A Network Protocol Cluster Analysis Method Based on Basis Functions

Wei Wang[1, a *] and Wenhong Zhao[2,b]

[1] Science and Technology on Communication Information Security Control Laboratory, Jiaxing Zhejiang 314001, China

[2] Nanhu College, Jiaxing College, Jiaxing 314001, China

[a]wwzwh@163.com, [b]whzhaonh@163.com

Keywords: Protocol Analysis; Base-function; Protocol State; K-Means

Abstract. Network protocol analysis can help security personnel analyze network vulnerabilities. However, the network protocol analysis is confronted with more and more kinds of protocols and more and more complex issues. Firstly, based on the basis function of the network protocol structure representation method, the improved k-means clustering algorithm based on the basis function is presented. Finally, a network protocol analysis process based on basis function is presented. Performance analysis shows that the proposed method is superior to traditional matching methods and statistical methods.

Introduction

Network protocol analysis is the core issue that many scholars focus on in the field of network management and network security [1-5]. Network protocol analysis by capturing the packets in the network, packet first and data fields provided detailed information and statistical results of protocol, and then the data classification and analysis, thus further help find network potential safety hazard, and in the event of a failure to provide network fault analysis information [6-9]. Network protocol analysis can make the network management personnel can quickly and accurately locate the cause of the problem, find out the cause of failure of network nodes, network protocols and network link, with the fastest speed to restore the normal operation of the network [10-11]. In addition, network protocol analysis can provide reliable basis for planning and adjusting network by analyzing network communication situation and network connection status, and reasonable allocation of network performance and resources.

However, due to the diversity and privatization of existing network protocols, the agreement analysts are faced with more and more kinds of protocols and more and more complex issues of agreement state space.

Many of the existing protocol analysis methods have mostly used string matching methods [2,5,7], and since this approach USES a large number of matching methods, the speed is slower. However, other methods of protocol analysis based on statistics have defects of low accuracy [3,4,6]. In addition, these methods cannot be used to analyze the private protocols of the network companies. In the case of continuous bittorrent, the data of Song [12] has proposed a method of data frame demarcation based on fingerprint characteristics by improving AC algorithm. Jin [13] et al., mining by frequent series and association rules, then calculate the minimum position difference of frequent string, and propose a frame head recognition technology based on association rules. Wang [14] in the capture of data for the data frame (e.g., the PPP frame, Ethernet frames, etc.) under the condition of a "identify specific circumstances unknown protocol based on association rules method", the method by mining association rules to recognize and identify the unknown protocol. These solutions can be set in the corresponding conditions get better effect, provide a useful reference for identifying unknown agreement, but they are all in a single agreement assumes that the analysis and experiment, in the practical application environment, to capture the unknown protocol data frame is often a variety of mixed

In order to solve the above problem, this paper designs a new protocol analysis process, and then are involved in the process of various methods are discussed, including: basis function based protocol structure characterization, protocol recognition based on clustering analysis, a self learning agreement

structure identification. Based on the protocol structure representation of the base function, the problem of protocol representation is mainly solved, and many protocols are used to reduce the complexity of the protocol description by using a limited number of basis functions. Based on clustering analysis, this paper analyzes the known structure protocols of nested structures and reduces the complexity of recognition analysis. Can self-learning protocol structure identification basis function building for private protocols, so as to take advantage of new basis function for characterizing, protocol access to its field structure, and add the new basis function library of protocol analysis, the final will be private protocol is converted to a known structure are analyzed.

The protocol structure representation method based on basis function

Because of the variety of feature fields of existing protocols, a small amount of information is required to express the characteristics of a large number of protocols, providing a basis for the analysis of the protocol.

The Walsh function is used as the basis function to characterize the protocol structure because of the characteristics of the function of Walsh function.

