Research on the Large-scale Network Intrusion Mode based on Principal Component Analysis and Drop Quality Sampling

Yanmei Zhang

Abstract

In this paper, we conduct research on the large-scale network intrusion mode based on the principal component analysis and drop quality sampling. With the growing of network security issues, invasion detection becomes the study hotspot. There are two main types of that invasion detection technology, the first is that misuse detection and the anomaly detection. Misuse detection can more accurately detect attacks, but high non-response rates, anomaly detection could detect the unknown attacks, but higher rate of false positives. Network invasion detection problem is summed up in the network data flow of discriminant problem, namely the judgment of network data flow is normal or malicious and in this sense here invasion detection problem can be understood as a pattern recognition problem. Our research integrates the PCA and sampling technique to propose the new idea on the IDS that is innovative and will promote the development of the corresponding techniques.

Keywords: Network Intrusion, Principal Component Analysis, Drop Quality, Sampling, Scale.

Introduction

Rapid collection for mass data processing applications, system in addition to having the requirement of real-time rapidity, because of the limitation of memory capacity in the process of collecting all cannot one-time acquisition into the memory, but to set certain disk buffer, acquisition and inventory protection, thus also requires system has fast speed of R/W disk. From the point of view of rapid processing, acquisition of the data to continuously distributed on disk, so that they can quickly R/W. Additional file system to manage disk space as large as possible. With the progress of data acquisition and transmission technology, make it possible to have a short period of time accumulated a large number of historical data. At the same time, the current Internet environment data stream processing application long-term, comprehensive and accuracy demand of enlarging the scale of historical data. For mass historical data real-time processing of the data flow demand with the contradiction between computing and storage capacity is insufficient to become a new challenge in the field of that cloud computing and data analysis. In the face of growing network security threats, traditional static security technology such as authentication mechanisms, such as encryption and firewall has been difficult to do. Intrusion detection as a kind of that important dynamic security technology can provide external attacks on computer systems and networks, internal attacks and wrong operation and fully tested.

- Misuse detection model, which is based on the characteristics of the test model: collection of abnormal operation behavior characteristics, to that establish related the feature library, when monitoring the user or the system behavior match the features in the library, the system will think that this kind of behavior is invasion.
- Anomaly detection model: first, summarize the behavior that should have normal operation characteristic, when the user activity and have the significant deviation from normal behavior that is considered to be invaded [1-3].

1 Guangzhou Vocational College of Science and Technology, Guangzhou, Guangdong, 510550 China
Intrusion detection technology is a new network security technology in recent years, the purpose is to provide real-time intrusion detection and take the corresponding protective measures, such as logging, disconnect the network connection, etc. It to detect and control technology for nature, it plays the role of active defense that is a very important part in network security. Based on the literature and review, we could summarize the methodologies into the listed aspects. (1) Bayesian technology. The priori information and sample data information is used to determine the posterior probability of the event, and according to the probability that the relationship between variables to build graph model, in order to solve the uncertain knowledge in intrusion detection. (2) The immune system. Defining the system normal pattern library, and then randomly generated many models with that the self-defined comparison, and if it matches any one self-model, then the random pattern is discarded. Otherwise, it will appear as a mature recognizer to match the system anomaly pattern. (3) Model predictive. The method has good adaptability to the change of user behavior. In IDS can detect attempts to predict rule learning period training system of the invaders has better ability in their own defense, detection speed. (4) The neural network. The basic idea is to use learning algorithm of neural network method to study on the relationship between the input and output vectors, and summed up new relationship between input and output, thus to judge. Neural network method can be easily express nonlinear relationship between variables, and automatically learn and update [4].

In this paper, we conduct research on the large-scale network intrusion mode based on the principal component analysis and drop quality sampling. In later sections, we will discuss in detail.

