2003 International Symposium on Advanced Intelligent Systems
September 25-28, 2003, Jeju, Korea

An Intelligent Intrusion Detection Model

Myung-Mook Han
KyungWon University
College of Software, San 65 Bokjung-Dong, Sujung-Gu,
Songnam, Kyunggi-Do, 461-701, Korea
mmhan@mail.kyungwon.ac.kr

Abstract - The Intrsuion Detecion Systems(IDS) are required the accuracy, the adaptability, and the expansion in the information society to be changed quickly. Also, it is required the more structured, and intelligent IDS to protect the resource which is important and maintains a secret in the complicated network environment. The research has the purpose to build the model for the intelligent IDS, which creates the intrusion patterns. The intrusion pattern has extracted from the vast amount of data. To manage the large size of data accurately and efficiently, the link analysis and sequence analysis among the data mining techniqes are used to build the model creating the intrusion patterns. The model is consist of "Time based Traffic Model", "Host based Traffic Model", and "Content Model", which is produced the different intrusion patterns with each model. The model can be created the stable patterns efficiently. That is, we can build the intrusion detection model based on the intelligent systems. The rules prodeuced by the model become the rule to be represented the intrusion data, and classify the normal and abnormal users. The data to be used are KDD audit data.

I. INTRODUCTION

The Intrusion Detection System(IDS) are required the accuracy, the adaptability, and the expansion. In order to protect the resource which is important and maintains a secret in the complicated network environment with satisfying the above conditions, it is required the more strucutred, and intelligent IDS[1].

Data Mining(DM), also called knowledge discovery in database, can be defined as efficiently discovering interesting patterns from large databases, and has been emerged as a promising new area for database research[2][3]. The knowledge by which is extracted DM can be applied in a wide range of domain including decision support system, marketing, and information security.

Data Mining(DM) techniques have the several methods. Among the methods, we have used the classification, link analysis, and sequence analysis in this paper. The data used in the experiment are KDD data, and the simulaion is conducted in the similar environment[4].

It is necessary to divide the data to the required parts for intelligent IDS, and we construct the model using the link analysis and sequence analysis. The model is consists of three models, which is "Time based Traffic Model", "Host based Traffic Model", and "Content Model", and produce the different intrusion patterns with each model[5]. These models are a model for the system to produce the variable rules automatically, and a model to create the stable and efficient patterns before the classification.

Genetic Algorithm(GA) has been spotlighted as discovering the solution to solve the combinatorial optimization problems[6][7]. GA is the search technique of a solution to imitate a process of biological evolution, in which the population of the plural coded gene to represent a candidate for the solution evolve gradually into the thing to be an object of optimization through the processof the change of generations and selection. The GA has an advantage in searching efficiently the wide scope of a state space in order to explore the solution from lots of initial states.

The purpose of our research is to develop an automated approach in building IDS. We have developed the mechanism that can be applied to a variety data sources to generate the intelligent intrusion detection model. That is, our approach is to apply data mining methods, based on GA, to the audit data to compute models that accurately capture the patterns of intrusions and normal activities.

The rest of the paper is organized as follows. In Sec. 2, we introduce IDS and DM. In Sec. 3, we discuss the proposed intrusion detection model. In Sec. 4, the experiments are explained, and section 5 concludes the paper with a summary.

II. INTRUSION DETECTION SYSTEM AND DATA MINING

A. Intrusion Detection System(IDS)

The type of IDS[8] can be divided abnormal intrusion detection and misuse intrusion detection, and abnormal intrusion detection method is divided ststistical approach, feature selection, predicitive pattern generation, neural network, and hybrid for action measurement method. Also, misuse intrusion detection method is classified pattern matching,

model-based detection, keystroke monitoring, state transition analysis, expert system, and conditional probability.

There is a dividing method based on the data source for detection area, except the above method which uses the detection method based on intrusion detection model.

