Research on Intrusion Detection Method base on Cloud Computing

Mengmeng Cai¹ and Honglin Wang²
¹Drug Research and Development Center, Shandong Drug and Food Vocational College, Weihai, China
²Weihai Marine and Fishery Bureau, Weihai, China

Abstract. As a new generation of technology cloud computing has been widely applied to various fields, and owing to the problem of the security, the security problems of the data in the cloud computing platform has been becoming the focus of attention. Therefore, it is vital of cloud computing that to put forward a security mechanism to ensure the security of data in the environment of cloud computing. The related technology of Intrusion Detection is researched, and the thoughts of combining misuse detection and anomaly detection together are adopting. Then a method of anomaly detection based on data mining is given. It is the improved decision trees algorithm, and achieves a real-time updates of invading the signature database by misuse detection. It was improved through the experiment the application based on improved decision trees can detect the attack effectively and guarantee the security of the cloud computing.

1. Introduction
With the development of Internet, the Cloud Computing has been one of the topics of technology in the information age. Of course, while people have the convenience, the problem of security has been the primary problem which needs to be solved. The Cloud Computing security incidents that occur frequently in recent years has caused people's fears. On February 24, 2009, Google’s Gmail broke out the global error. On February and July of 2009, the Simple Storage Service which can be called S3 of Amazon broke off twice. Rackspace's Clouding Services discounted on June, 2009 and suspended again on November at the same year. The Microsoft outbroke BPOS interrupt service on September, which is a major event of the Clouding Computing security of the Microsoft. On March of 2011, the leakage of the users' data of Google's mailbox happened again. On January, 10, 2013, Dropbox service failed and made the service break for about 16 hours. On January 28, 2013, the Facebook Site interrupted two to three hours, causing the users could not update the condition and information. On January 31, 2013, the Cloud Computing service of Amazon failed again, the page of Amazon.com displayed the error message text for up to one hour. On March, 2014, eBay suffered data leak seriously. The occurrence of these security incidents makes us realize the importance of the Cloud Computing to solve security. The existing method of how to protect the security of the Cloud Computing and how to ensure that the data stored in the cloud do not access from unauthorized user and hacked is that we can use comprehensive security techniques such as data encryption, identity authentication technology, VPN, firewall, and so on to ensure safety of the applications of the Cloud Computing efficiently. However, these technologies can not completely guarantee the security of the applications of the
Cloud Computing. This article details the Intrusion Detection method and proposes the Anomaly Detection method based on data mining- the improved Decision Trees and applies it under the Cloud Computing environments. The experiment shows that the method can effectively detect intrusions that are known or unknown and ensure the reliability of the Cloud Computing services.

1.1. The summary and features of the Cloud Computing

The Cloud Computing [1] is an on-demand use, on-demand pay- mode, which can provide IT resources and capabilities from the Internet. IT resources include servers, networks, applications, storage, services, etc. The reason for giving the name of the Cloud Computing is that it needs computing resources including physical resources, development and testing environment, data and applications to exist on the "Cloud". The "Cloud" is actually a network. The Cloud Computing is resources and services based on the network. In short, the Cloud Computing is a way to pool the cluster computing power that provides flexible and on-demand service of new Internet technologies and new business for internal and external users through the Internet.

The Cloud Computing technology system includes IaaS (Infrastructure as a Service), PaaS (Platform as a Service), SaaS (Software as a Service).

IaaS provides pay-elastic infrastructure services for business users and its core technology is virtualization. Virtualization technology can make a physical device dynamic virtual device divide into multiple dynamic logically independent virtual devices, and then realize the full reuse of software and hardware resources. It can also make all the resources form a unified pool of resources that is transparent to the users. The "Cloud" is counted like water, as users can easily use the water via the pipes to charge according to the amount of water used. IaaS is convenient to provide a variety of resources to the user via the network.

PaaS faces to the Internet developers and provides the users with the platform such as software development, testing, deploying, operating environment and applications, etc. as a service via the Internet. Its core technology is distributed and parallel computing.

SaaS faces to the individual users and provides a variety of online software services through the browser on the server side software as a service for users.

The Cloud Computing has these distinguishing features such as network-centric, to serve as provide ways, pooling and transparency resources, highly scalable and high reliability.