Walsh's function is defined as follows:

If use wal(k,t)(k=0,1,…) to represent the Walsh function of interval $t \in [0,1)$, it is defined as the following formula:

wal(2k,t) = wal(k,2t)+(-1)^{k}wal(k,2t-1), k=1,2,…

wal(2k+1,t) = wal(k,2t)+(-1)$^{k+1}$wal(k,2t-1), k=0,1,…

$$\text{wal}(0,t) = \begin{cases} 1, & 0 \le t < 1 \\ 0, & t < 0, t \ge 1 \end{cases}$$

After the Walsh function, we define the following transformation to convert ±1 of the Walsh function to 0,1 bit stream:

$$f(w) = \begin{cases} 1, & w = -1 \\ 0, & w = 1 \end{cases}$$

The transformation f(x) is used to obtain a set of orthogonal basis functions (k,t)(k=0, 1,...). :

base(2k,t) = f (wal(2k,t)), k=1,2,…

base(2k+1,t) = f (wal(2k+1,t)), k=0,1,…

base(0,t) = f (wal(0,t))

Due to the known structure of protocol of data can be represented as 0, 1 code strings, so you can use the different combination of orthogonal basis functions base on the model describe all known protocol, which can establish a basis function library and model combination of known structure agreement basis function library.

Below, we in the base function and transformation of $f(x)$, on the basis of using the existing basis function mode combination and received target protocol information, and the time of using sliding way to describe the target agreement - structure map, and then according to the time - structure map depicting basis function mode combination - match rate distribution, to determine whether the input data is the known structure of protocol data.

Suppose C is a set of combinations of basis functions,

For each of the existing base function combinations, $c_1, c_2, \ldots, c_{cn} \in C$. The time sliding mode is used to make the difference or operation with the input data, and the time of the input data is obtained.

To combine the different base function combinations of c_i, add the time of the input data - the structural distribution data, and get the structural matching value of each basis function composition pattern (c_i, m_i), $i \in \{1,2,\ldots,c_n\}$;

Using the structure matching of all known base function combinations to draw the combination of the base function pattern and the matching rate distribution;

Comparison of the maximum matching rate of m and some threshold t

If *m* is greater than or equal to *t*, the input data is considered to be the data of the known structure agreement, and the output of the module output to the hierarchical protocol analysis module is stopped.

If *m* is less than *t*, then the input data is considered as the data of the private agreement, and the output result of the self-learning protocol analysis module will be stopped

Among them, threshold *t* can be artificially specified, which affects the accuracy of protocol analysis.

Improved K-Means Clustering Algorithm

By the above method to calculate the approximate value of K and after the initial clustering center, K-Means algorithm, performance in a K value from zero knowledge environment, secondly, the algorithm does not need to random initial clustering center, K specified directly K clustering initial clustering center, so we can speed up the convergence rate of K - Means algorithm. The following steps are:

Improved K-Means for bit-stream protocol
Input: *n* bar data frames; *K*; *K* initial center Output: *K* clusters Steps: 1. For *i* =1 to *n*; 2. Calculate the distance d_i of the data frame x_i to each cluster center, and divide the data frame x_i into the nearest cluster; 3. Calculate the error sum of squares and E by formula $E = \sum_{j=1}^{K} \sum_{x \in C_j} \| x - m_j \|^2$; 4. Recompute the cluster center and calculate the error sum of squares and E* by formula (1). 5. Compare the absolute value of the difference between E and E *. If the threshold is less than the threshold, it goes to step 6, otherwise it goes to step 1; 6. Output K clusters.

After clustering the unknown protocol using the improved K-means algorithm, it is difficult to evaluate the unknown protocol without prior knowledge. In this paper, a result evaluation scheme using information entropy is presented. The calculation procedure of the algorithm is as follows:

CEAE(Cluster evaluation algorithm by entropy)
Input: class cluster containing n data frames Output: information entropy of the columns of this class Steps: 1. According to the data preprocessing method, two dimensional matrices are obtained and the corresponding two-dimensional array a[n][m] is established. 2. Loop through the groups 3. According to the column, count the number of occurrences of each byte in each column; 4. Loop through each column 5. Calculate the information entropy value of each column according to formula $E(X) = -\sum_{x \in S(x)} p(x) \ln(p(x))$; 6. Output information entropy per column.

The results are shown as the X-axis, the entropy of the column is the Y-axis, and the results of the clustering are analyzed. The entropy value represents the size of the information confounding degree, and in the case of large data frames, if it is the data frame of the same protocol, then the entropy value of some columns is close to 0. If there are multiple protocols mixed, the entropy value is nearly zero. Therefore, the method of calculating entropy value can be used to evaluate the good or bad of the unknown protocol clustering. The evaluation threshold low_entropy = 0.05, the more column entropy is smaller than low_entropy, the better the clustering effect will be.