With a dividing method based on the data source for detection area, we can divide host based intrusion detection system, multi-host based intrusion detection system, and network based intrusion detection system. Host based intrusion detection system uses the audit data that host system has been produced and collected to decide the intrusion, and a host system is the detection area. Multi-host based intrusion detection system uses the audit data that several host systems have been produced and collected to decide the intrusion. Since several host systems are the detection area, the information which is needed to decide the intrusion are exchanged among the host systems. On the other hand, network detection system collects the packet data of network and use them to decide the intrusion, and the whole area to be installed the IDS is the detection one.

B. Data Mining(DM)

DM is the essential part of the knowledge discovery in database. It extracts the valuable information and mutual association of data to be hidden using the pattern recognition, statistical method, and artificial method from the vast amount of data.

B.1 Association rule

If the transation of particular items in the particular item set has happened, the transation of particular itmes in the other particular item set is also happened. It is regarded this as the association. The analysis which tries to discover these phenomena is association analysis. That is, the essential of association rule algorithm is how much the scan times of database can be reduced.

The association rule is the rule which is reflected the association of each items in the transation represented as item set. After Agrawal introduced the algorithm, the advanced algorithm has been developed such as the reduction of search times in database, removing the limit of main memory[9]. The problem to search the association rules discovers the large item set that all of item set held the transation support degree over the minimum support degree determined previouly, and produces the association rules[10].

B.2 Frequent Episodes

Frequent Episodes is the operation to model the sequential patterns. It patterns the sequence of audit event to occure together based on the time. It is similar to the association rule, but uses the time window and discover the rules among the event produced during the certain times.

It is necessary to research the sequential pattern happened in the audit data frequently to understand not only the temporary and statistically properties of the most attack but also the user or the regular action of program. We use the frequent episodes method to represent and to discover the sequential phenomena of audit record patterns.

We found the 3 kinds of rules using the link analysis and sequence analysis, and explain the extracted steps in section 3 and 4.

B.3 Genetic Algorithms

Genetic Algorithms(GAs) are efficient and independent search method and have been used to learning classifier rules[11]. Also GAs are applied to concept learning, feature selection, adjust of parameter, and construction of feature.

The most of data mining system uses the modification of traditional machine learning algorithm. In machine learning, there are two purposes that it learns the complex system and makes the appropriate output of system. The machine learning based on genetic algorithm is called as GA machine learning or GBML(genetic based machine learning).

The aspect that machine learning method is basically different from the optimization problem is to seek a set of rules. Since the purpose of optimization problem is to seek the optimal solution, it is enough for one thing to converge the individual. But machine learning is not to seek a best rule, but to seek a set of rules cooperating each other. Generally, there are two approaches in GBML. The whole rule set is represented as an individual, and an individual of population is a rule set of candidate. And then it is natural that a new generation of a rule set is created by selection and genetic operator. That is, a traditional genetic algorithm is used and each entity in a population is a rule set which represents a complete solution of learning problem. This is called Pitt approach method[12].

At the similar time, Holland has developed a classifier system that an individual of population is a rule and a rule set is represented by a population. This method is called Michigan approach method[13]. The Michigan approach method has used a quite different evolution method, that a population consists individual rules and each rule represents a candidate solution of an overall learning task. That is, Pitt approach method is similar to evolutionary computation but Michigan approach method is to use a quite different method.

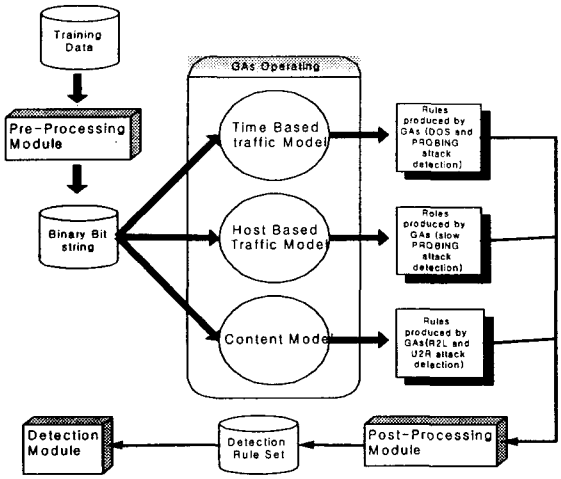

Fig. 1. The automatic rule generator

III. INTRUSION DETECTION MODEL

We are able to make the automatic rule generator through concretization of the intrusion detection model shown in Fig.1. With this we can decide the intrusion situation.

