Intrusion Detection System Based on data mining for Host Log

Ming Zhu
School of Computer Science and Technology
Donghua University
Shanghai, China
zhuming@dhu.edu.cn

ZiLi Huang
School of Computer Science and Technology
Donghua University
Shanghai, China
zilihuang@outlook.com

Abstract—The traditional intrusion detection technology is mostly based on the needs of Web log, using a single data mining to improve the algorithm analysis, which cannot be used in an unknown environment of zero-knowledge rule database, and the efficiency of detecting the potential threats and abnormal behavior is not significant. Therefore, the Paper proposes an intrusion detection system based on data mining for host log. In the premise of zero-knowledge rule database, the combination between ARIMA time series modeling and misuse detection and the combination between Apriori association algorithm and anomaly detection effectively solve the problem of intrusion detection of host system from two dimensions of real-time detection and post detection. In this Paper, the intrusion detection system is designed, and the detection efficiency and the rate of the proposed hybrid mining pattern algorithm and the single data mining algorithm are compared. The experimental results show that the detection rate of the intrusion detection method with hybrid mining pattern is improved by 30% at least, and when the log scale is larger, the expressed detection rate is faster and the system stability is stronger.

Keywords—intrusion detection; host log; ARIMA Time Series; Apriori algorithm;

I. INTRODUCTION

With the development of information technology, the security problem of computer follows. There exists hidden danger for computer in network in different degrees. Such security problems as loss of data, leaking of sensitive information and damage to the system are plaguing the people. Therefore, it is urgent to study the problem of computer intrusion detection [1]. Intrusion detection technology is the process of identifying the intrusion, which attempts, or is happening or has occurred. It could find the intruder and take the evidence of the intrusion. Therefore, in order to protect the system's security and investigate the system fault, viewing the system log is essential.

However, the implicated useful information of the massive log data is hard to discover only if by virtue of the method that the administrator views the log record. Therefore, it is urgent to help people to search useful information from the large amount of data and make user behavior analysis of a higher level by using new technology and tools, to make a better use of these data. However, the current database system can not find the relationship and rules in the data, even not predict the future development trend based on the existing data, leading to the "data explosion but knowledge poverty" phenomenon. Therefore, the application of date mining technology in log analysis [2] is the research hotspots in current research fields.

At present, there are two kinds of basic intrusion detection architecture in the network environment: intrusion detection system based on host and intrusion detection system based on network. The detection scope of intrusion detection system based on network [3] is limited, which generally could not detect the data packet of different network segments. The installation of multi sensor detection will greatly increase the cost of the system deployment, and may encounter the problem that the network encryption data makes it unreadable. In order to avoid these problems, this Paper is based on intrusion detection system based on host, which extracts the host log from the server as the data source of the intrusion analysis.

However, targeting at the host system, the mining method of traditional intrusion detection is to utilize the single data-mining algorithm to improve, and take the existing security strategy as the supervision set and acquire the rule of security features through cross training. These security rules have to rely on the original rule-base and the improvement of detection efficiency and the rate of the single mining method is limited, which cannot deal with the growing log scale. For example, Lee et al. [4] proposed intrusion detection method based on the data mining technology. The association rules mining and frequent pattern mining is applied into the intrusion detection research. The intrusion feature is expressed through rule learner and the first intrusion detection system based on data mining: JAM system is built. The Security Research Group of Purdue University [5] proposed the intrusion detection system based on host, its main method is to adopt the independently operating process group to detect respectively, compare the security behavior of rule-base and train these subjects, marking the abnormal behavior and transmitting the detection results to the detection center.

Therefore, intrusion detection system based on host proposed in the Paper adopts the ARIMA time series modeling to acquire the periodic log and predict the regular log scale, combining with the misuse detection [6] and preserving it in the knowledge database. Meanwhile, it adopts improved association algorithm Apriori to acquire the attribute set with large correlativity, combining with the abnormal detection [7] to match the existing rules in knowledge database, analyze and

judge the suspected potential threat. Thus, it could not only solve the detection problem of zero-knowledge rule base, but also can increase the efficiency and accuracy of intrusion detection with the expansion of the log scale so that the security of the host system can be improved.

