Analysis of Computer Security Forensics Based on Bayesian Network Intrusion Detection

Xiao Lijun
School of Information Technology and Media
Hexi University, Zhangye, Gansu, 734000, China
qie2871fuyong6@163.com

ABSTRACT: *With the rapid development of the Internet, network attacks and intrusion behaviors are becoming increasingly serious, and computer security issues have attracted much attention. In order to effectively respond to network attacks, researchers in the field of computer security have proposed various intrusion detection technologies. This article studies the application of intrusion detection technology based on Bayesian networks in computer security forensics analysis. By constructing a Bayesian network model, analyzing and inferring data such as network traffic and system logs, the detection and forensics of network intrusion behavior have been achieved. This method has the advantages of efficiency, accuracy, and adaptability, and can provide important technical support for the field of computer security.*

Keywords: Weighted Naive Bayes Method, Intrusion Detection, Computer Forensics Technology, Integrated Weighting

Received: 13 April 2023, Revised 28 June 2023, Accepted 19 July 2023

DOI: 10.6025/jisr/2023/14/4/87-95

1. Introduction

Intrusion detection is an important part of ICS network security defense, and it is an important means to protect system security. Intrusion detection has been concerned by experts and scholars at home and abroad [1]. Intrusion detection technology is a security measure to identify the integrity, confidentiality and availability of information resources. The purpose of intrusion detection is to classify normal events (normal) and abnormal events (anomalies) accurately in a large number of unknown network event data, so as to detect network attacks and reduce the false positive rate [2]. The disadvantage of misuse detection is that it is limited to the detection range of existing knowledge, and can't detect the attack behavior outside knowledge. Anomaly detection is based on the condition of the resource or the behavior of the user, rather than the detection criteria according to the specific behavior. In contrast, the applicability of anomaly detection can detect unknown attacks, rather than limited to misuse detection attacks, the main defect is the high rate of error detection, especially in the environment where many users, working conditions, system parameters and network structure are constantly changing [3]. At present, a variety of effective intrusion detection classification models have been proposed and adopted. With the rapid development and wide application of information network technology, the computer system and network security problems have brought unprecedented challenges to the human society, computer network crime cases have emerged in endlessly [4]. The main reason is that the technical level of the personnel in the IT industry is constantly improving, which makes criminal behavior covert and makes it more difficult to obtain evidence. The second is the existence and use of hacker software, which will increase the possibility of illegal use. Thus, the threshold of crime is reduced. Due to the increasingly serious problem of information security, the further development of network technology has been seriously hindered [5]. Therefore, it is very important to crack down on computer crimes. Because information security is of great significance to economic development and social stability, the computer security technology field is a new field, and computer forensics technology arises at the historic moment. As an important means, it will become a hot research topic.

2. Material and Methods

2.1. Bayes Network

In graph theory, the Bayes network is interpreted as a directed acyclic graph. In the graph, each node represents a characteristic attribute variable or a type attribute variable. When there is no conditional independence between nodes, there will be a directed edge between them to connect each other [6]. Each node maintains a corresponding joint probability table for it. The attached probability table of the node is the conditional probability of each attribute value in the range of attribute values when the parent node is known. If the node is a root node, its probability table represents the probability that each attribute value of the node occurs within the range of the attribute values, as shown in Figure 1.