A. Assocaition rule

The purpose to use association rule is to induce the relationship among attributes from the database table. Association rule is represented as *X ->Y,[c,s]* , here *X* and *Y* are the set of attribute values, respectively. Also, $X \cap Y = \varnothing$ and $s=support(X \cup Y)$ is support degree of rule, $c=\frac{support(XUY)}{support(X)}$ confidence degree.

The research has calculated the association rule from the audit record data such as the below one. Table 1 is shown the KDD data record, which is

Table 1. KDD data record

duration	protocol_type	service	Flag	src_bytes	.....	label
4	tcp	pop_3	SF	30	.....	guess_passwd
4	tcp	pop_3	SF	28	.....	guess_passwd
4	tcp	pop_3	SF	30	.....	guess_passwd
4	tcp	pop_3	SF	30	.....	guess_passwd
4	tcp	pop_3	SF	30	.....	guess_passwd
5	tcp	pop_3	SF	28	.....	guess_passwd
4	tcp	pop_3	SF	30	.....	guess_passwd
4	tcp	pop_3	SF	30	.....	guess_passwd
4	tcp	pop_3	SF	30	.....	guess_passwd
39	tcp	telnet	SF	270	.....	normal
⋮	⋮	⋮	⋮	⋮	⋮	⋮

represented the record of system usuage for the system user. Some attributes are indispensable for experiment, the others are not. The association rule has been used for divide and discover these attributes.

In the experiment, we use Apriori algorithm, which is the most popular one among the association rule algorithms.

The attribute set X occured frequently is the case $support(X) \geq minimum\_support$. That is, we have to produce the attribute set X that the result is shown over the minumum support degree. In the experiment, we set the minimum support degree 0.2.

Each attribute is indexed from the left as follows.

Table 2. The index of attributes

Num.	Attributes	Ind.	Num.	Attributes	Ind.
1	duration	F0	···	·····	···
2	protocol_type	F1	40	dst_host_rerror_rate	F39
3	service	F2	41	dst_host_srv_rerror_rate	F40

The attribute value of F0 is represeneted as combining the small character "f" and numeric number "0" such as f0. For example, "F3=f2" means that the attribute value of service, the third attribute from the left, has the second value.

B. Frequent Episodes

The set of event record to be recorded the time is given, each record of this is the set of attributes. The block $[t_1, t_2]$ is the continuous set of event record appeared between time stamp t_1 and t_2.

Support(X) is defined as the ratio of minimum appearance rate including X to the whole of event record. Frequent Episodes is represented as *X, Y ->Z,[c, s, w]*. Here, *X, Y, Z* are the set of attributes, and $s=support(X \cup Y \cup Z)$ is the support degree for the rule, $c=\frac{support(XUYUZ)}{support(XUY)}$ is the confidence degree for the rule. And $w(timewindows)=[t_1, t_2]$

Therefore, the purpose of frequent episodes is to seek the sequential rule among the attributes using these.

Z,[c, s, w]. Here, *X, Y, Z* are the set of attributes, and $s=support(X \cup Y \cup Z)$ is the support degree for the rule, $c=\frac{support(XUYUZ)}{support(XUY)}$ is the confidence degree for the rule. And $w(timewindows)=[t_1, t_2]$

Therefore, the purpose of frequent episodes is to seek the sequential rule among the attributes using these.

IV. EXPERIMENTS

We used the KDD data which is the data set used for the Third International Knowledge Discovery and Data Mining Tool Competition held in conjunction with KDD-99. The competition task was to build a network intrusion detector, a predictive model capable of distinguishing between bad connections, called intrusion or attacks, and good normal connections. This database contains a standard set of data to be audited, which includes a wide variety of intrusion simulated in a military network environment. That is, the connection is consists of TCP packet series in the begin and the end. With the confidential protocol, this includes the packet of start IP and destination IP in the packets, and the flows of data.