(a) network-centric: The Cloud computing is providing services to all users via the network, and its overall structure is composed of a network connection exists in the network.
(b) to serve as provide ways: The users just need simple import or export to satisfy their demands.
(c) pooling and transparency resources: For the offers of the Cloud Computing services, all resources are managed and controlled unifiedly and constitute a resource pool in order to provide on-demand service to users. For users, all resources are transparency and users just need to concern about whether their needs can be met.
(d) highly scalable and high reliability: In order to satisfy large numbers of users' need, resiliency services of the Cloud Computing can dynamically stretch to reflect its high scalability, while the "Cloud" can adopt with a fault-tolerant data protection services to high reliability.

The rapid development of the cloud computing has brought a lot of convenience. In addition, the cloud computing system and resources are usually in the form of plain text stored in a shared environment, encryption technology alone is unlikely to protect the security of their data. Therefore, it is the most important to find a viable and effective solution to prevent unauthorized users or lawless invasion of system resources such acts.

2. Intrusion Detection Technology

Target of intrusion detection technology is to take certain security policies to detect attacks various acts such as unauthorized user access to data and hackers. By detecting aggressive behavior to ensure the confidentiality of data, integrity and availability. The main strategy is divided into intrusion detection misuse intrusion detection [2] and anomaly intrusion detection [2].
2.1. Misuse Intrusion Detection

Typically the data collected with the predefined features in a variety of knowledge base invasion compared to a pattern, where the knowledge base is characterized by the behavior characteristics collected abnormal operation established, by comparison, if found to have aggressive behavior, it is determined that the invasion. Misuse intrusion detection of known intrusion is very effective, but can not detect unknown intrusions. A method of misuse intrusion detection process as shown in Figure 1:

![Figure 1. Misuse intrusion detection.](image)

2.2. Abnormal Intrusion Detection

Firstly, to establish a user profile, the subscriber profile is characterized in having a normal operation, then the current active user with pre-established comparing these normal operation detects whether the invasion. Anomaly detection for known attacks is very sensitive, but the key is to determine the threshold. Therefore, this paper will combine detection and anomaly detection into their works shown in Figure 2.

![Figure 2. Working principle of the misuse detection and anomaly detection.](image)

Firstly data packets are converted by the misuse detection unit analytical testing, if they meet the intrusion features, directly take the appropriate strategy, otherwise it is handled by the abnormality detection unit, if unusual, and it is judged as suspicious behavior.

For anomaly detection unit, we use data mining techniques. Data Mining [5] is also known as knowledge discovery in databases. Data mining technology is mainly for people to extract meaningful information from large databases, such knowledge to be extracted are generally expressed as concepts, rules, forms and patterns. Intrusion detection mainly has two directions; the first is one tap from the audit data in the invasion of rules and patterns, and detection methods combined with pattern matching; the second is applied to anomaly detection, first identify the user profile, and the establishment of user normal behavior signatures.
A decision tree is based on a data classification data mining algorithms. Therefore, an improved decision tree will offer into intrusion detection system used to predict the invasion.

3. Decision Tree based on Intrusion Detection

3.1. The thought of Decision Tree Algorithm
Decision Tree [7] is a data mining algorithm, which uses a recursive way from top to bottom. When learning a classification model for the training set, the first is compared to the internal nodes of the decision tree based on the first attribute value, and then the branch for different attribute value determination by the coupled node tree the leaf node to reach the final conclusions for.

The resulting decision tree is a tree structure, composed of the root, non-leaf nodes and leaf nodes. The top is the root node of the tree and each corresponds to a value of the property on a branch. Each tree represents a non-leaf node test on an attribute, each branch represents a test output. Leaf node of the tree represents the distribution of the class.

3.2. Decision Tree algorithm based for independence test
The enter of the tree construction algorithm is a training set, which consists of large amounts of data structure, exist as the way of "Properties-value". The training focused on each piece of data that contains multiple attributes features and a class label, structure decisions is a binary tree or tree. Non-leaf nodes represent attributes the decision tree, each branch is all the value of the property.

There are two concepts associated with the decision tree algorithm: information entropy, information gain.