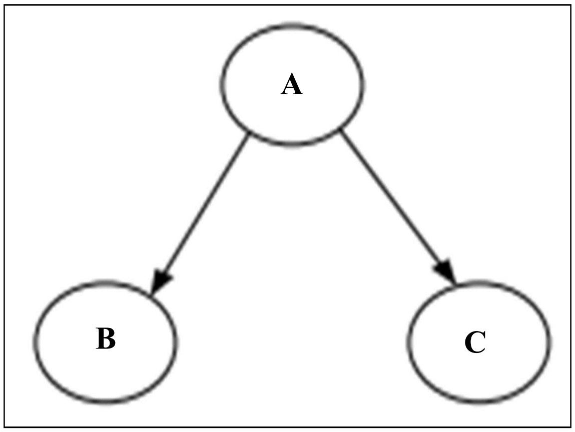

Figure 1. Computer forensics process

2.2. Naive Bayes Classification Algorithm

The naive Bayes classification algorithm is a relatively simple and effective classification method, which takes the Bayes theorem as the theoretical basis; its performance is comparable to the widely used algorithms such as neural networks, and decision trees, and even better performs in some fields [7]. The working process is as follows:

A feature vector represents each data sample.

The original data sample *X* is classified. In general, *X* is divided into the class with the greatest posterior probability value; in essence, the maximum value of *P*(*Ci*|*X*) is found, that is:

$$P(C_i \mid X) = \frac{P(X \mid C_i)P(G_i)}{P(X)} \tag{1}$$

The maximum value of $P(Ci|X)$ is calculated by maximizing the $P(X|Ci)P(Ci)$. If the prior probability is unknown, they are considered as equal probability, namely, $P(C1) = P(C2) = \ldots = P(Cn)$. Otherwise, the probability formula can be calculated by the prior probability formula:

$$P(C_i) = \frac{S_i}{S} \tag{2}$$

Among them, *Si* is the number of training samples, and *S* is the total number of training samples.

In order to reduce the computing time, the assumption of class conditions is independent when the number of attributes of the attribute set is relatively large, that is, each attribute value is independent of each others.

$$P(X \mid C_i) = \overset{n}{\underset{k=1}{\pi}} P(X_k \mid C_i) \tag{3}$$

If *Ak* is a discrete property, the probability can be calculated by formula (4):

$$P(X_n \mid C_i) = \frac{S_{ik}}{S_i} \tag{4}$$

Among them, *Sik* represents the number of training samples whose attribute *Ak* is *Xi* and belongs to class *Ci*, and *Si* represents the total number of training samples in class *Ci*. If *Ak* is a continuous property, it is generally considered to be the Gauss distribution.

X is classified; it is necessary to calculate the $P(X\,|Ci)P(Ci)$ of each class *Ci*. If the sample *X* is assigned to class *Ci*, the following conditions should be satisfied:

$$\mathrm{P}(X\,|C_i)P(G_i) > \mathrm{P}(X\,|G_j)P(G_i),\ 1 \le j \le m,\ j \ne i \tag{5}$$

Where *m* is the total number of classes. In other words, the class *Ci* that makes the maximum value of $P(X\,|Ci)P(Ci)$ is the class *X* belongs to.

2.3. Improved Synthetic Weighted Naive Bayes Algorithm

Covariance attribute weighting coefficient. Decision attributes refer to attributes that have a significant impact on classification. The conditional attribute refers to the rest of the attributes. In addition, the degree of correlation between different conditional attributes and decision attributes is also different. The system ρ which is composed of the decision attribute *X* and conditional attribute *Y* reflects the closely correlated degree the properties of *X* and *Y*, the greater the ρ is, the greater the effect of conditional attribute *Y* on the decision attribute *X* is, and vice versa. The formula of the correlation coefficient between attributes is:

$$\rho = \frac{Cov(X, Y)}{\sqrt{D(X)D(Y)}} \tag{6}$$

In order to ensure that the weight coefficient is positive, the weight is set as:

$$\alpha_1 = |\rho| = \left| \frac{Cov(X, Y)}{\sqrt{D(X)D(Y)}} \right| \tag{7}$$

Improved synthetic weighted coefficient. This paper proposes a new weight calculation method by comparing the above methods and combining the influence of different attribute values on classification results. N_{A_k} represents the number of the value of attributes A_k, and $N_{(A_k = m)}$ represents the number of sample objects whose attribute A_k is valued as *m*, and $N_{(A_k = m \cap Ci)}$

represents the number of sample objects whose attribute A_k is valued as m and belongs to class *Ci*. According to the different values of attributes, the weights are designed, and the weighted coefficient formula is expressed as:

$$\alpha_2 = \frac{\dfrac{N_{(A_k = m \cap Ci)}}{N_{(A_k = m)}}}{N_{A_k}} \quad (8)$$