Protocol analysis process

This section of protocol analysis process based on basis function first characterization of ideas for the analysis of the network protocol data analysis, determine the structure is known protocol data or private data. Then according to the characteristics of the known structure of protocol and private for processing respectively, deal with known structure, due to the current network of the protocol hierarchy nested design train of thought, so we used a layered protocol analysis method; For private protocols, we use the thought of self-learning to construct new base functions and characterize the protocol data, and then transform it into a protocol of known structures for analysis.

The specific process is described as follows:

Enter the protocol data of the target network;

Characterizing the structure of the target protocol based on the existing base functions and the combination modes of the existing libraries;

Determine whether the structure distribution map of the target protocol has a matching basis function combination mode

If there is, the input is the protocol data of the known structure, and then (4)

If no, input is private agreement data, and (5)

The protocol data is analyzed and stopped by the layered method.

Adopt the self-learning basis function and the combination mode extension method to establish the new base function, base function combination mode, update the base function library and the combinatorial mode library, and (2).

Experimental Results and Analysis

To test the effect of the model, the tcpdump dataset released by Lincoln lab was used to test the data. The test data was divided into four groups:

Group 1:5 protocols: DNS, HTTP, NTP, SMTP, SSH;

Group 2:5 protocols: FTP, icmp_error, irc, NBSS, Telnet;

Group 3:9 protocols: arp, DNS, HTTP, LLC, loop, NTP, rip, SMTP, SSH;

Group 4:9 protocols: FTP, icmp_data, icmp_error, irc, NBSS, NBSS, pop, syslog, Telnet;

Parameter lowestSimilar (L): Set value threshold to determine whether two sets are merged. If the similarity of the two sets is greater than or equal to lowestSimilar, the two sets are merged or not merged. The greater the threshold value, the greater the K value, and the smaller the K value.

For the first set of data, set the parameter lowestSimilar(L) to change from 0.1 to 1.0, with each increment of 0.05, the corresponding K value is calculated.

As shown in Fig. 1, the pointcut coordinates (0.65, 7). To specify the initial center of K=7 , there are 500 pieces of data, of which 429 are correct clustering and the correct rate is: C = 429/500 * 100% = 85.5%.

As shown in Fig. 2, for the second group of data, using the same method, the tangent point is (0.65, 6), and the calculated value of K is 6, and the accuracy of using k-means clustering is: C = (464/500)*100%=92.8%.

As shown in Fig. 3, for the third set of data, the tangent point coordinates (0.68, 8), calculated K= 8, use k-means algorithm to cluster, the correct rate is: C = (652/900)*100%=72.4%.

As shown in Fig. 4, , for the fourth group data, the pointcut coordinates (0.678, 10), calculated K= 10, use k-means algorithm to cluster, the correct rate is: C = (760/900)*100%=84.4%.

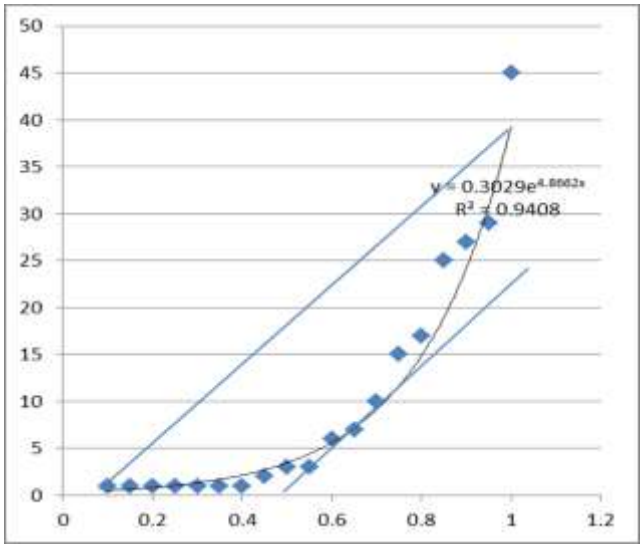

Figure 1. L-k curve (group 1)

