Front. Electr. Electron. Eng. China 2010, 5(1): 85–90
DOI 10.1007/s11460-009-0071-5

RESEARCH ARTICLE

Xinliang WANG, Fang LIU, Luying CHEN, Zhenming LEI

Anomaly traffic detection of database network based on flow statistical features

Abstract The traditional intrusion detection system has the problem of high false positive rate and false negative rate. This paper deeply analyzes the differences of statistical features between single-flow and multi-flow on the database network, and presents a group of features that are easy to acquire and can be used to detect the anomaly in database network efficiently. By applying this group of features in Fisher algorithm for anomaly detection, the false positive rate and false negative rate are dramatically reduced. Simultaneously, the model made by using the group of features has the advantages of low algorithm complexity, good detection result and strong generalization ability. Experimental results show that there is higher accuracy when using the features of single-flow and multi-flow to construct the anomaly detection model than only using single-flow features.

Keywords anomaly detection, flow, statistical feature, Fisher, statistical package for social sciences (SPSS)

1 Introduction

In this paper, a network mainly constructed of database servers is described as a database network. Nowadays, network security incidents happen frequently, and the database as the data heart of the enterprise is facing more and more threats. However, the database in many enterprises has not been properly secured, and malicious hackers are utilizing some very simple attack methods to violate the database. For example, attackers can crack weak passwords or utilize default user name and password to invade the system, and they can also utilize vulnerabilities of unused or unnecessary database service and function to achieve the root privilege or attack the database by using structured query language SQL injection [1,2] and so on. For endless attacks, most system administrators try to protect the database server against vulnerabilities by updating and optimizing the security performance of web application and so on. However, the effect is not very ideal.

Currently, the main intrusion detection techniques are misuse detection and anomaly detection. The former is applied in many signature-based commercial systems [3–5] for detecting intrusion by analyzing the signatures of known attacks; however, it cannot detect unknown attacks. The latter detects the anomaly through identifying the host or network abnormal behavior patterns [6]. It assumes that behaviors between attackers and legitimate users are different to some extent, and we can detect unknown attacks by building normal and abnormal behavior modeling. However, it is not strict and has the problem of high false positive rate and false negative rate. Meanwhile, some well-known commercial systems can detect the database SQL injection attacks efficiently. However, database is suffering various attacks, and it is not enough to satisfy the need of database network anomaly detection only by using signature-based technology. In this paper, the statistical features of single-flow and multi-flow can be used to construct a more accurate anomaly detection model based on the features of the database network.

2 Data acquisition

As shown in Fig. 1, many database servers and web servers exist on the database network, and data acquisition equipment is deployed on the network entrance to collect all the in-out traffic directly accessing database servers. Different network databases usually use different ports, and the web application based on the database server is almost developed based on the browser/server (B/S) model. If the system load is not high, the web application and database will be mostly installed on one server;

Received June 30, 2009; accepted September 10, 2009

Xinliang WANG (✉), Fang LIU, Luying CHEN, Zhenming LEI
School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China
E-mail: wangxinliang@bupt.cn

otherwise, they will be installed on different servers. In order to ensure the speed of accessing the database, web application and database are usually installed in the same internal network because it can save bandwidth more effectively and improve access speed. Normal users generally access databases indirectly by the web applications that generate database traffic lying in the internal network, and the data acquisition equipment cannot capture normal internal traffic. Of course, there may be a small amount of traffic caused by the database administrator using external IP to maintain the database remotely. Therefore, the anomaly traffic is dominant in the collected database traffic; in contrast, normal traffic occupies a very small proportion. The main point of the research in this paper is how to detect the anomaly traffic of database network effectively, classify the normal and anomaly traffic accurately, and reduce the false positive rate and false negative rate.

This paper chooses SQL server as the focus of research and makes an intensive study. The traffic acquisition equipment was deployed on the entrance of an Internet data center, and it monitored all the in-out traffic of the SQL server database network in real time. The database traffic was collected in real time on March 12th, 2009, March 13th, 2009, March 15th, 2009, and March 19th, 2009, and collected data are shown in Table 1.