Although the upper formula calculates the weights according to the influence of different values of each attribute on the classification, it considers the frequency relation of attribute value and does not consider the influence of the content of attribute value on classification. The covariance theory mainly uses the attribute value content to express the correlation between attributes; therefore, the combination of the two methods will yield more reasonable and accurate weighting coefficients.

Therefore, the improved comprehensive weighting coefficient is defined as: $\alpha = \dfrac{\alpha_1 + \alpha_2}{2}$

2.4. Computer Forensics

The characteristics of digital evidence. Digital evidence is any information or check value that is saved or transmitted in digital form. Compared with traditional evidence, digital evidence also has the characteristics shown in Table 1.

Characteristics	Meaning
Universality	Criminal evidence can be found in different regions and countries by using the Internet.
Concealment	Digital data is not directly visible to the naked eye, and must be provided with appropriate tools.
Fragility	In the process of searching for digital evidence, the original data may be modified or lost.
Diversity	It is reflected in the diversity of digital evidence storage and the diversity of forms of expression.
High-tech technology	With the development of IT technology, more and more scientific and technological means are used in crime hacker attacks. Faced with such a severe situation, digital evidence forensics technology is constantly updated and changed to meet the ever-changing computer technology.

Table 1. Characteristics of Digital Evidence

Computer forensics process. Due to the vulnerability of digital evidence, its authenticity and security are questioned. Therefore, in the entire forensic investigation, it is necessary to follow certain operating procedures and techniques so as to ensure that the proposed evidence is properly analyzed and the original evidence is not tampered. The process of computer forensics is shown in figure 2. Computer forensics can be divided into nine stages as shown in table 2.

Related technologies of computer forensics. Computer forensics technology has been widely recognized and recognized in computer information security. Although the technology still has some limitations, computer forensics has become an indispensable part of dealing with computer crimes with the development of computers and networks. The main techniques of computer forensics are introduced in Table 3.

Forensic stage	Concrete content
Detection and determination	In the network, through the combination of IDS and network deception technology, some behaviors in the network are monitored in real time, once a suspicious attack is found, it is immediately identified and determined; once it is determined to be illegal attacks, it is needed to access to data and analysis in real-time, and then implement appropriate measures. In addition, it is required to collect evidence as much as possible to ensure system safety.
Obtaining legal authorization	Since the evidence obtained from the computer will be used in the law, the process of investigation and evidence collection must be authorized by law, and each process of the investigation must be supervised by law. The technical scheme used in forensics and the software used should be legally recognized, otherwise, the evidence collected may be recognized as invalid data by the court.
Seal up and protect	Computer forensics must protect the computer against digital evidence and avoid restarting or running the application. The target computer system should be stored in a restricted access area to avoid any changes, injuries, data breaches and virus infections, and should stay away from the magnetic field, because the high magnetic field may cause data loss in the disk.
Data backup	In order to avoid the destruction of the original data, the original media is used instead of the original medium, and the original medium is used as the mirror image, and the byte stream backup method is adopted. Linux or DD command of UNIX system, DOS Disk Copy and other special tools such as NTI Safe-Back, Norton Ghost, and Encase in Guidance Software are used for disk backup; these backup tools can even copy bad sectors and CRC check error data.
Collecting data	The sources of digital evidence are: systems, networks, and other digital devices. The evidence extracted from the system mainly includes the existing normal files; hidden files, password-protected files and encrypted files; system log files; chat room logs; E-mail; backup media, etc..
Analytical evidence	The diversification of computer operating systems, application programs, and data format and data type makes the analysis of evidence very complicated. The main data analysis technology is to find and match keywords or key phrases in the acquired data stream or information flow. According to the system time and standard recording intervals recorded in step (2), the time line is established to determine correlations between events.
Submitting conclusion	Forensics data should be submitted to reliable digital evidence in court, and this phase should be taken according to relevant policies and regulations. When forensic evidence is submitted to the court, it is necessary to conduct a detailed record of the whole investigation and evidence collection procedures to illustrate that the evidence is reliable. In addition, in the process of investigation and evidence collection, the two-person rule should be followed, so as to prevent tampering with information and ensure the chain of custody.
Summary	The system of attack is restored; the countermeasures to prevent similar cases are put forward; and the safety measures are improved; the technical summary and experience are summarized.

