Applying Multiple Time Series Data Mining to Large-Scale Network Traffic Analysis

Weisong He,Guangmin Hu, Xingmiao Yao ,Guangyuan Kan and Hong Wang
School of Communication and Information Engineering
University of Electronic Science and Technology of China
Chengdu, P.R. China
{weisonghe,hgm}@uestc.edu.cn

Hongmei Xiang
Chongqing College of Electronic Engineering
Chongqing, P.R. China

***Abstract*—Minimize false positive and false negative is one of the difficult problems of network traffic analysis. This paper propose a large-scale communications network traffic feature analysis method using multiple time series data mining, analyze multiple traffic feature time series as a whole, produce valid association rules of abnormal network traffic feature, characterize the entire communication network security situation accurately. Experiment with Abilene network data verify this method.**

***Keywords*—network traffic analysis, principal components analysis, time frequency analysis, symbolic time series analysis, multiple time series data mining.**

I. INTRODUCTION

Network traffic anomalies has the feature of erupting suddenly without known signs, which can bring great damage to network or computers in network in a short time. Therefore, one of the prepositions to ensure a safe network is to detect network traffic anomalies fast and accurately, determine the reasons that causes them and make reasonable response to them in time. To decrease abnormal network traffic and reduce or eradicate attack of denial of service, large-scale communication network routing and switching equipment must possess abilities of detecting and analyzing network traffic behavior. The features of large-scale communication network traffic are high speed and immense data while anomalous traffic is small and scattered in multiple links, which is hard to be detected among normal traffic. Moreover, parameters for analysis are limited. All these make anomaly identification very difficult.

Large-scale communication network traffic anomaly detection is mainly on three levels: packet level, flow level and network-wide level. The advantage of packet level traffic analysis is to provide elaborate information about user performance on the finest level of granularity and basic information about application layer, which in favor of description of anomalous features and fault diagnosis. For example, Snort[2], a signature-based intrusion detection system, summarize packet content in which special attack will appear as one attack feature in artificial ways, and this special attack can be determined if a packet has the same feature when intrusion detection system matches packet content. Ke Wang and Salvatore J.Stolfo [3] adopt the way of using statistical distribution of bytes ASCII code of packet to distinguish the content difference between normal packet and abnormal packet, then the normal connects and abnormal events. Masaki Ishiguro [4] distinguishes network worms attack or port scan through observing packet frequency to specific IP address by Bayesian classification method. But, because of the features of wide distribution, high speed and massive data possessed by large-scale communication network, capturing packet is hard to be implemented on large-scale network. Flow level traffic analysis is based on flow classification, collecting statistical information of each flow and providing performance information of users on medium granularity, which makes characterizing, detecting, diagnosing and restoring network convenient. The idea of flow level traffic analysis is to separate events, group abnormal types, search distributive model of anomaly and analyze anomaly pattern. Since Netflow is a good compromise for traffic analysis based on SNMP and packet in its performance and accuracy, the mainstream method of flow level traffic analysis is to be based on Netflow. Network-wide traffic analysis uses the global traffic information, including path traffic, link traffic information, etc. for example, [5] deploys method of multi-way subspace to identify links with traffic anomaly.

Early network traffic analysis mainly focused on the laws that single traffic feature (such as value of traffic, counts of bytes) changes. But single traffic feature time series can not characterize large-scale communication network traffic completely and accurately, so problems like high false positives and false negatives can not be avoided. Lakhina [5] adopted method of multi-way subspace to detect and identify links with traffic feature anomaly. However, this study only used clustering analysis to obtain types of anomaly and did not study correlations among multiple traffic features.

Our contributions lie in: (1) aiming at traffic features of large-scale communication network, make every traffic feature simple time series; then take multiple traffic feature as a whole to analyze and study through multiple time series data mining. (2) search motif correlation pattern among anomalous segments of multiple time series within the same time interval by Multiple Time Series Data Mining(MTSDM for short in the following), analyze correlation patterns of multiple traffic feature anomaly and describe network security situation of large-scale network accurately and qualitatively.

978-1-4244-1674-5/08 /$25.00 ©2008 IEEE

The rest of the paper is organized as the following. In Section II, we illustrate the process of network traffic analysis and three-level network traffic analysis frame. In section III we illustrate data preprocessing. In section IV we introduce multiple time series data mining method. In section V, we provide experiment and in section VI a conclusion.