3 Packet analysis and feature extraction

3.1 Packet analysis

It is hard to determine accurately whether the traffic is abnormal or not only by using single-flow features. Based on database traffic data collected on March 19th, a deep study on the distributions of the flow duration and the number of packets were made. The statistical result shows that the duration of 97% abnormal flows is shorter than 1 s, and the duration of normal flows is generally longer. By the analysis of the packet number of flows, we find that the packet number of 84% abnormal flows is between one and five, and the packet number of 97% normal flows is larger than five. That is to say, the packet number of flows can reflect the differences between the normal flows and abnormal flows to some extent. Although the number of packets and duration can be used to differentiate between normal flow and abnormal flow, it is not enough to determine the types of the flow only by using the two features. Before launching the attacks, most attackers will scan the servers or open ports to search for loopholes, and the duration of this kind of attack is shorter and the packet number of flows is smaller. While the attackers confirm their attack targets, they will probably set out long-duration and continuous attacks that generate attack flows of longer

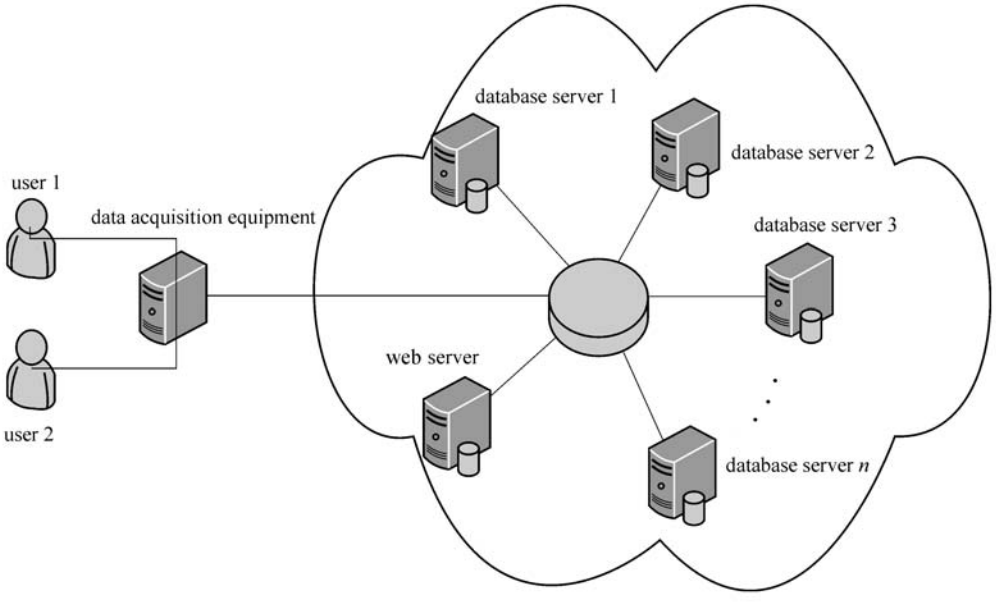

Fig. 1 Database network topology

Table 1 Data collection

date	count of all flows	count of abnormal flows	count of normal flows
March 12th	40456	40284	172
March 13th	34955	34791	164
March 15th	41928	41928	0
March 19th	28506	28324	182

duration and larger packet number. Therefore, it is not accurate to detect abnormal flow only by analyzing the duration and the packet number of flow.

The behavior of the attacker and the normal database user is different. For example, a normal user can query or modify the database after finishing the security certification, but the attacker may need to probe to get the password repeatedly. The two types of behaviors will lead to the differences in every section of the session of the transport layer (transmission control protocol (TCP) layer). For example, in order to access the database, normal users should first establish a TCP connection with the database through three-way handshake in which synchronize (SYN) packet, acknowledgement (ACK) packet, and SYN/ACK packet are generated. After the connection is built, it is going to transfer the data which are mainly constituted of ACK packets. After all the data are completely transmitted, the user will utilize the four-way handshake to close the connection. In contrast, the attackers will possibly scan the database before attacking, and the packets scanning result is mainly constituted of SYN packets, SYN/ACK packets, reset (RST) packets and RST/ACK packets. The well-known SYN flooding attack establishes the half-connection for denial of service, and the sessions it produces mainly consist of SYN packets and SYN/ACK packets. By comparing the types of packets and proportion of different packet types, we find that some significant differences that are caused by different access methods and behavior ways exist between normal users and abnormal users. Therefore, we can effectively improve the accuracy by regarding each type of packet number as classification features.

The features of the single-flow cannot reflect the relationship of the multi-flow (a set of flows in specific condition), and the flows normal user and abnormal user generate by accessing the database have a certain internal logical relationship. Therefore, it is not enough to detect anomaly only by analyzing the features of single-flow, and will affect the anomaly detection ability. On the database network, there should generally be only one or two external IP to maintain one database server. If one external IP accesses many database servers or many ports of one server at the same time, its behavior will deviate from the normal user behavior and cause the differences of multi-flow features. Therefore, it can effectively improve the accuracy of anomaly detection by applying multi-flow features for anomaly detection.