Table 2. Computer Forensics Process

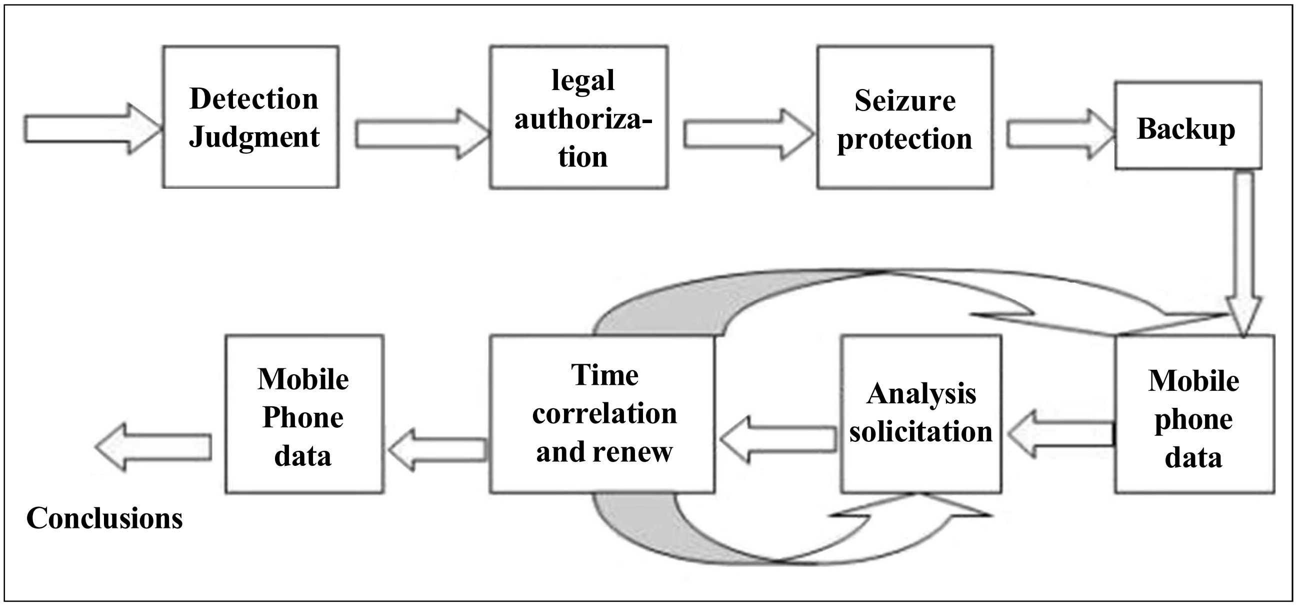

Figure 2. Computer forensics process

Forensics Technology	Concrete application
Source	The main sources of information for computer forensics include computer host systems, networks, and other electronic devices.
Classification	Computer forensics methods (forensics time) include static forensics and dynamic forensics. Static forensics is also known as passive forensics and afterwards, forensics, and dynamic forensics is also known as active forensics and real-time forensics. These two kinds of forensic methods depend on each other and have different emphases.
Forensic Tools	In the process of computer Judicial Forensics at home and abroad, the three most commonly used analysis and forensics software are respectively: EnCase 7, FTK 3 and X-ways Forensics. EnCase 7 is characterized by a very high degree of freedom, the user can quickly get to write script statements associated with the case of suspected data.
Forensics model	There are many kinds of computer forensics models, but the most commonly used is the law enforcement process model. The forensic model (forensics process) is divided into five stages, one of which is preparation.