II. The Process of Large-scale Communication Network Traffic Analysis

A. *Overview*

The process of large-scale communication network traffic analysis is shown in Fig. 1.

Basically, in this paper, the process of large-scale communication network traffic analysis consists of the following five steps:

- Compute entropy of several flow level traffic features collected over each time bin.
- Apply Principal Component Analysis and subspace method to entropy time series.
- Apply time frequency analysis method and Piecewise Aggregate Approximation and Symbolic Aggregate approximation to anomaly time series.
- Apply association rule mining to symbolic sequence.
- Real-time monitoring with valid motif association pattern.

The first 1,2,3 step is data preprocessing stage, step 4 is data mining stage, step 5 is the outcome of the mining and network traffic monitoring with the valid association rules.

B. *Level of Network Traffic Analysis*

The network traffic analysis consists of three levels, from bottom to top are packet level, flow level, network-wide level.

1) Packet Level Traffic Analysis: Each packet including time stamp,IP address or prefix, port number, protocol type, bytes and content. IP header information includes traffic volume by IP addresses or protocol, burst of the stream of packets, packet properties (e.g., sizes, out-of-order). TCP header information includes traffic breakdown by application (e.g., Web), TCP congestion and flow control, number of bytes and packets per session. Application header information includes URLs, HTTP headers (e.g., cacheable response?), DNS queries and responses, user key strokes and so on.

2) Flow Level Traffic Analysis: Basic information about the flow include source and destination IP address, port number, packet and byte counts, start and end times, ToS, TCP flags. Information related to routing includes next-hop IP address, source and destination AS.

3) Network-wide Level Traffic Analysis: Network-wide level traffic analysis combines traffic, topology, and state information. Network-wide level traffic analysis is mainly referred to traffic matrix analysis in this paper. Traffic Matrices (TM) reflect the traffic volume of OD (origin-destination) flow in a large-scale communication network. Information on the size and locality of flows contained in traffic matrix is crucial for monitoring network and diagnosing problems.

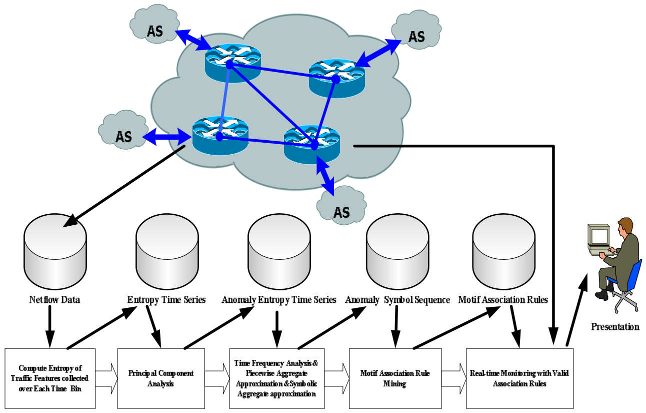

AS
AS
AS
AS
Netflow Data
Entropy Time Series
Anomaly Entropy Time Series
Anomaly Symbol Sequence
Motif Association Rules
Presentation
Compute Entropy of Traffic Features collected over Each Time Bin
Principal Component Analysis
Time Frequency Analysis & Piecewise Aggregate Approximation &Symbolic Aggregate approximation
Motif Association Rule Mining
Real-time Monitoring with Valid Association Rules

Figure 1. The process of large-scale communication network traffic analysis

TABLE I. Description of 6 Feature Time Series

Series	Description
H(srcPort)	Entropy of source port distribution
H(dstPort)	Entropy of destination port distribution
H(srcIP)	Entropy of source IP address distribution
H(dstIP)	Entropy of destination IP address distribution
H(octets)	Entropy of octets distribution
H(prot)	Entropy of protocol type distribution.

III. Data Preprocessing

A. *Information Entropy*

The data packets within each bin of five minutes are summarized by a set of aggregate features. A flow is formally defined as a 5-tuple: source address, destination address, source port, destination port, and type of protocol. We focus on six fields: source address (sometimes called source IP and denoted srcIP), destination address (or destination IP, denoted dstIP), source port (srcPort), destination port (dstPort), octets and type of protocol.

Entropy is a metric that captures the degree of dispersal or concentration of a distribution. A wide variety of anomalies will impact the distribution of one of the discussed IP features as in:

$$H(X) = -\sum_{i=1}^{n} P[X = x_i]\log(P[X = x_i]) \qquad (1)$$

$P[X = x_i]$ is the probability of event $x_i \in X$ occurring. For example, the probability of seeing IP 129.173.192.0 is defined to be number of packets using IP 129.173.192.0 divided by the total number of packets in the given time interval.

