies if the user is an intruder or not.

II. Related Works

Recently, researchers have explored data mining as a new valuable and reliable approach to the IDS [6]. They are exploring the chance to improve data reduction, discovery and detection capabilities covering hidden patterns, deviations of known or unknown attacks, and maximizing the cost-benefit relationship for an ID deployment.

In 2008, Jiuhua presented a framework in his paper [1] to show how to effectively find normal and abnormal behavior from data using intrusion detection and how to effectively generate automatic intrusion rules. To accomplish the task the author shows that there is a need to study various data mining algorithms such as correlation analysis algorithms, sequence analysis algorithm, classification algorithms, etc. The system uses data mining technology to analyze databases, it collects complete network data such as protocol type and IP addresses and extracts the related behavior characteristics and rules, then the author suggested creating detection model to detect the actions. System status and corresponding behavior rules made through data mining are described based on ontology for sharing rules. Complete results are stored in rules library. Analysis model created by data mining engine and rules from rules library, then the conclusion is sent to decision-making center. Decision-making center judges network behavior and carries out corresponding solution scheme.

In [7], Wenjun divided data into three categories: intrusion data, normal data and unknown data. Because most of the network data is normal data, the author had filtrated out this data before using any mining technique. The system is designed as follows: normal intrusion detection engine filter out the normal data and intrusion data, then distributing intrusion detection engine detects unknown data. Lastly, all detection results are sent to the decision-making routine. The paper has to improve data mining's efficiency by decreasing data mining quantity, add normal intrusion detection in order to improve real-time, and structure a new IDS. This IDS integrates normal intrusion detection technology and distributes intrusion detection technology. Normal intrusion detection improves the system's real-time greatly, and distribute intrusion detection detects new intrusion effectively.

In their proposed system Leu and Hu [8] introduced Intrusion Detection and Identification System (IDIS) that mines log data to identify commands and their sequences that a user habitually submits and follows as the user's forensic features. When an unknown user logs into a computer, the IDIS starts monitoring the user's input commands to detect whether he or she is issuing an attack. The results show that the recognition accuracy of the students of the computer science department is up to 98.99% with extremely low error rate.

In our paper the system checks the data, the login time, the duration for each attribute times and pay attention to the attribute regardless of the contents, this data are going to be filtered together at the same time (please refer to figure 1). The data mining engine works using classification and pattern recognition approaches in order to obtain the results. Those results will consider specific user as an intruder or not.

III. Methodology

The objective of this paper is to study employee's behaviors who work in a specific organization. The proposed method works in a way where it analyzes the behavior for each person in the organization and these behaviors will be studied to identify his or her unique and distinguished working habits. Obviously, in our study the employees work in different departments, which are represented in the database as attributes. These departments or attributes could be administration, finance, HR, etc….The behaviors classification method is achieved based on the transactions reports which we created randomly to represent a company portfolio and activities reports (departments folders accessing times and durations) for each employee. These activities are inter-related by each employee's accounts, access times, access durations, and information access habits. Once the classification is completed, we can define rules based on the features of each department to identify any abnormal behavior that may break those rules using several data mining techniques.

The employees' activities are concerned with their access to the department's data folders on their online transaction system, where each department is restricted in access to its own folders according to the department restriction needs. HR for instance, is not allowed to access profits folder whereas the finance can. Folders accessibility is presented in table 1

Table 1: folders accessibility for the employees

Department	Salary	Names and private information	Bonuses	Profits	Holidays
Human resource	No	Yes	No	No	Yes
Finance	Yes	No	Yes	Yes	No
Human resource administration	Yes	Yes	Yes	No	Yes

The employees' records have to go through pre-processing technique in order to be ready for feature extraction and patterns recognition, and then some rules might be constructed. After the rules are generated the system can examine the organization's data and detect the intruders accordingly. System framework is shown in figure 1.

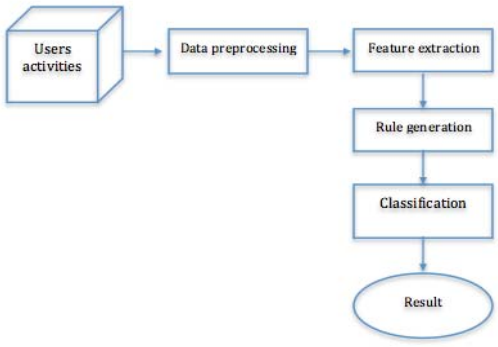

Figure 1: Data mining engine framework

IV. IMPLEMENTATION

To accomplish this task, we had to write two programs. The first code performs the pre-processing function, whereas the second one reads the clean (pre-processed) data, then classifies employees according their departments then compares their records to the extracted rules.