The Proposed Methodology

The Signal Sampling. Signal sampling can use the Nyquist sampling theorem, but the theory is used for high bandwidth signals, they paid a high price: for 1 MHZ signal only 2 MHZ sampling rate, the difference only 1 MHZ, and the signal of 1 GHz surged to 2 GHz sampling rate, up to 1 GHz, Nyquist sampling theorem in the sampling rate requirement is doubled, the high frequency signal sampling rate demand higher. The actual source most signals are analog signals, all the source signal generation set is composed of a signal. For a signal collection, under what conditions, can use the sample value of the signal in time domain and restore sequence said out, is a fundamental problem of signal theory.
Answered for band-limited signal, sampling theorem of sampling points per unit time as the signal frequency band twice or more than twice signal on any of the signal will be sample value sequence. Random transform sampling due to the time of using the so-called stability trigger point as measuring datum, so on the implementation technology of the input signal waveform has great limitations: it is only applicable to cycle there two points corresponding to the same level of the signal, the complex signal failure; On the other hand, it is referred to as random sampling, actually is to use the principle of equidistant trigger pulse and the input signal is not relevant, so sampling blind area exists.

A signal in time domain can be expressed as:

\[ x(t) = A\sqrt{2} \sin(2\pi ft + \phi) \]  

(1)

In most cases, the signal sampling before is impossible to know the actual value of that the signal frequency, but know the signal frequency band. In signal frequency under the condition of the actual value, only at the initial sampling period to determine sampling period and the signal is sampled for the first time to estimate the actual value of the signal frequency.

Corresponding to the fundamental wave signal is signal collection, reflects the characteristics of the entire signal collection, while the fundamental wave that coefficient corresponding to the specific signal, this type of signals of the different reflected in the corresponding wavelet coefficients are different. In fact, the analog signal is most actual source has the decomposition characteristics, can prove that the classical theory of the band-limited signals have the same features, thus solving the problem of this kind of signal sampling theory has practical significance and according, the revised sampling function could be listed as the follows [5-6].

\[ x(n) = A\sqrt{2} \sin\left(2\pi (n-1)T_s / N + \phi\right) \]  

(2)

**The Principal Component Analysis.** The principle of principal component analysis is to try to into a new set of the original variables were independent of each other a few variables, at the same time, according to the actual need to take out a few less the sum of the variables as much as possible to reflect the original statistical methods of information called principal component analysis or called principal component analysis also is a kind of mathematical processing dimension reduction method.

![Figure 2. The Demonstration of the PCA.](image)

PCA robustness consideration mainly has two general aspects. Consider how to achieve output independent among principal components. Because, the eigenvalue decomposition is equivalent to the linear transformation covariance matrix into a diagonal matrix, its off-diagonal element is zero. So we can put a multi-input problem is decomposed into the several independent single input issues to consider as the follows.

\[ f(x) = \frac{1}{2^{n/2} |S|^n} \exp\left\{ -\frac{1}{2} x^T S^{-1} x \right\} \]  

(3)
The existing main method is based on the known input sample distribution, the introduction of appropriate nonlinear processing link the so-called nonlinear PC the algorithm is put forward. So, it is considered the higher-order statistics characteristic of the input, so as to realize the output of the principal components independent of each other. Intuitively, it refers to sample focused most of the sample distribution difference too few samples, their existence makes the PCA results will appear a lot of errors. "Bad" reasons are various, such as sudden and random noise, measure and record the occasional error and so on. In addition, because the number of the samples is limited, even if all the samples are produced by the same distribution may also be because of a lack of sample making in which a small amount of sample actually the "bad" samples with energy function of the formula 4.

$$J_{\text{energy}} = E \| e \|^2 = e \| X - U \|^2$$  

(4)

The Network Data Collection. In network management, often to the network transmission of data analysis, to detect the abnormal situation and take timely measures to ensure the normal operation of the network, in order to achieve this goal, often need to capture, save the datagram. However in the network transmission of data quantity is huge, especially the datagram low-level protocol, its number is amazing. Due to the amount of data to be processed is very large, if not by means of the database system, to filter and analyze these data is very difficult, in addition to deal with this amount of data, some small database has been difficult to do. Network data acquisition circuit includes the analog multiplexer switches, front signal amplification filter module, A/D converter, MCU controller, level conversion circuit, from the station controller and I/O channels, etc. Sensors equipment physical information into corresponding electrical signals, after transform into digital signals, microcontroller after processing by the network went out of the station controller for data transmission.