A set of 6 aggregate features used in this paper are listed as Table I.

B. *Principal Components Analysis and Subspace Method*

The subspace method is an effective approach to separate normal from anomalous network traffic. Principal components analysis (PCA) [6] is usually the main method applied to exert such function.

We regard the data points as a cloud with n dimensions, and the first main component PC_1 is the direction point with the greatest change. PC_2 is the direction points with greatest change which stand on the orthogonal direction of PC_1. This process does not finish until all the main components are discovered, and n refers to the dimensions of data. Please make sure all the main components converge with one another orthogonally, and thus form an orthonormal basis. The data change most on the first number axis direction, while the one on the second axis change less than the first one, and so do the other components.

The subspace method uses these PCA mentioned above to define the normal subspace and the anomalous subspace. For some m, the normal subspace is the space spanned by PC_1 through PC_m, and the abnormal subspace is similarly the space spanned by PC_{m+1} through $PC_n (m \le n)$.

C. *Time Series Representation and Time Frequency Analysis*

1) Piecewise Aggregate Approximation: The basic idea of Piecewise Aggregate Approximation (PAA) [7][8][9] is that it represents the time series as a sequence of rectangle basis functions. It is a dimensionality-reduction representation method in essential as in

$$\overline{p_i} = \frac{N}{n} \sum_{j=\frac{n}{N}(i-1)+1}^{\frac{n}{N}i} p_j \qquad (1)$$

n is the length of sequence, N is the number of PAA segments. $\overline{p_i}$ is the average value of the i^{th} segment.

2) Symbolic Aggregate approximation: The basic idea of Symbolic Aggregate approximation (SAX)[7][8][9]is that it converts the time series into an discrete symbolic sequence. Having transformed a time series data into the PAA, we can apply SAX to obtain a discrete symbolic representation. Since normalized time series have a Gaussian distribution, we can determine the "breakpoints" that will produce c equal-sized areas under Gaussian distribution curve [10].

The breakpoints will be found out by looking them up in Table II. Once the breakpoints have been obtained we can disperse time series into discrete symbolic series. We firstly obtain a PAA manner of the origin time series. All PAA coefficients that are below the smallest breakpoint are transformed to the symbol "a", and all coefficients greater than or equal to the smallest breakpoint but less than the second smallest breakpoint are transformed to the symbol "b", etc. Fig.2 illustrates the idea.

TABLE II. A LOOKUP TABLE THAT CONTAINS THE BREAKPOINTS

c \ β	3	4	5	6	7	8	9	10
β_1	-0.43	-0.67	-0.84	-0.97	-1.07	-1.15	-1.22	-1.28
β_2	0.43	0	-0.25	-0.43	-0.57	-0.67	-0.76	-0.84
β_3		0.67	0.25	0	-0.18	-0.32	-0.43	-0.52
β_4			0.84	0.43	0.18	0	-0.14	-0.25
β_5				0.97	0.57	0.32	0.14	0
β_6					1.07	0.67	0.43	0.25
β_7						1.15	0.76	0.52
β_8							1.22	0.84
β_9								1.28

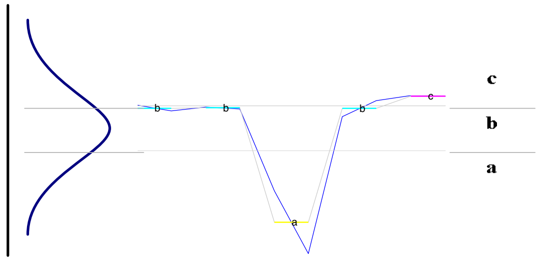


Figure 2. A time series is dispersed by firstly obtaining a PAA approximation and then using predetermined breakpoints to map the PAA coefficients into SAX symbols.

Note that in this example the 3 symbols, "a", "b" and "c" are approximately equiprobable as we desired. We call the series of symbols representing a subsequence a word.

A subsequence P with length N can be represented as a word $\overline{P} = p_1, \ldots p_n$ .Let α_i denote the i^{th} element of the alphabetize, $\alpha_1 = a$ and $\alpha_2 = b$ .Then the transformation from a PAA approximation P_{appr} to a word $\overline{P}$ is obtained as in

$$\overline{p_i} = \alpha_j , iif \ \beta_{j-1} \le p_{appr} \le \beta_j \qquad (3)$$

The PAA representation is only an intermediate step of obtaining the SAX.

