Intrusion Detection System Based on Data Mining

ZHAN Jiuhua
Physics and Electron Communication Department
Leshan Teachers College, Sichuan, 614000, China
lstcwd@163.com

Abstract

Analyzed recent IDS models, the development of IDS (Intrusion Detection System), and the current and gives a brief introduction to DM (Data Mining) technology. Presented a framework of IDS based on data mining for resolving the current problems IDS is facing. The system that performs anomaly detection can detect intrusions known and unknown, reduce omissions and misstatements, improve accuracy and speed of intrusion detection and has good adaptive capacity and scalability.

1. Introduction

In the information age, more and more people use the computer as a tool for crime. Of these, hacking is the most typical. And it has been for many companies and countries had brought great losses. Intrusion detection technology is to prevent such acts. In the traditional IDS (Intrusion Detection System) based on error detection, intrusion mode is usually pre-defined by the security experts. And experts need to rely on manually to update detection mode security. However, in the face of increasing network data flow, relying solely on security expert with eyes to discover all the invasion models is unrealistic. Because it cannot update the model timely, resulting in the intrusion detection rate of false positives has increased significantly. Thus, the need for automated tools to discover new invasion mode, or by semi-automated tools to support security experts found invasion mode. Using data mining technology can effectively solve this problem.

2. Data Mining Description

Data Mining is an analytic process designed to explore data in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. The ultimate goal of data mining is prediction, so predictive data mining is the most common type of data mining and one that has the most direct business applications. The process of data mining consists of three stages: the initial exploration, model building or pattern identification with validation/verification, and deployment.

Exploration: This stage usually starts with data preparation which may involve cleaning data, data transformations, selecting subsets of records and performing some preliminary feature selection operations to bring the number of variables to a manageable range. Then, depending on the nature of the analytic problem, this first stage of the process of data mining may involve anywhere between a simple choice of straightforward predictors for a regression model, to elaborate exploratory analyses using a wide variety of graphical and statistical methods in order to identify the most relevant variables and determine the complexity and/or the general nature of models that can be taken into account in the next stage.

Model building and validation: This stage involves considering various models and choosing the best one based on their pt goal which are often considered the core of predictive data mining.

Deployment: That final stage involves using the model selected as best in the previous stage and applying it to new data in oredictive performance. This may sound like a simple operation, but in fact, it sometimes involves a very elaborate process. There are a variety of techniques developed to achieve tharder to generate predictions or estimates of the expected outcome.

Data mining is an interdisciplinary course, including machine learning, mathematical statistics, neural networks, databases, pattern recognition, rough sets, fuzzy math and related technology. Data mining has many basic theories: Data Reduction, Data Compression, Pattern Discovery, Probability Theory,

This paper is supported by the Young Foundation of Education Department of Sichuan China(NO.2006B074)

0-7695-3090-7/08 $25.00 © 2008 IEEE
DOI 10.1109/WKDD.2008.12

Microeconomic View, Induction Database, etc. In various fields Data mining technology has been widely applied.

3. IDS Research

An IDS is the means that attempts to detect intrusions and collect the evidences of intrusion for data restoration and event treatment. In practice, every intrusion detection system has its own assumption on normal behaviors and abnormal behaviors. Intrusion detection technology is a means that surveillance systems running, and found all kinds of attempted attacks, attacks or attacks, in order to ensure the confidentiality of system resources, integrity and availability. Traditional intrusion detection systems there are two major classes: Anomaly Detection and Misuse detection.

Anomaly detection assumes that there are significant deviations between intrusions and normal behaviors and that normal behavior are consistent in a short period. If we can define each one can be accepted behavior, so every one can not be acceptable behavior should be seen as act of Invasion. Usually, normal operation should have certain characteristics. Usually, the normal operation should have certain characteristics. Serious deviation from the normal behavior of the user's activities should be considered invasion. Because no need to define each act of invasion, anomaly detection can effectively detect unknown invasion. In short, Anomaly Detection has a lower rate of underreporting, but also has a relatively high rate of false positives. Misuse detection assumes that the distinct features of intrusions, which differ from that of normal behaviors, can always be explored. If all unacceptable behavior can be defined, each can be matched with actions would cause alarm. Under Misuse Detection model the IDS collects the non-normal operating characteristics, and built relate database features. When the action and record in the libraries of the consumer match mutually, the system thinks that this kind of action is invasion. Faced with the known attacks it can detailed and accurately report the attack types. However, when faced with unknown types of attacks it has only limited effect. In addition, the characteristics library must be constantly updated. In summary, this detection model's advantage is the low rate of false positives and its deficiency is a high rate of underreporting. So the key of misuse detection is how to discover and express the distinct intrusion features, and of anomaly detection is how to find a set of statistical metric of normal behaviors.

The following characteristics are identified as desirable for an intrusion detection system: robustness, lightweight, efficiency, fault-tolerance, adaptability, scalability and configurability. These features are a natural result of diversity, distributivity, locality and dynamical. There are three architectures of intrusion detection system: centralized architecture, hierarchical architecture, and distributed architecture. Among them, the distributed architecture is promising. There exist different models for intrusion detection. The early models include statistical model and expert system model. The recent models include agent-based model data mining model, artificial neural network model, machine learning model, and immunity-based model. In order to overcome the traditional limitations of Intrusion Detection System, Intrusion Detection System design should adopt a more systematic and automated approach, based on data mining IDS is such an effective method. IDS based on Date Mining have a behavioral model through widely checking data. So it can accurately capture the actual invasion and normal behavior. This automated method no longer need to manually analysis and coding the invasion mode and no longer need to choose statistical methods by experience when build the normal using model. It is a major advantage of the same data mining tools can be applied to multiple data stream. This will be conducive to the construction of resilient Intrusion Detection System.