Table 3. Related Technology Application of Computer Forensics

3. Results

3.1. Structure Learning

In this paper, a structure learning algorithm based on mutual information is proposed, which can be used to obtain the Bayesian network structure by measuring the dependence between attributes. The specific algorithm steps are as follows:

Step 1

The training sample set consists of N feature attributes: $\{X1, X2, ... Xn\}$. Supposing that R is a collection of characteristic attributes, and C is a collection of decision attributes. The mutual information $I\ (Xi, Xj)\ (i{\neq}j)$ in the case of $(Xi, Xj){\in}R$ and the conditional mutual information $I\ (Xi, Xj\ |C)$ in the case of the known decision attribute of C are calculated by using the previously mentioned mutual information and conditional mutual information formula.

Step 2

A constant $e1$ is taken in the real number range, which is a decimal fraction greater than zero, if $I\ (Xi, Xj) > e1$, an undirected edge is added between the attribute nodes Xi and Xj. After a round of operation, the draft $P1$ without the direction is obtained.

Step 3

In the real number range, a constant $e2$ is taken, and if it satisfies $I\ (Xi, Xj\ |C) < e2$, it can be considered that the condition of Xi and Xj is independent in the case of known decision attribute C. In this case, the algorithm can eliminate the edges between the two nodes; and after a round of traversal and trimming operations, a new $P2$ without direction is finally obtained.

Step 4

A constant $e3$ is taken in the range of the real, if it meets $I\ (Xi, C)\ \text{-}I\ (Xj, C) \leq e3$, then the directions between node Xi and node Xj have following two conditions: $Xi{\rightarrow}Xj$ or $Xj{\rightarrow}Xi$, namely, $ijXX$. If $I\ (Xi, C) - I\ (Xj, C) > e3$, and $I\ (Xi, C) > I\ (\mathrm{Xj}, C)$, then $Xi{\rightarrow}Xj$, otherwise, $Xj{\rightarrow}Xi$. The network structure graph $P3$ is obtained based on $P2$.

3.2. Intrusion Detection Data Set

In this paper, the experimental data was the *KDD'99* intrusion detection data set, the training dataset contained 7 weeks of network traffic, with 5000000 connection records; the training set contained the network traffic for 2 weeks, and there were 2000000 connection records. This study simulated 5 major types of network attacks.

3.3. Results and Analysis

To verify the correctness and effectiveness of the proposed algorithm, experiments and analysis were carried out. To ensure the efficiency of the implementation, 20000 records were randomly selected in the experiment, and the 20000 records were randomly divided into 5 groups, each group of data was used as training data, and 70% was used as the test data. An experimental building environment was carried out in the process of the Windows operating system platform; the processor was Inteli5, CPU frequency was 1.9 GHz, the memory was 4 GB, and the software tool was Weka. In the experiment, through the discretization and attribute reduction of the continuous attributes embedded in Weka software, the final condition attribute was obtained: serv-ice; flag; src-bytes; dst-bytes; dst-host-srv-count; diff-srv-rate. Various classification algorithms were used for experiments; the experimental results were obtained, as shown in Table 4.

Sample	1	2	3	4	5	**Average**
J48	97.43	97.41	97.54	96.57	98.23	97.436
NB	96.23	96.76	95.21	96.01	95.65	95.972
ASWNB	98.46	98.48	98.55	98.79	98.67	98.590
CWNC	98.76	98.45	98.68	98.89	99.01	98.758

Table 4. Comparison Table of Accuracy (%)

In this experiment, the improved naive Bayes classification algorithm was compared with other classification algorithms in Intrusion detection. It can be seen from table 1 that the proposed classification algorithm based on improved naive Bayes improves the classification accuracy compared with other classification algorithms, and proves that the proposed algorithm is effective and feasible.