The major 4 kinds of attack in the KDD data are divided as follows.

- DOS : denial-of-service, (syn flood etc.)
- R2L : unauthorized access from a remote machine, (guessing password etc.)
- U2R : unauthorized access to local superuser(root) priviledges (buffer overflow etc.)
- PROBING : surveillance and other probing (port scaning etc.)

The detection models, which is time based traffic model, host based model, and content model, are the classification model specialized the each intrusion type.

The produced rules are evaluated in the test data, which are included 38 kinds of attacks. Among them 14 attacks are regarded as new type.

Fig.2 is shown the performance of intrusion rule produced in each model. Time based traffic model is 92%, host based model is 97%, and content model is 94% in the detection rate, and 0.02, 0.03, 0.05% in the false alarm rate.

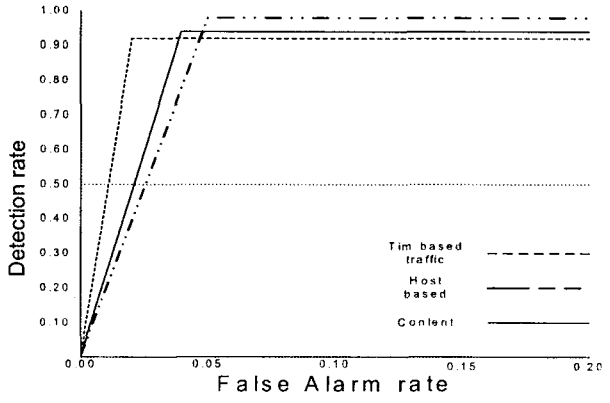

Fig.2 error and false alarm rate in the 3 models

V. CONCLUSION

The paper has proposed a data mining method for constructing the intelligent intrusion detection system. That is, our approach is to apply data mining techniques, based on GA, to the audit data to compute models that accurately capture the patterns(rules) of intrusions and normal acitivities. During the process, we apply the GA-based classifier system to KDD data, and the produced rule has been shown the 94.3% detection rate in each model.

REFERENCES

[1] Wenke Lee, Salvatore J, Solfo, Data Mining Approaches for Intrusion Detection, In Proc. of the 7th VSENI Security Symposium, San Antonio, TX, Jan., 1998.

[2] Michael J.A. Berry and Gordon Linoff, Data Mining Techniques for Marketing, Sales, and Customer Support, Wiley Computer Publishing, 1997

[3] Pieter Adriaans and Dolf Zantinge, Data Mining, Addison-Wesley, 1996.

[4] //kdd.ics.edu/databases/kddcup99/kddcup99.html

[5] Wenke Lee, Salvatore J.Stofo, A Framework for Constructing Features and Models for Intrusion Fetection Systems, In Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1999.

[6] D.E.Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning, Addison Wesley, 1989.

[7] J.H.Holland, Adaptation in Natural and Artificial Systems, Ph.D. thesis, Univ. of Michigan, Ann Arbor, Mich., 1975

[8] Denning, Dorothy E, An Intrusion Detection Model, IEEE Transaction on Software Engineering, Feb.,1987.

[9] R.Agrawal, T.Imielinski, and A.Swami, Mining association rules between sets of items in large databases, In proc. of the ACM SIGMOD Conference on Management of Data, pp.207-216, Washington,D.C., May, 1993.

[10] R.Agrawal and R.Strikant, Fast algorithms for mining association rules, In Proc. of the 20th International Conference on Very Large Data Bases(VLDB94), pp.487-499, Santiage, Chile, Sep.,1994.

[11] S.Wilson, Classifier Systems and the Animat Problem, Machine Learning, Vol.2, pp.199-228, 1987.

[12] Stephen F.Smith, A Leaning System based on Genetic Adaptive Algorithms, Ph.D. thesis, Univ. of Pittsburgh, 1980.

[13] John H.Holland, Escaping brittles: the possibilities of general purpose learning algorithms applied to parallel rule-based systems, Machine Learning, an artificial intelligence approach, 2. 1986.