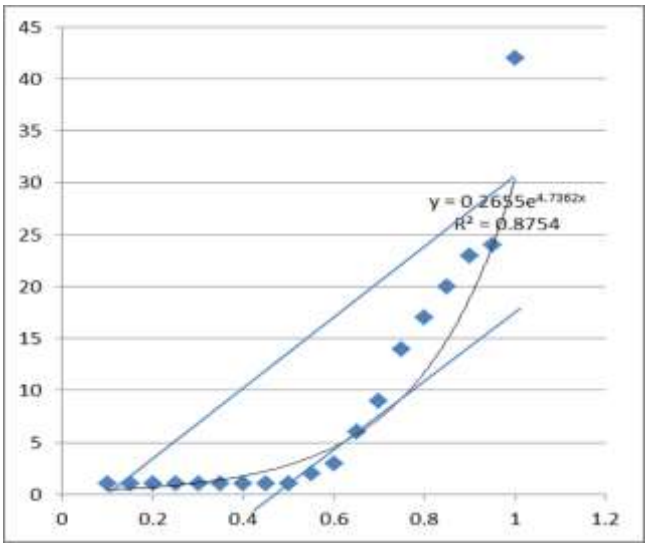

Figure 2. L-k curve (group 2)

As can be seen from the above figures, the calculated value of K is close to the real value in the method proposed in this paper, and the clustering analysis can also get higher clustering accuracy with the calculated K values.

Fig. 5 shows the result of using the matching method [2], statistical method [3] and the proposed method to analyze the protocol data, where the horizontal coordinates are the number of protocol data entered, and the ordinate is the analysis time. The experimental data is derived from the Intranet data of an enterprise, which contains encrypted network data and unencrypted data. It can be seen from the figure that the proposed method of network protocol based on basis function is better than the matching method and statistical method.

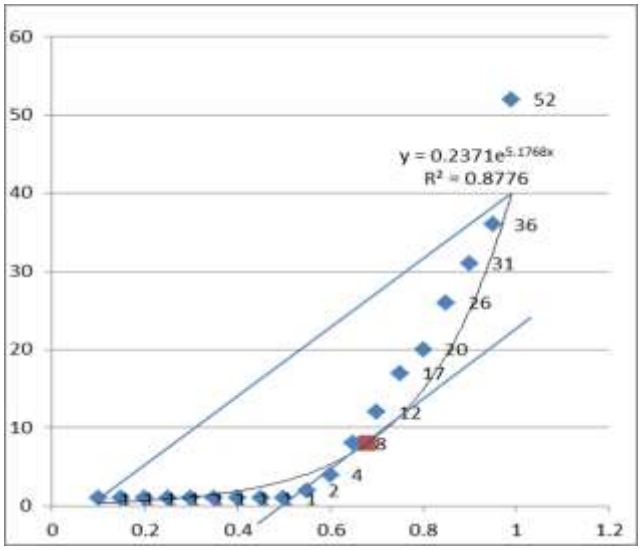

Figure 3. L-k curve (group 3)

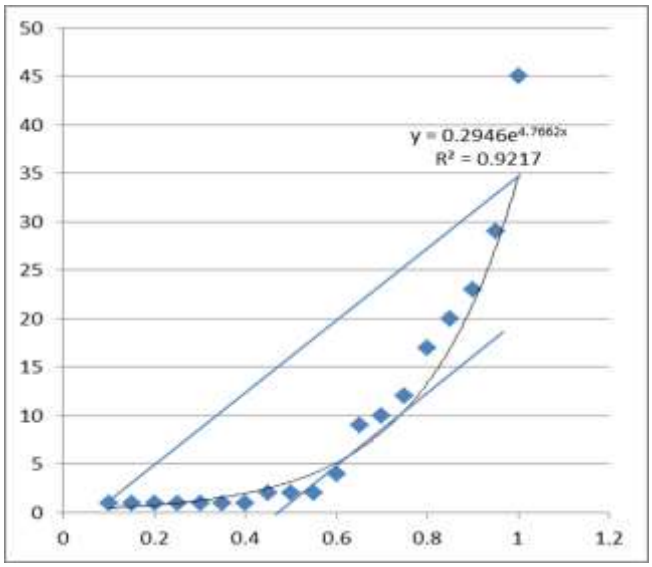

Figure 4. L-k curve (group 4)