II. PROPOSED MODEL AND ALGORITHM

In order to analyze and dig the potential intrusion information of host log, the misuse detection method based on ARIMA time series modeling and the anomaly detection method based on the Apriori association rule is described in detail.

A. ARIMA time series modeling

The full name of ARIMA model is Autoregressive Integrated Moving Average Model [8]. In this Paper, the basic idea of ARIMA modeling is to regard the historical statistics as a time series, which is a serial of random variables dependent on time t, in the premise that the serial of host log quantity is stable in the forecast time. There exists interdependence, correlation and regularity between these variables. If the statistical model reasonable as far as possible could be established based on these host logs, these models could be used to explain the regularity of the log scale and these log data acquired could be used to predict the future log data, namely to find the abnormal log, and to achieve the goal of misuse detection.

In ARIMA(p,d,q) model, AR is the autoregression, p is autoregression item; MA is the moving average, q is the number of terms of moving average, d is the differential times made by the time serial of host logs in stability.

Where, autoregression AR(p) model and MA(q) model is respectively shown as:

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t \tag{1}$$

$$X_t = \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \tag{2}$$

Where, nonnegative integer p is the autoregressive order of log, $\varphi_1,...,\varphi_p$ is the autoregressive coefficient, nonnegative integer q is the moving average order of log, $\theta_1,...,\theta_q$ is the moving average coefficient; X_t is the correlation serial of host logs, ε_t is $WN(0,\sigma^2)$.

Then the ARMA model is expressed with lag operator, $L_i X_t = X_{t-i}$. Therefore, AR(p) model and MA(q) model is respectively shown as:

$$\varepsilon_t = (1 - \sum_{i=1}^{p} \varphi_i L^i) X_t = \varphi(L) X_t \tag{3}$$

$$X_t = (1 + \sum_{i=1}^{q} \theta_i L^i) \varepsilon_t = \theta(L) \varepsilon_t \tag{4}$$

Where, φ represents the polynomial $\varphi(x) = 1 - \sum_{i=1}^{p} \varphi_i x^i$, θ represents the polynomial $\theta(x) = 1 + \sum_{i=1}^{q} \theta_i x^i$.

Finally, the ARMA(p,q) model is expressed as:

$$(1 - \sum_{i=1}^{p} \varphi_i L^i) X_t = (1 + \sum_{i=1}^{q} \theta_i L^i) \varepsilon_t \tag{5}$$

ARIMA(p,d,q) model is the extension of ARMA(p,q) model. Its essence is to carry out d times differential processing for the unstable historical data of host logs Y_t and acquire new stable log serial X_t, and then make X_t fit ARMA(p,q) model, and restore the original d times differential, the predicted date of future log amount of Y_t could be acquired.

Therefore, it is known from equation (5) that the general expression of ARIMA(p,q) is shown as:

$$(1 - \sum_{i=1}^{p} \varphi_i L^i)(1 - L)^d X_t = (1 + \sum_{i=1}^{q} \theta_i L^i) \varepsilon_t \tag{6}$$

The basic process that ARIMA(p,d,q) model conducts the modeling for the time serial of host log amount is as follows:

STEP 1. Its stability is recognized according to the scatter diagram of time serial, autocorrelation function and partial autocorrelation function diagram of host log. The stationary processing is carried out for data of unstable the time serial until the value of the autocorrelation function and the partial autocorrelation function is not significant and non-zero.

STEP 2. The model recognition is conducted by the autocorrelation function $\hat{\rho}_k$ and partial autocorrelation function $\hat{\varphi}_{kk}$ of serial X'_t after computation pre-processing. The specific computation equation is shown as:

$$\hat{\rho}_k = \frac{\sum_{t=1}^{N-k} X'_{t+k} X'_t}{N} \tag{7}$$

According to the computation results of equation (7) and the model recognition principle in Table I, the suitable model of X'_t could be determined.