Let S be the set of training samples, all samples which belong to the n-th categories, respectively C_1, C_2, ..., C_n, to represent the n categories, Then the entropy S (entropy) can be as follows:

\[
\text{entropy}(S) = -\sum_{i=0}^{n} p_i \log_2 p_i
\]  

(1)

P_i represents the probability of class C_i. Entropy can be used to measure the training set. When it is not pure, entropy is larger, the higher the purity of the training is not set, but the decision tree to reach the goal is to make each branch as low as possible without the purity of the sample after the subset that entropy of each subset as small as possible.

A set of the S property is divided into m subsets, then according to the entropy S can be divided as follows:

\[
\text{entropy}(S,A) = \sum_{i=1}^{m} \frac{|S_i|}{|S|} \text{entropy}(S_i)
\]  

(2)

S_i is the i-th subset S A divided according to the obtained, | S | and | S_i | denote the number of samples S and S_i.

Information gain is used to measure the change in the value of the desired partition before entropy, and then the use of property A gain information on S can be divided as follows:

\[
\text{Gain}(S,A) = \text{entropy}(S) - \text{entropy}(S,A)
\]  

(3)

Gain (S,A) represents because of knowing the value of the property resulting from the expected value of A Minimizing Entropy. The bigger Gain (S,A), the more information that attribute A can provide and the better for the sort.
The chi-square independent test is used to determine the relationship between two categorical variables. However, this article is to study the results of comparison of the two algorithms. Hence, this paper selected in the test for independence is the chi-square test. The chi-square test is calculated as follows:

$$X^2 = n \sum_{i=1}^{r} \sum_{k=1}^{c} \frac{n_{ik}^2}{n_i n_k} - 1$$  \hspace{1cm} (4)$$

In the formula, $n_{ik}$ denotes the $i$-th attribute value of the property; $n_i$ and $n_k$ are the row or column value of the property and on behalf of the relevance of this attribute and decision attribute.

3.3. Classification Algorithm combining Information entropy and independence test

The formula can be shown as follows:

$$G(x_i) = x^2 + \text{Gain}(x_i)$$  \hspace{1cm} (5)$$

In the formula, $G(x_i)$ represents classification select Properties, size $\text{Gain}(x_i)$ on behalf of information gain, as shown in equation (3). Training focused on the use of the properties are calculated, and finally select the category property $G(x_i)$ the greatest attribute as a root node classification, recursive partitioning, until the termination condition is met.

By the decision tree classification algorithm constructed relatively small due to the fitting of the phenomenon and the training examples exist, so the test results decline, while the main method is to avoid fitting after tree pruning, pruning methods there are many here is not too much introduction, this paper is presented after Quinlan[17] REP approach to pruning.

After the tree structure is completed, the classification of unknown samples to determine its category. Analyzing the process starts from the root node of the decision tree, from top to bottom in turn searches until the corresponding leaf node, this time, and the leaf nodes of the class label represents the category of the unknown sample.

As decision trees can effectively detect unknown intrusions, therefore, we propose a decision tree based detection to update signature database method, we all know that misuse of the known intrusion detection has good detection efficiency, and the unknown the invasion is often not very sensitive. In this paper, we can use the tree structure of the unknown intrusion detection and judgment, and make the appropriate records, the record is added to detect misuse of signatures in order to achieve real-time updates of intrusion signature database, while retaining the misuse detection efficiency while expanding its ability to detect unknown attacks, again when such a request, will be by misuse detection unit determines directly, eliminating the need for the anomaly detection process again, but also to further improve the efficiency of detection.

4. Application and results of decision trees

When using decision trees for intrusion detection, intrusion detection training set is used in KDD Cup99 [18]. As shown in Figure 3, the data set includes test data and training data to identify unidentified. In this paper, experimental platform for Windows XP, CPU for Intel Core i3, RAM 4GB, HDD 500GB, MATLBA experimental environment.
Each record experimental training examples before the forty-first represent characteristic attributes, the last represents the class attribute. Experiments using two detection rate and false alarm rate index as a measure of the mechanism results. Detection rate is the probability of an attack when the system can correctly detect the intrusion. False positive error rate is normal behavior will determine the probability of occurrence of an attack.

\[
\text{Detection rate} = \frac{\text{number of correctly detected intrusion}}{\text{total Number of invasion}}
\]

\[
\text{False alarm rate} = \frac{\text{false alarm rate of false positives}}{\text{total number of normal behavior}}
\]

Through simulation experiments, firstly train the test set. Then for the test set and improved decision tree algorithm use decision tree algorithm for intrusion detection, the last set of tests were obtained curve is shown in Figure 4.