A. Data pre-processing:

Data pre-processing is a data mining tool to prepare data to be ready for processing [9]. It can be done by selecting valid attributes for the analysis and ignoring any attribute that might be irrelevant to the data analysis phase. This process is called task-relevant data.

According to data mining concepts [9], sometimes, during data analysis and data pre-processing, the data you wish to analyze by data mining techniques maybe:

1- Incomplete (lacking attribute's values or certain attributes of interest).
2- Noisy (containing errors, or outlier values that deviate from expected).
3- Inconsistent in contents with other databases.

Incomplete, noisy, and inconsistent data are commonplace properties of large real world databases and data warehouses. Incomplete data can occur for a number of reasons, such as:

- Data may not be included simply because it was not considered important at the time of entry.
- Data that were inconsistent with other recorded data may have been deleted.

Missing data, particularly for tuples with missing values for some attributes, may need to be inferred. There are many possible reasons for noisy data such as: [9]

- The data collection instruments used may be faulty.
- There may have been human or computer errors occurring at data entry.
- Errors in data transmission may also occur.

Data cleaning routines work to "clean" the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies. If users believe the data are dirty, they are unlikely to trust the results of any data mining that has been applied. Furthermore, this data may cause confusion for the mining procedure, resulting in unreliable output or logical errors. [9]

In [9] the outlier is defined as "Outlier is very often; there exist data objects that do not comply with the general behavior or model of the data. Such data objects, which are obviously different from or inconsistent with the remaining set of data, are called outliers". Outliers can be caused by measurement or execution error. For example, the display of a person's age as "−999" could be caused by a program default setting of an unrecorded age.

Many data mining algorithms try to minimize or eliminate the outliers. But this could result in the loss of important information hidden. In other words; the outliers may be of particular interest, such as in case of fraud detection, outliers may indicate fraudulent activity. Thus, outlier detection and analysis is an important data mining task, known as outlier mining. The outlier-mining problem can be divided under two sub problems [9]:

1. Define what data can be considered as inconsistent in a given data set,
2. Find an efficient method to mine the outliers so defined.

In the proposed system, there might be some outlier values, for example, a finance employee in some cases may access a profits folder for a critical reason in unusual time according to his behavior record, so the access time would be unexpected and considered as outlier. In the proposed system, it will not affect the record value at all, because we are considering average numbers so the difference would be in minutes. This exception or outlier is already handled in the system where in matching the rule with the user's record, it concerns only with the hours part and ignores the minute part. Therefore, from what have been described earlier, a useful pre-processing step is to run our data through data cleaning routines.

B. Feature Extraction:

This part deals with extracting attributes' values such as time. Each folder has five attributes. Salary folder for instance, has the following attributes:

- Earliest possible access time.
- Latest possible access time.
- Minimum duration.
- Maximum duration.
- Average duration.

These periods are extracted for each employee to be used later in the rule generation process.

C. Rule generation:

Calculating the feature average number for all employees under their departments, it is our proposed method to generate the rules. Therefore, all of the employee's timings have to be within the range of the calculated average. Because as we mentioned before the values here are averages and the average is calculated by dividing the sum of timings for all the employees by the number of the employees. So the resulted number will be the average number or time that the employees are usually will be on the system. We introduced tolerance when comparing features with the rules to solve the problem. This tolerance allows rules to accept some differences in minutes but not in hours. This average equals to zero when there is no access to a folder.

For more details about rules, we first define user record contents. Each record has 26 values explained in table 2.