TABLE I. Recognition Principles of ARMA(p, q) Model

Model	Autocorrelation	Partial autocorrelation
AR(p)	Trailing, exponential damping or oscillation	finite length, truncation (p steps)
MA(q)	finite length, truncation (q steps)	Trailing, exponential damping or oscillation
ARMA(p.q)	Trailing, exponential damping or oscillation	Trailing, exponential damping or oscillation

STEP 3. Parameter estimation and hypothesis test.

To check whether the model meets the stability and reversibility, requiring the roots of equation (8) (9) is outside the unit circle:

$$\varphi(B) = 1 - \sum_{j=1}^{p} \varphi_j B^j = 0 \tag{8}$$

$$\theta(B) = 1 - \sum_{j=1}^{q} \theta_j B^j = 0 \tag{9}$$

Then further, whether the residual serial of above models is the white noise shall be judged according to equation (10). If not, the model recognition shall be re-carried out, if it is, the prediction model of reliability shall be acquired through testing.

$$X'_t = \hat{\varphi}_1 X'_{t-1} + ... \hat{\varphi}_p + \varepsilon_t - \hat{\theta}_1 \varepsilon_{t-1} - ... - \hat{\theta}_q \varepsilon_{t-q} \tag{10}$$

STEP 4. The prediction shall be conducted by using the tested model. If the current log amount matches with the predicted model within threshold value, the periodic log task shall be filtrated with the script, which is deemed as the safety status, if it not matches, it shall be fed back to the administrator as the warning.

B. Apriori association rule algorithm

In the algorithm of association rule, the most classic is the Apriori algorithm [9]. The basic idea of analyzing the log by using the algorithm is to find out all the parameter attribute of host log, and the frequency of itemset of attribute shall at least be same with the predefined minimum support. And the strong association rules are generated by the frequency set, which must satisfy the minimum support and minimum confidence. All the rules of item only including the set are generated by using the predicted rules generated by the frequency set, among which there is only one item at the right of every rule. The definition of middle rule is adopted here. Once these rules are generated, the rule larger than the minimum confidence given by the user could be left. In order to generate all the frequency set, the method of recursion is used.

In the process of mining association rules, the most important step is to generate frequent itemset. Most of the time of Apriori is consumed in the process of repeated scanning of the database. The Paper proposes an improved support algorithm of frequent itemset Apriori based on the association analysis. First, Apriori acquires the probability of all 1 sets by scanning the database. All 1 sets that are larger than the user specified minimum support are frequent 1 itemset. Then the probability of all 2 itemset, 3 itemset,…k itemset are calculated by the algorithm; thus, it acquires all the candidate frequent itemset; later, the algorithm scans the database to verify the support of candidate frequent itemset, and finally get the association rules.

Definition 1: $P_1,P_2,...,P_n$ is the probability of every log attribute $A_1,A_2,...,A_n$, the probability that two attributes A_k, $A_m(P_k<P_m)$ appear in the same log is P_{km}. If for any $k(1<k\leqslant n)$ and any $1\leqslant i_1<i_2<...i_k\leqslant n$, $P(A_{i1}A_{i2}...A_{ik})=P(A_{i1})P(A_{i2})...(p(A_{ik})$, then $A_1,A_2,...,A_n$ is that the log attribute is fully uncorrelated.

The itemset of candidate frequency of log refers to the itemset of log attribute that the probability is larger than the minimum threshold vale given by the user.

If A_k and A_m log attribute is fully uncorrelated, then $P_{km}=P_k\times P_m$. If A_k and A_m log attribute is fully correlated, then P_{km} is the smaller value of P_k and P_m, namely P_k, so $P_k\times P_m\leqslant P_{km}\leqslant P_k$. Under the condition of given P_k and P_m, the value of P_k shall be estimated.

