Advanced Science and Technology Letters
Vol.134 (CIA 2016), pp.120-124
http://dx.doi.org/10.14257/astl.2016.134.21

Research of Network Intrusion Detection System based on Machine Learning and Rough Set Theory

Yang Hui-jun[1]

[1] Department of Information Service, Anhui Institute of International Business,
Hefei AnHui, 231131, China

Abstract. Data mining for intrusion detection is one of the most cutting-edge researches which focus on network security, database, and information decision-making. Due to the emergence of new forms of attacks and intrusion on the network, we need a new intrusion detection system which would be able to detect new and unknown attacks. In the paper, by studying the characteristics of network data intrusion, we put forward a intrusion detection system based on Rough set theory, and detect anomaly action in network. This method can extract detection rule model of the network connection data, dealing with incomplete data and basic discrete data exit in data mining effectively. The experiments results show that, models, methods and generation framework proposed in this paper can effectively detect network intrusion.

Keywords: Network Intrusion, Rough Set Theory, Machine Learning, SVM

1 Introduction

Intrusion detection technology is divided into two categories: anomaly detection and misuse detection. Anomaly detection uses quantitative way to describe acceptable behavior characteristics, distinguishing abnormal behavior characteristics and those contrary to normal behavior action. The advantage of anomaly detection is it can discover new and unknown intrusion, simultaneously, is has a certain learning ability. Misuse detection uses foregone system and attack pattern of application to detect intrusions. The feature of misuse detection is it can detect intrusion with high accuracy, but misuse detection can only discover foregone intrusions.

In this paper, we introduce Rough set theory, and establish network intrusion detection model based on Rough set theory. Simultaneously, we mainly describe Machine Learning based on Bayesian Classification, neural networks-based, and association rules mining intrusion detection technology. We proposed Anomaly Detection System based on Data mining. ADSDM is able to mine suspicious behavior within port numbers, application layer data, and network data protocol. In data mining, we mainly notice association rules data mining method based on weak rule. This method is used to detect attacks those have less abnormal operation and may not be readily detected. At the same time, the influence between network communication time, direction, port number, and host address is used to establish Bayesian networks with various attributes nodes, then regards the networks as abnormality discriminator.

ISSN: 2287-1233 ASTL

During the experiment, we figured out that association between several different data mining methods works more effectively than usage of one method. For instance, combination of Rough set theory and the genetic algorithm can improve intrusion detection accuracy that will be introduced in the later sections.

2 Rough Set Theory and ADSDM

Rough set theory assumes that knowledge is a kind of ability to classify objects, and knowledge must be associated with various classification models among specific or abstract parts of the world. These parts are called universe discussed. Network intrusion detection system includes two stages: pattern generation and pattern detection. During pattern generation, network data collection module is to collect network connection data, and data selection module is to select target data from all connection data including dimension, attribute, and data type. Rough set theory is used in data preprocessing module and reducing knowledge module. During pattern detection stage, the system uses generated detection rules to detect current data and alarm for abnormal behavior.

We choose KDD CUP 99 Data Sets as data source, and there are 41 attribute values in each network connection record which equal to condition attribute in information system. The data in this data source are complete. This algorithm process shows as follows:

(1) Randomly select Initial Population P, and cycle following operation:
(2) P -Selection, then we obtain P1, P2, P3;
(3) P1-Crossover, then we have Q1;
(4) P2-Mutation, then we have Q2;
(5) Inversion- P3, then we have Q3;
(6) Compute new population P in terms of (2),(3),(4),(5);
(7) If genetic fitness in P no longer increases, end;
(8) Else if, return (2).

Attributes of network behavior are not independent each other. Bayesian network describes associated conditional probability distribution, and shows relationship between each variables. There are two parts make up Bayesian network, that is Directed acyclic Graph(DAG) and Conditional Probability Table(CPT). Network intrusion detection system includes two stages: pattern generation and pattern detection. During pattern generation, network data collection module is to collect network connection data, and data selection module is to select target data from all connection data including dimension, attribute, and data type. Rough set theory is used in data preprocessing module and reducing knowledge module. During pattern detection stage, the system uses generated detection rules to detect current data and alarm for abnormal behavior. The algorithm is shown as follows.