3.2 Feature extraction

Definition 1 [7] Flow consists of a set of packets generated in two-way communication of the application processes of two hosts under the constraint of overtime. That is to say, flow can be identified by the same couple of IP addresses, the same protocol (such as TCP, user datagram protocol (UDP), Internet control message protocol (ICMP), and so on) and the same couple of process ports.

Definition 2 [7] Uplink flow consists of the communication process where the source address which produces the flow sends message to the destination address; downlink flow consists of the communication process where the destination address which receives the flow sends message to the source address.

Definition 3 Source-source (S-S) set [7] consists of the flows regarding the source address of the classified flow as the set's source address. Source-destination (S-D) set [7] consists of the flows regarding the source and destination address of the classified flow as the set's source and destination address. Source-port (S-P) set consists of the flows regarding the source address and source port of the classified flow as the set's source address and source port.

The two kinds of flow statistical features used in this paper are respectively single-flow features and multi-flow features.

1) Single-flow features are utilized to detect whether the flow is normal, including the duration, the TCP packet number, TCP packet number of different types, and so on, as shown in Table 2.

Table 2 Parameters of single-flow features for anomaly detection

flow feature	description of each flow feature
duration/s	session duration
TCP packet	TCP packet count of session
SYN packet	SYN packet count of session
SYN/ACK packet	SYN/ACK packet count of session
ACK packet	ACK packet count of session
FIN packet	FIN packet count of session
FIN/ACK packet	FIN/ACK packet count of session
RST packet	RST packet count of session
RST/ACK packet	RST/ACK packet count of session

2) Multi-flow features consist of the features defined in the S-S set, S-D set, and S-P set, as shown in Table 3.

Table 3 Parameters of multi-flow features for anomaly detection

flow feature	description of each flow feature
S-S set	count of flow, source port, destination IP and packet in the set
S-D set	count of flow and packet in the set
S-P set	count of flow and packet in the set

The accuracy of anomaly detection can be effectively improved by combining the single-flow features and multi-flow features, and multi-flow features can fully reflect the

differences on the behaviors of different network users. The experiment result shows that it can effectively reduce the rate of false alarm by using single-flow features and multi-flow features, and the extracted features have linear separability. Therefore, we can use the low-complexity linear Fisher classification algorithm to detect the database network anomaly traffic.

4 Fisher discriminant analysis

4.1 Basic principle of Fisher discriminant analysis

The statistical model of discriminant analysis can be abstractly and generally described as follows: there exists m P-dimensional population $(G_1,G_2,...,G_m)$, and its distribution can be described as $F_1,F_2,...,F_m$. A new sample that is described as $\boldsymbol{x} = (x_1,x_2,...,x_p)$ may be one of the samples of the population $G_1,G_2,...,G_m$. Fisher discriminant analysis [8,9] can be utilized to solve the problem of how to determine which population the sample comes from according to its p factors.

The basic idea of the Fisher discriminant analysis is projection that projects the P-dimensional variable to the D-dimensional space $(D < P)$ and classifies it in the D-dimensional space. The principle of projection is to make the differences of the same type as small as possible and the differences of different types as large as possible. Assuming that there are n samples respectively belonging to class 1 to class m $(m < n)$, and each sample has p variables to determine the classification, the expression of linear discriminant equation is as follows:

$$y = c_1x_1 + c_2x_2 + \cdots + c_px_p,$$

where $c_1,c_2,...,c_p$ are the undetermined coefficients of the discriminant function.

4.2 Sample statistics of P-dimensional space X

In this paper, the sample will be classified as normal or abnormal. Among the statistics of samples in the P-dimensional space X, ω_i is the type of one sample, and N_i is the number of samples in one class.

1) Mean vector $\boldsymbol{m}_i$:

$$\boldsymbol{m}_i = \frac{1}{N_i}\sum_{\boldsymbol{x}\in\omega_i}\boldsymbol{x},\ i = 1,2.$$

2) Within-class dispersion matrix $\boldsymbol{S}_i$ and all within-class dispersion matrix $\boldsymbol{S}_\omega$:

$$\boldsymbol{S}_i = \sum_{\boldsymbol{x}\in\omega_i}(\boldsymbol{x}-\boldsymbol{m}_i)(\boldsymbol{x}-\boldsymbol{m}_i)^{\mathrm{T}},\ i = 1,2, \tag{1}$$

$$\boldsymbol{S}_\omega = \boldsymbol{S}_1 + \boldsymbol{S}_2. \tag{2}$$

3) Inter-class dispersion matrix $\boldsymbol{S}_\mathrm{b}$:

$$\boldsymbol{S}_\mathrm{b} = (\boldsymbol{m}_1-\boldsymbol{m}_2)(\boldsymbol{m}_1-\boldsymbol{m}_2)^{\mathrm{T}}. \tag{3}$$

4.3 Method of discriminant analysis

y is a sample in the one-dimensional space Y, and the sample statistics in Y converted from the P-dimensional space X are as follows.