Table 2: user record contents

Value	Refers to
1	user ID number
2	earliest possible access time to the salary folder
3	latest possible access time to the salary folder
4	minimum duration spent in the salary folder
5	maximum duration spent in the salary folder
6	average duration spent in the salary folder
7	earliest possible access time to the staff names folder
8	latest possible access time to the staff names folder

Value	Refers to
9	minimum duration spent in the staff names folder
10	maximum duration spent in the staff names folder
11	average duration spent in the staff names folder
12	earliest possible access time to bonuses folder
13	latest possible access time to bonuses folder
14	minimum duration spent in bonuses folder
15	maximum duration spent in bonuses folder
16	average duration spent in bonuses folder
17	earliest possible access time to profits folder
18	latest possible access time to profits folder
19	minimum duration spent in profits folder
20	maximum duration spent in profits folder
21	average duration spent in profits folder
22	earliest possible access time to holidays folder
23	latest possible access time to holidays folder.
24	minimum duration spent in holidays folder
25	maximum duration spent in holidays folder
26	average duration spent in holidays folder

The first value, which is the user ID, is used in the classification process, which classifies the user to his\ her referring department (HR, finance, or administration). The rest 25 values are used in the rule generation process. To generate rules out of the attributes values we described before, the average value is calculated for each attribute for all users according to their department. As we mentioned before these average numbers describe the overall timing behavior of the user (e.g. monthly, or yearly), so the numbers are not referring to a specific day or action. Then the result will be 25 calculated averages that represent the department rule values.

We have three categories of employees; consequently we have three rules for each of them. The generated rules are shown in figure 2.

Rule1 (HR) =
{0.0,0.0,0.0,0.0,0.0,0.0,7.31,8.26,0.52,0.63,0.58,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,7.48,8.27,0.65,0.73,0.96 }

Rule2 (finance) =
{7.35,8.47,0.52,0.52,0.52,0.0,0.0,0.0,0.0,0.0,0.0,7.42,8.22,0.54,0.54,0.54,7.53,8.36,0.52,0.47,0.63,0.0,0.0,0.0,0.0,0.0,0.0 }

Rule3 (HR administration) = {
7.45,8.35,0.20,0.41,0.31,7.07,8.26,0.40,0.40,0.40,7.39,8.06,0.27,0.78,0.53,0.0,0.0,0.0,0.0,0.0,0.0,7.25,8.34,0.55,0.55,0.55}

Figure 2: Generated rules depending on average features (timings)

D. Classification and pattern recognition:

Here comes the role of the second code, which performs the classification process for each employee according to his\her departments, then performs a pattern recognition process with the generated rules. Results are determined after that.

V. RESULTS

#

In order to evaluate the system performance, we run several tests using small and large amount of data. This test is performed to inspect the system performance in term of efficiency and accuracy. The system has been tested under three different cases:

- 100 regular records,
- 1000 intrusion records,
- 10,000 that include 9,000 regular records
- 1,000 intrusion records.

For the 100 records, the tests showed that only four records were classified as intruders out of 100 as shown in figure 3, this is because the constructed rules were based on the average of all the employees' data and a very few number of them didn't match the average and the minutes tolerance as well, figure 4 shows finance department user and its' rule. It shows the difference in tolerance between (average duration times at department that extend to hours part), thus, it is presented as intruder. For the 1,000 intrusion records, the system classifies all of them as intruders, with 0% error rate as shown in figure 5. For the 10,000 records the system shows 1,274 intruders out of 10,000 as shown in figure 6, which is an excellent result. The following chart shows the three cases results.

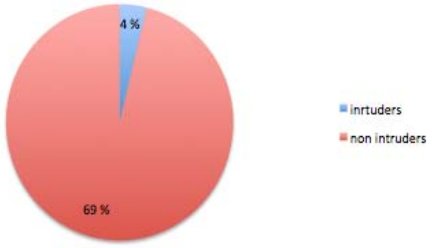

Figure 3: Tested 100 regular data records results

Rule2 (finance) =
{7.35,8.47,0.52,0.52,0.52,0.0,0.0,0.0,0.0,0.0,0.0,7.42,8.22,0.54,0.54,0.54,7.53,8.36,0.52,0.47,0.63,0.0,0.0,0.0,0.0,0.0,0.0 }

Finance record =
{7.09,8.30,0.58,1.00,0.79,0.00,0.00,0.00,0.00,0.00,0.00,7.36,8.36,1.00,1.00,1.00,7.54,8.06,0.27,0.50,0.39,0.00,0.00,0.00,0.00,0.00,0.00}

Figure 4: intruder sample compared with its' rule, values show the difference in tolerance (duration extend to hours part)