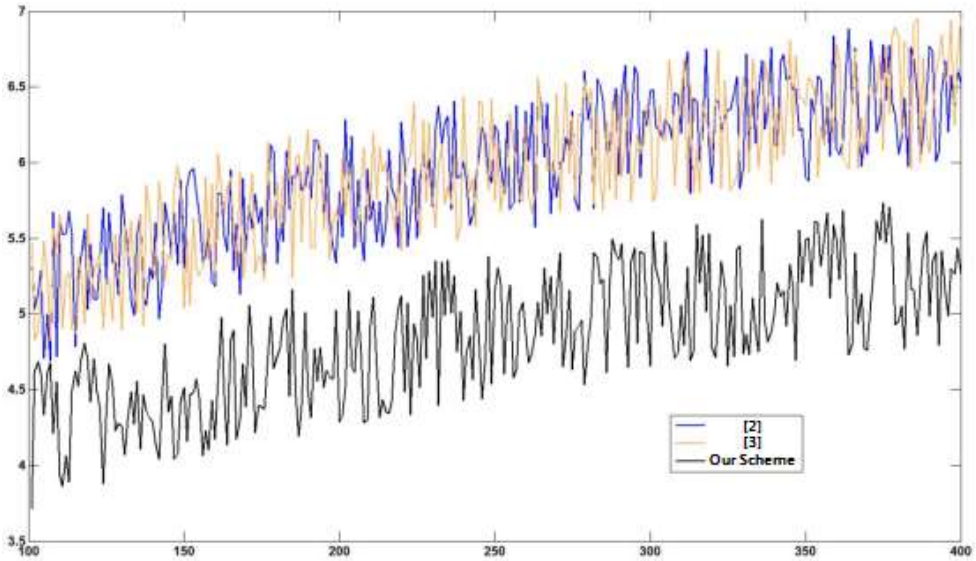

Figure 5. Compare the results of the protocol analysis

Conclusion

In order to solve the problem of network in the field of security protocol analysis, this paper will deal the data into the description of the basis function and the mode of combination way, and then puts forward the basis function based protocol structure characterization methods, based on the pattern of basis function combination layer protocol analysis method and self-learning basis functions and their combination patterns of extension methods, finally was given by using the method of protocol analysis process. The performance analysis experiment shows that the proposed method can be used to analyze the network protocol quickly and accurately.

References

[1] Wireshark user manual, http://man.1upaworld.com/content/network/wireshark/
[2] L. Xu. Design and implementation of network protocol analysis system. Computer programming skills and maintenance. 2009, 8:74-76.
[3] L. Chen, J. Gong, C. Xu. An overview of algorithm protocol recognition algorithms. Computer science. 2007, 7, 3(7):73-75.
[4] S.H. Chen, J.S. Su. Research on protocol recognition based on content analysis. Journal of national defense technology university. 2008, 30(4):82-87.
[5] H.P. Fan, L. Xu, S.H. Chen. The application layer protocol recognition acceleration based on regular expressions. Computer research and development. 2008, 45:438-443.
[6] Q. Li, Q. Zhao,K.Q. Xiang et al. Research on the network protocol recognition algorithm based on decision tree [J]. Microcomputer information. 2009, 25(9-3): 25-26.
[7] M.H. Shi, W.W. Yan, Y.G. Li. Analytic function of CORBA communication protocol based on Ethereal. Computer engineering and design. 2005, 26(12):3236-3239.
[8] S. Owen, P. Brereton, D. Budgen. The Protocol analysis: A neglected practice. Journal of Communications of the ACM, 2006, 49 (2) : 117-122
[9] T. Abbes, A. Bouhoula, Michael Rusinowiteh. Protocol Analysis in the International Conference on Information Technology: Coding and Computing. USA, 2004, 1253-1261.
[10] J.C. Wang, F. Li, M.Y. Shen. Protocol analysis algorithm design and implementation of network data packets. Computer technology and development. 2006. 16(4): 77-80
[11] F. Risso, L. Degioanni. An architecture for high performance network analysis. Proceedings of the Sixth IEEE Symposium on Computers and Communications.
[12] S. Jiang. Discovery of unknown protocols in wireless network environment. Chengdu: university of electronic science and technology, 2013.
[13] L. Jin. Research on unknown frame-head recognition for bittorrent. Shanghai: Shanghai jiaotong university, 2011.
[14] Y. Wang, Y.M. Wu, F. Li, et al. Correlation analysis and identification of unknown protocols for bitstream data [J/OL]. Computer application research, 2015,32(1):243-248.