Input: and
Output: Alarm Information
For each
Step 1: Compute the probability in terms of CPT values and -support data:

```
            c,
              Step 2: If
             then{//      is   normal,   thus,   this   suspicious   behavior   is
normal(misinformation)
                        {//c, where      is weight
                          }
                     }
                     else
                          Alarm Information
                     Endif.
```

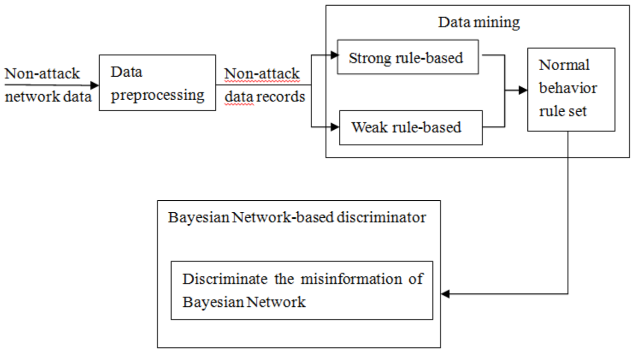

Fig. 1. Training stage of ADSDM

3 Network Security Situational Awareness System

Data source used for knowledge discovery derives from two parts: alarm event sets generated from simulation attacks and historical alarm event sets. Knowledge discovery means discovery and selection situation relation knowledge from alarm event sets.

(1) Reduct and Filtering for Alarm Events.

Evidence Theory(DS)-based stablishment of alarm event filtering mechanism is a statistical analysis for program in terms of alarm event confidence. Firstly, program automatically count distribution of different alarm events. Secondly, based on DS, and delete insignificant events.

(2) Knowledge Discovery

We obtain situation knowledge from alarm event sets in terms of frequent pattern and sequential pattern. Frequent pattern is relation between several event attributes, and its aim is to get the regularity of event attributes. Sequential pattern is sequence relation between events, and its aim is to figure out time series relation and causal relation. In general, we consider relation between attack, sip, dip, sport, protocol.

In order to avoid confusion and convenience of narration, we introduce some related definition.

(1) Security situation: Security situation information includes time dimension and spatial distribution dimension.

(2) Security event: Alarm information generated from network security situation sensor, and initiated by intrusion behavior. Security event can be denoted as a multivariable equation.

(3) Security situation modeling: Analyze all alarm events generated from security sensor, and then from network security situation.

4 Experiments and Analyses

We chose KDD CUP 99 Data Sets as data source, and there are 41 attribute values in each network connection record which equal to condition attribute in information system.

In this experiment, we used 31 attributes in these 41 attributes values. There are 7 basic attributes, 11 connection content attributes, 18 attributes based on host and time. The training data set has 13,107 connected records, and test data set has 26,214 records. Distribution of data connection is displayed in Table 1.

Table 1. Distribution of Data

	Training data	Test data
Type	Percent(%)	Percent(%)
Normal	59.952796	9.727629
Probe	1.213095	1.224527
DoS	38.735037	87.308521
U2R	0.015259	0.003815
R2L	0.083925	1.735714

5 Conclusion

People's work and life has been closely linked with the network. However, there are a lot of potential threats on the internet, and our security need protection. In this paper, we proposed several method to detect network intrusion, and we also introduced some algorithm to figure out this network problem. During the experiment, we figured out that association between several different data mining methods works more effectively than usage of one method. For instance, combination of Rough set theory and genetic algorithm can improve intrusion detection accuracy. The results of our experiments show that the methods and system proposed in this paper effectively detect attacks from attacker or other IP addresses, simultaneously, these methods can also protect our security.

Acknowledgement. This work was supported by the Natural Science Foundation of higher education in Anhui Province (KJ2016A127), and the Professional leaders Research Foundation of Higher Vocational Colleges in Anhui Province (2014 Year).

References

1. Heberlein, L. T., Dias, G., Levitt, K., Mukherjee, B., Wood, J., Wolber, N.: A Network Security Monitor. In: Proceedings of 1990 Symposium on Research in Security and Privacy, 1990
2. Barbara, D., Wu, N., Jajodia, S.: Detecting Novel Network Intrusions using Bayes Estimators. First SIAM International Conference on Data Mining, 2001
3. Hart, J. W., Pei, J., Yin, Y W.: Mining frequent patterns without candidate generation [J]. ACM SIGMOD Record, 2000. 29(2): 1-12
4. Dempster, A.: Upper and lower probabilities induced by multi-valued mapping [J]. Annals of Mathematical Statistics, 1967, 38(2): 325—339
5. Vapnik, V N.: Statistical learning theory. Adaptive and learning systems for signal processing, communications and control, New York: Wiley, 1998