1) Within-class sample mean m_i:

$$m_i = \frac{1}{N_i}\sum_{\boldsymbol{x}\in\omega_i}y,\ i = 1,2. \tag{4}$$

2) Within-class dispersion S_i^2 and all within-class dispersion S_ω:

$$S_i^2 = \sum_{\boldsymbol{x}\in\omega_i}(y-m_i)^2,\ i = 1,2. \tag{5}$$

In order to facilitate the classification, the samples should be separated as much as possible when projecting from the P-dimensional space X to the one-dimensional space Y. That is to say, the inter-class mean (m_1-m_2) should be as large as possible, and the within-class dispersion should be as small as possible. Therefore, the Fisher criterion function [9] is defined as follows:

$$J(\boldsymbol{\omega}) = \frac{(m_1-m_2)^2}{S_1^2+S_2^2},$$

where $\boldsymbol{\omega}$ is the transformation matrix from the space X to space Y,

$$y = \boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x},$$

$$J(\boldsymbol{\omega}) = \frac{\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{S}_\mathrm{b}\boldsymbol{\omega}}{\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{S}_\omega\boldsymbol{\omega}}.$$

When $J(\boldsymbol{\omega})$ is the largest, we can get transformation matrix $\boldsymbol{\omega}^*$, then according to $\boldsymbol{\omega}^*$, the discriminant function can be obtained and can be used to determine the type of the new sample by discriminant analysis.

5 Discriminant analysis based on statistical package for social sciences (SPSS)

5.1 Discriminant analysis

For anomaly traffic detection, it is not only related to the features of single-flow, but also related to the relationship of multi-flow. In this paper, single-flow features and all features of the network were respectively used to build the model of anomaly detection, and the detection result was analyzed and compared in detail. The data collected on March 12th was utilized as the training data to build the

Fisher linear model, and the data collected on March 13th, March 15th and March 19th were utilized to estimate the accuracy and generalization ability of the model.

The result of anomaly traffic detection based on features of single-flow is shown in Table 4 which consists of the classification function coefficients (Fisher linear discriminant function coefficients). According to Table 4, various types of linear programming models are built as follows (q_1 and q_2 are respectively used to describe abnormal and normal discriminant function):

$$q_1 = 7.8\times10^{-6}x_1 - 0.032x_2 + 9.271x_3 + 0.358x_4 + 0.028x_5 - 0.579x_6 - 9.2, \quad (6)$$

$$q_2 = -3.9\times10^{-5}x_1 - 0.298x_2 + 3.6x_3 + 1.721x_4 + 1.561x_5 + 0.048x_6 - 23.213, \quad (7)$$

where q_1 and q_2 can be computed by utilizing the above model for anomaly flow to be detected. If $q_1 > q_2$, the flow to be detected is determined to be abnormal flow; If $q_1 < q_2$, the flow to be detected is determined to be normal flow. According to Table 4, Fisher linear discriminant function is only related with flow duration, TCP packet, SYN packet, RST packet, ACK packet, and RST/ACK packet, but not related with other features. In Table 4, the coefficient of SYN packet is the largest, and the coefficients of RST/ACK packet, ACK packet, and RST packet are larger. It means that the discriminant function is most sensitive with the number of SYN packet which can better reflect the result of anomaly detection to some extent.

Similarly, the coefficients of Fisher linear discriminant function based on single-flow and multi-flow features are shown in Table 5. The experimental result shows that the basis of discriminant function is more reliable when detecting anomaly based on 14 attributes selected from the total 17 attributes. Meanwhile, the number of destination IP in S-S set and number of flows in S-D set have the largest proportion among the coefficients of multi-flow features, and by contrast, the proportion of other features is smaller. According to the experimental result, although the coefficients of multi-flow features are smaller than the coefficient of single-flow features, multi-flow features can effectively reduce the false positive rate of anomaly detection.

5.2 Classification result

The classification result is shown in Table 6. When using single-flow features to classify the database network traffic collected on March 12th, the result shows that the accuracy rate of anomaly traffic detection is 98.4%, and the accuracy rate of normal traffic detection is 76.7%. However, when single-flow features and multi-flow features are both applied for anomaly detection, experimental result shows that the accuracy rate of anomaly traffic detection is 99.7%, and the accuracy rate of normal traffic detection is 100%. Therefore, it can effectively improve the accuracy of traffic detection and reduce the false positive rate and false negative rate by applying multi-flow features for anomaly traffic detection.