Given parameter a is the probability that A_k and A_m is fully correlated, parameter b is the probability that A_k and A_m is fully uncorrelated, a+b=1 and 0<a, b<1, then P_k could be expressed as:

$$P_{km} = a\times P_k + b\times P_k \times P_m \tag{11}$$

The problem is simplified as the value of parameter a and parameter b. Given $S=\{S_1,S_2, ...,S_m\}$ is the list of A_m value, $S_1,S_2,...,S_m$ is acquired from the log database; $X=\{X_1,X_2,...,X_m\}$ is the list of A_k value, $X_1,X_2,...,X_m$ is also acquired from the log database. Then the calculation formula of correlativity is shown as:

$$a_{km} = \frac{\min\limits_i \min\limits_k \Delta_i(k) + \rho \max\limits_i \max\limits_k \Delta_i(k)}{\Delta_i(k) + \rho \max\limits_i \max\limits_k \Delta_i(k)} \tag{12}$$

Where, a_{km} is the association coefficient of log attribute of A_k and A_m, $\triangle_{ik}=|S_i-X_i|$, ρ coefficient is given by the user, $\rho \in(0,1)$

The specific procedure of improved Apriori algorithm is as follows:

STEP 1. Establish new array P[n], the initial value of the element is 0; scan the entire log database, calculate the probability $P_1,P_2,...,P_n$ of every log attribute $A_1,A_2,...,A_n$. Assign the array element P[1],P[2],...,P[n] as $P_1,P_2, ...,P_n$;

STEP 2. Set the minimum probability V1, if P[i] is larger than the threshold value, then Ai is number of the frequent 1 itemset, P[1],P[j],...,P[m] is the probability of frequent 1 itemset.

STEP 3. The probability of a item ser of all log attributes could be calculated according to P[1],P[j],...,P[m] and the calculation formula of correlation coefficient of log attribute. Given the threshold value of minimum probability V_1, if the probability of I itemset exceeds the threshold value, then it is the candidate frequency 2 itemset; otherwise, its probability is set as zero. The support of all candidate frequency 2 itemset could be recorded according to array $P_2[i]$.

Given V_{k-1} is the threshold value of minimum probability of candidate frequency (k-1) itemset of log, there is equation (13):

$$V_{k-1} = a\times \max(P) + b\times \min(P)\times \max(P) \tag{13}$$

Where, P is the set of array $P_{k-1}[1]$, $P_{k-1}[2]$,...,$P_{k-1}[m]$.

STEP 4. Repeat above Step1, 2, 3, calculate the probability of k itemset A_1,A_2,...,A_k from k=2 to k=n.

STEP 5. Scan the log database again; calculate the support of candidate frequency itemset of log generated in Step 4.

Set new array D[m], record the candidate frequency itemset of log, the initial value is 0, m is the number of candidate frequency itemset

STEP 6. If D[k] is larger than the threshold value of minimum support, namely it is the frequency itemset.

Where, step 5 and step 6 could be used to guarantee the probability and support of candidate frequency itemset of log, to meet the threshold value input by the user.

STEP 7. Acquire the association rules of log attribute.

Thus, improved Apriori algorithm greatly reduces the times of scanning the database and improves the efficiency of association rule algorithm.

III. SYSTEM STRUCTURE

To apply data mining and intrusion detection technology to host log analysis, framework design for Intrusion Detection System in this article is shown in Figure 1.

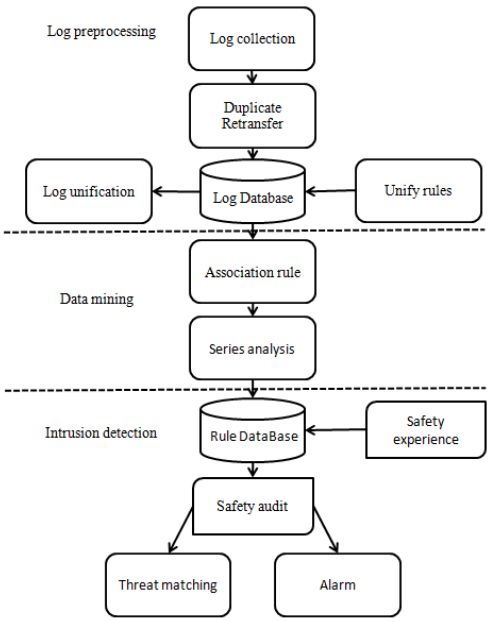

Fig.1. Intrusion Detection System Framework

Main modules of such framework include:

(1) Log preprocessing: collect host log in log database; duplicate, retransfer and store log to temporary cluster in batches as per day at during idle time; in accordance with required format, clear the obtained log, unify format and input it in analysis host to fulfill the purpose of log preprocessing.