4. Data Mining Process of Intrusion Model

A research focus in Intrusion Detection realm is how to effectively find normal and abnormal behavior from abundant raw date and how to effectively generate automatic intrusion rules after collecting the raw data from network. To accomplish the task must study various data mining algorithms such as correlation analysis algorithms, sequence analysis algorithm, classification algorithms, etc. Correlation analysis data mining algorithms can be used for found the relationship between records of the network attributes. Sequence analysis data mining algorithm can find the timing relations between network connections records. Taking use of correlation analysis and sequence analysis data mining algorithm can find normal behavior patterns for Anomaly Detection model. Classification data mining algorithm can identify normal and invasion behavior rulers from training date. The process of Data Mining based on Intrusion Detect System is shown in Fig. 1.

Firstly, the process of Data Mining of Intrusion Detect System collects complete network data such as

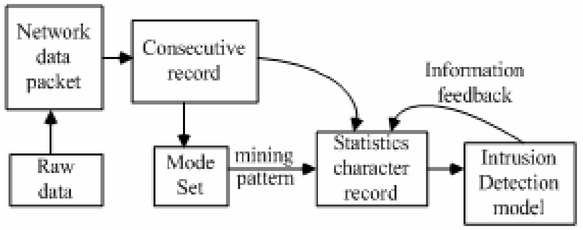

Fig. 1 Data Mining Process of Intrusion Detection

protocol type, purpose IP addresses and signs, and use correlation analysis and sequence analysis of data mining algorithms to deal with these connections, then can find association rules and sequence rules, and then can find the normal patterns of behavior which can be used for abnormal intrusion detection. Secondly, taking the normal behavioral patterns as standard to filter network connection date can gain the relatively high purity invasion data. Thereby, training data set can be rebuilt. Finally, the classification algorithm is used to rule mining, which can product rulers to be used for Misuse detection. Intrusion detection engine modules on the one hand takes categories rules as the judgment basis of misuse detection for detecting real-time network date scream., on the other hand, takes use of sequence rules and association rules mining module as the judgment basis of Anomaly Detection to judge whether network data is normal or not and sends the detecting feedback result to training data collection module.

5. Framework of Date Mining-based IDS

Intrusion detection system based on data mining which consists of following major components: data acquisition module, data sensor module, data pre-processing module, database module, data mining and rules describing based on ontology, rules library, detection engine, decision-making center, etc. The framework of IDS based on Date Mining is shown in Fig. 2.

Data sensor module is mainly responsible for collecting extern data from network, and then gives them to data pre-processing module to accomplish the following work: data filtering, data integration and conversion, data specification, data discretization and hierarchical data generation. Then, data which has been processed is stored in database. The system takes use of data mining technology to study database data and extracts the related behavior characteristics and rules, and then creates detection model. System status and corresponding behavior rules made through data mining are described based on ontology for sharing rules and complete presentation and are stored in rules library. Analysis model created by data mining engine and rules from rules library, then the conclusion is send to decision-making center. Decision-making center judges network behavior and carries out corresponding solution scheme.

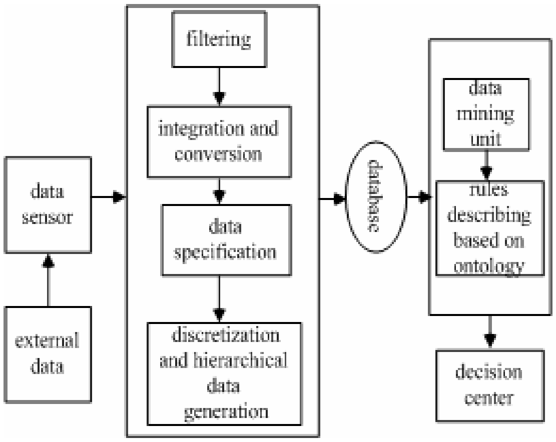
filtering
integration and conversion
data specification
discretization and hierarchical data generation
data sensor
external data
database
data mining unit
rules describing based on ontology
decision center

Fig. 2 Framework of IDS Based on Date Mining

6. Conclusion

In this paper, an Intrusion Detection System based on data mining technique has been proposed which performs anomaly detection, and then misuse detection is executed on the filtered network traffic. The system shows diversity, distributivity, adaptability, robustness and lightweight. It is a feasible framework and tries to explore a new approach to intrusion detection system. But, many problems about this system to be resolved in the future, such as how to automatically efficiently validate gained rulers, how to further prove efficiency of date mining, etc.

7. References

[1] ZHAO J Z, HUANG H K. An intrusion detection system based on data mining and immune principles, *Proceedings of International Conference*, 2002, 1:524-528

[2] Wenke Lee. A Data Mining Framework for Building Intrusion Detection Models. *In IEEE Symposium on Security and Privacy*, pp.120-132, 1999

[3] Hofmeyr, Forrest. Architecture for an Artificial Immune System. Department of Computer Science, UNM, Albuquerque, NM 87131, April 25, 2000

[4] Dasgupta D. Immunity-based Intrusion Detection System: A General Framework. *In the Proceedings of the 22th International Information System Security Conference*, pp.18-21, Oct. 1999

[5] Spafford. Zamboni. Intrusion Detection Using Autonomous Agents. Computer Networks, vol. 34, pp. 547-570, 2000