4. Discussions

The data set used in this paper is the *KddCup99* data set, and this data set is very important experimental data in the field of intrusion detection, it consists of nearly 5 million samples, each of which corresponds to a network connection record. Each sample contains forty-one network feature attributes, and the forty-second attribute column is a mark of class attributes. The data set includes four types of attacks: *DOS*: denial of service attack; *R2L*: login without remote access from a remote computer; *U2R*: native super user access without permission; *Probing*: monitoring some computer ports. One of the normal types is marked as: Normal. First of all, the data is cleaned up, and the sample number is set to *n*, and the samples are divided into *k*=1+3.32**log*2 (*n*) groups. If the value of the attribute value 1 is in the range of [*min* ((*max-min*)/*k*), *min* (11)((*max-min*)/*k*)], then l is the result of the discretization of the attribute value. The partial probabilities obtained by calculation are shown in Table 5; after Bayes classification, the correct classification rate is shown in Table 6.

***X*9**	***X*1**	***P*(*X*1/*X*9)**	***X*9**	**X1**	***P*(*X*1/*X*9)**
1	0	0.6749	5	4	0.0099
1	1	0.2052	2	49	0.8060
2	4	0.0023	1	3	0.2796
3	0	1.00	1	2	0.7783
4	0	1.00	1	12	0.0001
1	36	0.0001	5	17	0.0090
5	0	0.8558	2	30	0.0002

Table 5. Partial Probabilities Calculated

Record Type				
Normal	Normal	Probe	R2L	U2R
92.13%	90.82%	81.02%	82.53%	79.12%

Table 6. Accuracy of Classification

5. Conclusion

The rapid development of the Internet has brought great convenience to people’s lives but also produced a lot of security problems. Therefore, the research on intrusion detection has attracted more and more attention. Based on the weighted naive Bayes method, an intrusion detection method which can better apply to complex system networks was proposed in this paper. This method constructs the Bayes network by using dependency relationships among attributes and classifies the samples, and it is a simple Bayes classification algorithm based on improved weighted coefficient. Finally, the experimental results show that the proposed algorithm can effectively improve the classification accuracy compared with other classification algorithms.

References

[1] Santos, I., Brezo, F., Ugarte-Pedrero, X., et al. (2013). Opcode sequences as representation of executables for data-mining-

based unknown malware detection. *Information Sciences, 231*, 64-82.

[2] Sindhu, S. S. S., Geetha, S., Kannan, A. (2012). Decision tree based light weight intrusion detection using a wrapper approach. *Expert Systems with Applications, 39*(1), 129-141.

[3] Davis, J. J., Clark, A. J. (2011). Data preprocessing for anomaly based network intrusion detection: A review. *Computers & Security, 30*(6), 353-375.

[4] Panda, M., Patra, M. R. (2009). A novel classification via clustering method for anomaly based network intrusion detection system. *International Journal of Recent Trends in Engineering, 2*(1), 1-6.

[5] Rehman, A., Saba, T. (2014). Evaluation of artificial intelligent techniques to secure information in enterprises. *Artificial Intelligence Review, 42*(4), 1029-1044.

[6] Ganapathy, S., Kulothungan, K., Muthurajkumar, S., et al. (2013). Intelligent feature selection and classification techniques for intrusion detection in networks: a survey. *EURASIP Journal on Wireless Communications and Networking, 2013*(1), 271.

[7] Linghu, H. Y., Chen, M., Wang, H. H., et al. (2009). Bayesian network intrusion detection method based on credibility of mutual information. *Computer Engineering and Design, 14*, 011.