(2) Data mining: perform association rule mining and time series analysis for unified data in accordance with selected algorithm and given parameters; continuously adjust parameters recursive analysis, cache its analysis results and cache the obtained log with the greatest dubiety of including mining knowledge.

(3) Intrusion detection: contrast data mining and analyzed dubious log with safety rule and known safety experience in threat database one by one; perform safety audit for log data within a certain period; produce corresponding alarm or add new safety rule.

Framework of data mining in intrusion detection model is shown in Figure 2: firstly, collect host log, obtain user behavior data set and unify log in accordance with algorithm requirements; secondly, respectively use sequence and association analysis algorithm to perform data mining for conventional events, construct safety rules bank and obtain intrusion data set; thirdly, use association analysis for rule mining and identify normal behavior and intrusion behavior; besides, continue to mine intrusion data set, extract intrusion mode, construct characteristics detection model for intrusion data and continuously update such model as per newly obtained data so as to be used for misuse detection.

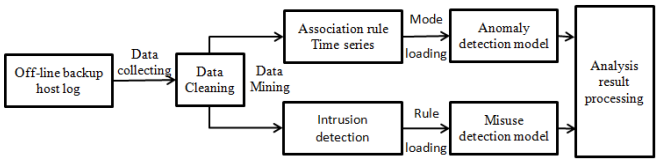

Fig.2. Application Framework of Data Mining in Intrusion Detection

Intrusion detection engine detects host log in accordance with safety rules bank as judgment basis for misuse detection on one hand, and uses association and sequence mode as anomaly detection basis on the other hand to judge whether host log is normal or intrusive and feeds back the result to analysis result processing module.

IV. EXPERIMENTAL RESULTS

Since different data mining algorithms and characteristics of intrusion detection technology have nothing in common, we design a series of tests to verify stability and detection rate of such Intrusion Detection System and contrast with single mining algorithm in detection rate.

In this test, data samples all come from host log of the same kind in production office environment of intranet.

In data mining parameters setting, degree of support for association rule is set as ‰4; degree of confidence %3; error threshold of time sequence ‰5; about 200 items of safety rule entries are added in rules bank.

TABLE II. Detection Rate Contrast in Intrusion Detection

Scales of log	Association rules	Time Series	Intrusion Detection	A & T & I
2000	38.5%	15.4%	10.7%	68.4%
5000	39.1%	15.4%	19.4%	69.6%
10000	39.2%	16.1%	14.6%	75.6%
20000	41%	16.8%	16%	79.1%
40000	40.2%	17.5%	15.7%	83.5%

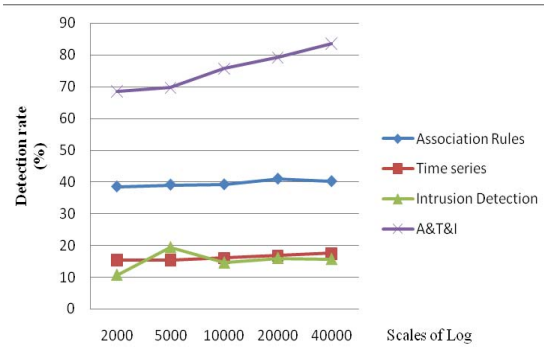

Fig.3. Intrusion Detection Efficiency Test

Table II shows the comparison of intrusion detection rates based on different methods. The first column shows the number of sample logs while all other lines indicate detection rates obtained with different methods based on the same log volume. Intuitive line comparison chart is shown in Figure 3.