Coming from the sensor measured signal amplitude range is from several millivolts to dozens of millivolts, and therefore must be in front of the A/D converter of level with appropriate preamplifier bias circuit, in order to make full use of the full range resolution of A/D converter. After the amplifier stage added active filter, reduces the power supply system, space electromagnetic interference noise, further improved the precision of data acquisition. Our purpose is on the premise of not to use that hardware bag collector, through improved measurement system architecture and software structure, improve software caught performance make it can be applied to the high speed network environment.

The Large-Scale Data. The output of the digital sensors generally has two forms: according to certain ways of encoding digital signal and the pulse signal. Before a way of digital coding and has a one-to-one relationship between measured signal and its measurement precision is fixed after a way of pulse width and the frequency correspond to the size of the measured signal, sampling system to measure the pulse width or frequency is needed to get the value of the measured signal, its measuring accuracy is determined by the sampling system [7].

In phase control, on the basis of the phase controller between the stages by adjusting the worker threads to make full use of the CPU, improve the overall throughput. Thread resources between phase adjustment also uses the feedback control method, the existing SEDA method by observing the phase data within the input buffer size to adjust the thread pool size, increase the thread when the buffer data more than the threshold value, on the other hand, reduce the thread. This method without considering the system CPU usage the information between each stage of the global control that cause the CPU utilization to improve the global performance cannot be converted to optimization.

Ensure accuracy and improve anti-jamming improvement methods: in practice, the output of the digital sensors are not always in a state of stable signal output, the general situation is to switch from the signal source or before sampling to digital sensors with electricity, which will result in sampling start sensor output pulse width has a gradual and stable process, in addition, other interference can cause during the sampling process of the single pulse width and other pulse difference is very big, although these factors caused by sampling signal mutation can be fixed in the subsequent processing,
but can get close to the actual signal when sampling correct samples values is very important, it can be appropriate improvement through the composite sampling method.

**The Network Intrusion Mode.** Intrusion detection is extremely beneficial complement to the firewall. Intrusion detection is considered to be the second after the firewall security gate, in the case of does not affect the network performance to network monitoring so as to provide the internal attack, exterior attack and misoperation real-time protection, greatly improved the security of the network. Intrusion detection system can be divided into centralized intrusion detection systems and intrusion detection system which could be demonstrated as the follows respectively.

- Distributed into the number of field operating data analysis and detection system components to be monitored is proportional to the number of hosts. With the increase of the complexity of network structure, network resources are generally distributed, and intrusion activities also gradually with a collaborative, so a centralized IDS is exposed its limitations [8].
- Centralized into the detection system in a fixed number of field data analysis, including the host-based and that network-based IDS. Host-based IDS audit data analysis of the host and monitor a host directly, often combined with the operating system.

Statistical analysis of the key from the description of the state and the behavior of the system or network select a set of the statistical measurement attributes, and according to the historical data to establish the normal range. The advantage of this approach is basically don't need much a priori knowledge about invasion of technical details, have a mature statistical techniques can be applied, the trace of a set of statistics on the system resources occupancy rate is low. Statistical analysis of the limitations of first reflects on the question of statistical measures. With collection point of view, Snort characteristic features of the library of one of the characteristic collection, when a message to Snort will check these four parameters, look to the current feature set inside of feature match, if don't match we transferred to another feature matching in the collection. If a message matches the characteristics of the current collection of four parameters, on the various characteristics of the current feature set for a single match, the characteristics of the parameters of the other part is in order to match. When the current feature set match, the packet transfer on to the next feature set to repeat the above work, so on. On this basis, PCA will enhance the IDS system well.

**Conclusion**

In this paper, we conduct research on the large-scale network intrusion mode based on the principal component analysis and drop quality sampling. With appearance of new attack methods, especially the emergence of some collaboration intrusion behavior, to study brings new topics in the field of the intrusion detection. Early centralized intrusion detection system cannot effectively prevent this kind of invasion. Therefore, research of distributed intrusion detection system is very necessary. In this paper, we integrate the principal component analysis and drop quality sampling to enhance the basic IDS system. In the future research, we will simulate the method for testing effectiveness.

**References**