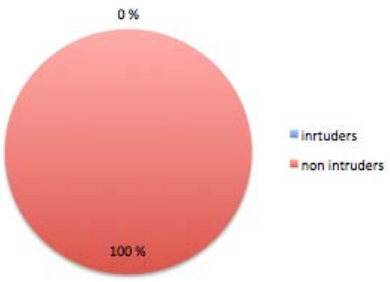

Figure 5: Tested 1,000 intruders data records results

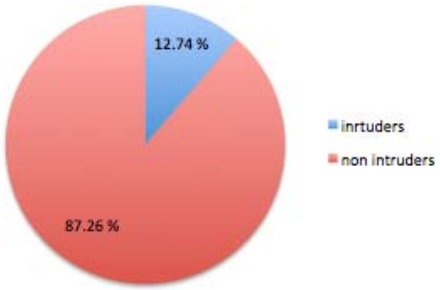

Figure6: Tested 10,000 (9,000 regular and 1,000 intruders) data results

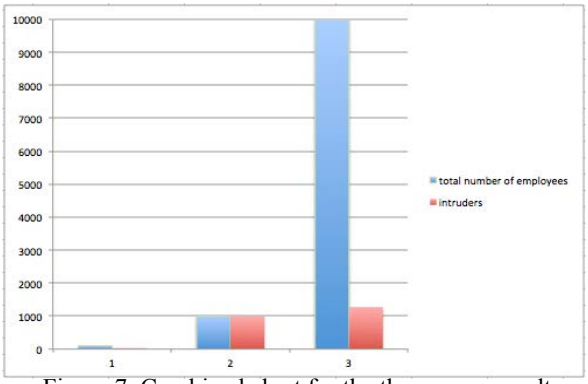

Figure 7: Combined chart for the three cases results

VI. CONCLUSION

This paper gives an overview of applying data mining techniques to intrusion detection systems. We have described a framework of mining patterns and constructing features from data. Generating rules were based on the user accessibility to his\her department folders and the average duration times that spent at those folders. The classification technique has been approached in order to classify each user to his or her referred department. After classification and rules generation are processed, the rule of pattern recognition begins to match the constructed rules with user records to determine if a particular user is an intruder or not. The system has been tested under different amounts and different cases of data. The results are showing a promising future in data mining fields, which produced the efficiency and accuracy in our system.

REFERENCES

[1] Zhan Jiuhua . "Intrusion Detection System Based on Data Mining ". Physics and Electron Communication Department. Leshan Teachers College, Sichuan, 614000, China.Workshop on Knowledge Discovery and Data Mining, 2008.

[2] R. Heady, G. Luger, A. Maccabe, and M. Servilla. " The Architecture of a Network Level Intrusion Detection System". Technical report, Department of Computer Science, University of New Mexico, August 1990

[3] LI Min. "Application of Data Mining Techniques in Intrusion Detection", An Yang Institute of Technology.

[4] WenkeLee(1), SalvatoreJ.Stolfo(2), PhilipK.Chan(3), Eleazar Eskin(2),Wei Fan(4),Matthew Miller(2), Shlomo Hershkop(2), andJunxinZhang(2). "Real Time Data Mining-based Intrusion Detection", (1) Computer Science Department, North Carolina State University, Raleigh, NC 27695 , (2) ComputerScience Department, Columbia University, New York, NY 10027, (3) Computer Science Department,Florida Instituteof Technology,Melbourne,FL32901, (4) IBM T.J.Watson Research Center, Hawthorne, NY 10532, 2001.

[5] Tan, Steinbach, Kumar. "Introduction to Data Mining",2004.

[6] Maldonado Dary Alexandra Pena. "Data Mining: A New Intrusion Detection Approach".GIAC Security Essentials Certification Practical Assignment Version No 1.4 Option 1 ,June 19th 2003

[7] Liu Wenjun . "An Security Model: Data Mining and Intrusion Detection". Department of Computer Science & Technology, Nanchang Institute of Technology, Nanchang, Jiangxi, 330099, China. 2nd International Conference on Industrial and Information Systems, 2010.

[8] Fang-Yie Leu and Kai-Wei Hu. "A Real-Time Intrusion Detection System using Data Mining Technique". Department of Computer Science and Information Engineering, Tunghai University, Taiwan.

[9] Han, Jiawei and Kamber, Micheline. "Data Mining Concepts and Techniques", University of Illinois at Urbana Champaign , second edition , 2006 , p 47, p 48, p451