In this paper, based on the training data, all the features were utilized to build the Fisher linear model used for anomaly detection of testing data set. The detection result is shown in Table 7. For only abnormal packets existing on March 15th, false positive rate is represented as NO. The experimental result shows that the Fisher linear model keeps lower false negative rate and false positive rate at different times, and has good generalization ability for anomaly detection.

Table 4 Single-flow classification function coefficients

types of flow	duration/s	TCP packet	SYN packet	RST packet	ACK packet	RST/ACK packet	constant
abnormal	7.8×10^{-6}	−0.03	9.271	0.358	0.028	−0.579	−9.200
normal	-3.9×10^{-5}	−0.30	3.600	1.721	1.561	0.048	−23.213

Table 5 Single-flow and multi-flow classification function coefficients

types of flow	duration/s	TCP packet	SYN packet	RST packet	ACK packet	SYN/ACK packet	FIN/ACK packet	RST/ACK packet
abnormal	-9.90×10^{-6}	−0.020	7.385	1.748	0.115	−2.313	1.466	0.027
normal	0	0.218	−1.900	139.900	1.027	−137.300	41.841	−4.394
types of flow	number of flows in S-S set	number of ports in S-S set	number of destination IP in S-S set	number of packets in S-S set	number of flows in S-D set	number of packets in S-D set	constant	
abnormal	−0.003	−0.011	0.114	0.001	0.294	−0.010	−21.524	
normal	0.081	0.165	−0.272	−0.008	−17.597	0.873	−4388.550	

Table 6 Anomaly detection result of training sample

date	single-flow or all features	false negative rate/%	false positive rate/%
March 12th	single-flow	1.6	23.3
March 12th	all	0.3	0

Table 7 Anomaly detection result of testing sample

date	single-flow or all features	false negative rate/%	false positive rate/%
March 13th	all	9.10	0.60
March 15th	all	1.70	NO
March 19th	all	2.60	0

6 Conclusions

In this paper, a group of easy-extracted and linear separable features of related flows are proposed to detect the anomaly traffic of a database network. The remarkable features of the detection model are as follows:

1) The used features are easy to be extracted, and have low vector dimension, low complexity, and high-speed process.

2) The extracted features have linear separability, and we can utilize Fisher criteria to build a model to detect the data flow. The detection model is simple, and has low complexity and good generalization ability.

3) The algorithm can provide high accuracy for anomaly detection and classification of database network traffic.

4) The anomaly detection does not rely on the well-known port and packet keyword matching, and is well applicable to many private protocols and encryption protocols.

This model can effectively solve the problem of how anomaly traffic of database network is accurately detected, and has high availability and reliability. Furthermore, we will optimize the anomaly detection algorithm, reduce the algorithm complexity and make a deep research on how to apply the detection model in other network environments.

Acknowledgements This work was supported by the Key Project in the National Science and Technology Pillar Program (No. 2008BAH37B04), and the 111 Project (No. B08004).

References

1. Anley C. Advanced SQL injection in SQL server applications. Next Generation Security Software Ltd. White Paper, 2002
2. Anley C. More advanced SQL injection. Next Generation Security Software Ltd. White Paper, 2002
3. Ilgun K, Kemmerer R A, Porras P A. State transition analysis: a rule-based intrusion detection approach. IEEE Transactions on Software Engineering, 1995, 21(3): 181–199
4. Lindqvist U, Porras P A. Detecting computer and network misuse through the production-based expert system toolset (P-BEST). In: Proceedings of the 1999 IEEE Symposium on Security and Privacy. Oakland: IEEE Computer Society Press, 1999, 146–161
5. Roesch M. Snort-lightweight intrusion detection for networks. In: Proceedings of the 13th USENIX Conference on System Administration. Washington: USENIX, 1999, 229–238
6. Krügel C, Toth T, Kirda E. Service-specific anomaly detection for network intrusion detection. In: Proceedings of the 2002 ACM Symposium on Applied Computing. New York: ACM Scientific Press, 2002, 201–208
7. Lin P, Yu X Y, Liu F, Lei Z M. A network traffic classification algorithm based on flow statistical characteristics. Journal of Beijing University of Posts and Telecommunications, 2008, 31(2): 15–19 (in Chinese)
8. Johnson R A, Wichern D W. Applied Multivariate Statistical Analysis. New Jersey: Prentice Hall, 2002
9. Duda R O, Hart P E, Stork D G . Pattern Classification. 2nd ed. New York: Wiley-Interscience, 2001