According to Table 2, the highest log detection rate is only about 40% by using a single detection method. However, when intrusion detection technology based on hybrid mining method, detection rate ranges from 68.4% to 83.5%. In addition, as shown in Figure 3, with the expansion of log size, detection rate is on increase by hybrid mining method. Therefore, hybrid mining method has advantage in intrusion detection and better self-adaptability thus providing higher detection rate.

Similarly, we compared intrusion detection method based on hybrid mining method and traditional intrusion detection method. In the test, we increased sample log volume to approximately 20000,40000,60000,80000. Then, we collected and analyzed the time spent on these two methods as shown in Figure 4.

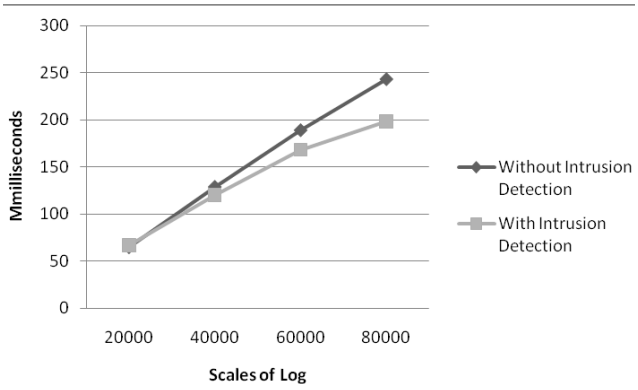

Fig.4. Intrusion Detection Time-consuming Test

Figure 4 shows that when the log size is small, time consumption detected is almost the same because both methods have equivalent basic rule base. With the increase of log size, the system combined with intrusion detection method could adaptively add new rules to rule base thus dispensing with repeated mining; therefore, it is faster than data mining method. In addition, with the increase of log size, growth rate of time consumption is not obvious for this method. For this reason, compared with traditional intrusion detection, hybrid mining method has higher detection rate.

V. CONCLUSIONS

This paper aims at intrusion detection of host log through hybrid mining method. ARIMA time series, Apriori association algorithm and intrusion detection method are integrated and applied to intrusion detection system of host system in order to improve system security. Such detection method could not only solve the problem of zero rule knowledge base detection and improve log detection efficiency by at least 30% but also well cope with rapid increase of log size thus enhancing the stability of intrusion detection system.

VI. ACKNOWLEDGMENT

I would like to express my sincere gratitude to my supervisor, Ming Zhu, a respectable and erudite scholar, who has provided me with valuable instructions for this paper.

References

[1] Rowland C H. Intrusion detection system: U.S. Patent 6,405,318[P]. 2002-6-11.

[2] Barbara, Daniel, et al. "ADAM: Detecting intrusions by data mining." In Proceedings of the IEEE Workshop on Information Assurance and Security. 2001.

[3] Vigna G, Kemmerer R A. NetSTAT: A network-based intrusion detection system[J]. Journal of computer security, 1999, 7(1): 37-71.

[4] Lee W, Stolfo S J, Chan P K, et al. Real time data mining-based intrusion detection[C]//DARPA Information Survivability Conference & Exposition II, 2001. DISCEX'01. Proceedings. IEEE, 2001, 1: 89-100.

[5] Spafford E H, Zamboni D. Intrusion detection using autonomous agents[J]. Computer networks, 2000, 34(4): 547-570.

[6] Depren O, Topallar M, Anarim E, et al. An intelligent intrusion detection system (IDS) for anomaly and misuse detection in computer networks[J]. Expert systems with Applications, 2005, 29(4): 713-722.

[7] Garcia-Teodoro P, Diaz-Verdejo J, Maciá-Fernández G, et al. Anomaly-based network intrusion detection: Techniques, systems and challenges[J]. computers & security, 2009, 28(1): 18-28.

[8] Box G E P, Pierce D A. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models[J]. Journal of the American statistical Association, 1970, 65(332): 1509-1526.

[9] Agrawal R, Srikant R. Fast algorithms for mining association rules[C]//Proc. 20th int. conf. very large data bases, VLDB. 1994, 1215: 487-499.