As can be seen from the experimental results figure, compared to the traditional algorithm, the improved algorithm for aggressive behavior has a high detection rate.

5. Conclusion
With the continuous development of the Internet, the occurrence of emergence of cloud computing environments under a variety of intrusion is not completely blocked. Therefore, this article will misuse...
detection and anomaly detection technology integration, an improved decision tree algorithm to determine the occurrence of intrusion and invasion of behavior judged update to misuse intrusion detection unit features library that implements the features of invasion Live Update library. The experiment shows the improvement of the decision tree algorithm clever use of further improved performance and efficiency of detection.

References
[1] Zhang Weimin, Tang Jianfeng, Luo Zhiguo, Qian Ling. Cloud Computing: The profound change in the future [M]. Bei Jing: Science Press, 2009.
[2] Xian Yongju. Intrusion Detection [M]. Xi'an: Xi'an University of Electronic Science and Technology Press, 2009. 8.
[3] Feng Dengguo, Zhang Min, Zhang Yan, Xu Xia. Study on Cloud Computing Security [J]. Journal of Software, 2011, 22 (1): 71-83.
[4] Yang Jian, Wang Haihang, Wang Jian, Yu Dingguo. Research Review of Cloud Computing Security Issues [J]. Journal of Chinese Computer Systems, 2012, 33 (3): 472-479.
[5] Liu Qiang. A design and implementation of intrusion detection system based on data mining [D]. University of Electronic Science and technology, 2013.
[6] Wang Xin, Huang Yanlei. A new intrusion detection mechanism [J]. Computer Systems & Applications, 2011, 20 (5): 65-68.
[7] Tom M. Mitchell, Zeng Huajun, Zhang Yinkui. Machine learning [M]. Bei Jing: Machinery Industry Press, 2003.
[8] Wei Guangke. Analysis of anomaly-based intrusion detection technology [J]. Computer Engineering and Design, 2005, 26 (1): 107-109.
[9] Kang Shiyou. Intrusion Detection Model based on data mining and feature selection [J]. Microelectronics & Computer, 2011, (08): 74-76.
[10] Zhang Hui, Xing Peizheng. Analysis of Information Security in Cloud Computing Environment [J]. Computer Technology and Development, 2011, 21 (12): 164-166, 171.
[11] Zhou Qian, Yu Jiong. Defense system trust model based on Cloud Computing [J]. Journal of Computer Applications, 2011, 31 (6): 1531-1535.
[12] Chen Kang, Zheng Weimin. The present situation and the case of cloud computing [J]. Journal of Software, 2010, 27 (2): 429-433.
[13] Jiang Wei, Gao Zhixin, Li Benxi. Intrusion Detection Based on multi dimension cloud model [J]. Computer Engineering, 2006, 32 (24): 155-156.
[14] Ahmed Patel, Mona Taghavi, Kaveh Bakhtiyari, Joaquim Celestino Junior. An intrusion detection and prevention system in cloud computing: A systematic review [J]. Journal of Network and Computer Applications, 2013, 36 (1): 25-41.
[15] Huai bin Wang, Hai yun Zhou, Chundong Wang. Virtual Machine-based Intrusion Detection System Framework in Cloud Computing Environment [J]. Journal of Computers, 2012, 7 (10): 2397-2403.
[16] Abd alal Alkhalid, Talha Amin, Igor Chikalov. Optimization and analysis of decision trees and rules: dynamic programming approach [J]. International Journal of General Systems, 2013, Vol. 42 (6): 614-634.
[17] Quinlan J R. Simplifying decision trees [J]. International Journal of Man-Machine Studies, 1987, (27): 221-234.
[18] KDD99. KDD99cupdataset [DB/OL]. http://kdd.ics.uci.edu/databases/kddcup99, 1999.