3) Wavelets Transform and Wavelet Packet Transform: Suppose $x(t) \in L^2(R), \psi(t)$ is the basic wavelet function, and $\psi_{a\tau}(t) = \frac{1}{\sqrt{a}} \psi(\frac{t-\tau}{a})$ is the shift and scale extension of the basic wavelet function as in

$$WT_x(a,\tau)=\frac{1}{\sqrt{a}}\int x(t)\psi^*(\frac{t-\tau}{a})dt=<x(t),\psi_{a\tau}(t)> \qquad (4)$$

is called as the wavelet transform of $x(t)$. The basic idea of discrete wavelet transform (DWT) is to transform a discrete time signal into a discrete wavelet representation. Discrete wavelet transform converts an input series $x_0, x_1,..., x_k$ into one high-pass wavelet coefficient series and one low-pass wavelet coefficient series (of length $n/2$ each) given by as in (5) and (6):

$$H_i=\sum_{k=0}^{m-1}x_{2i-k}\bullet s_k(z) \qquad (5)$$

$$L_i=\sum_{k=0}^{m-1}x_{2i-k}\bullet t_k(z) \qquad (6)$$

Where $s_k(z)$ and $t_k(z)$ are wavelet filters, m is the length of the filter, and $i=0,1,...[n/2]-1$. In practice, such transformation will be used recursively on the low-pass series until the desired number of iterations is reached.

Wavelet Packet Transform (WPT) is a method which makes time frequency decomposition of signal. WPT is of self-adaptive of signal, which can effectively display the time frequency property of signal. Just by orthogonal mirror filter we can obtain WPT decomposition. Assume the signal $y(t)$, we can obtain

$$\begin{cases} y_{2n}(t)=\sqrt{2}\sum_k h(k)y_n(2t-k) \\ y_{2n+1}(t)=\sqrt{2}\sum_k g(k)y_n(2t-k) \end{cases} \qquad (7)$$

Function set $\{y_n(t)\}$ is called as wavelet packet which is the result of whole decomposition of all bands on various scale of origin signal. Let $k=n-2^j$, then $y_n(t)=y_{2^j+k}(t)$ is the result of decomposition of k -band on scale j. WPT may consist of various orthogonal basis, wavelet basis is the typical case. Among all combination, the least entropy is the good basis. The decomposition of good basis can represent the time-frequency of signal which imply that the method is adaptive for signal.

4) Choi-Williams Distribution: In order to reduce the disturbed components of the Chio-William distribution[11], we should have a research on the factors which make the disturbed value minimum. The Choi-Williams is proposed to solve this problem as in

$$C(t,f)=\int_{-\infty}^{+\infty}e^{j2\pi ft}\int_{-\infty}^{+\infty}\sqrt{\sigma/4\pi\tau^2}e^{-\frac{\sigma(\mu-t)^2}{4\tau^2}}x(\mu+\frac{\tau}{2})x^*(\mu-\frac{\tau}{2})d\mu d\tau \qquad (8)$$

Wigner-Ville distribution $C(t,f)$ satisfies the edge conditions and shift characteristics but does not satisfy weak and limited support characteristics. However,when $\sigma\rightarrow\infty$ it would satisfy weak and limited support characteristics.

5) Pseudo Wigner-Ville Distribution: For a given time, Wigner-Ville Distribution can describe the global distribution of a signal. In addition, for a given frequency, it can also equally measure all the frequencies either higher or lower than the given frequency. In fact, we are not able to study all the integrals between $-\infty$ and $+\infty$ but the ones in a limited range. When studying the distribution shape of a certain time (t) we should study the features of signals near the given time. In a sense, it means to condense the cross-item of multivariable signal by adding some windows and deleting the non local components. Finally we change Wigner-Ville distribution into local distribution. Pseudo Wigner-Ville distribution characterizes [11]a local behavior of a signal as the following formula, so it is convenient for us to mine the local features of the fault signals as in

$$PW_x(t,f)=\int_{-\infty}^{+\infty}h(\tau)x(t+\frac{\tau}{2})x^*(t-\frac{\tau}{2})e^{-j2\pi f\tau}d\tau \qquad (9)$$

$h(t)$ is the window function.

IV. Multiple Time Series Association Rules Data Mining

$I=\{i_1,i_2...,i_k\}$ is a set of items. X is an *itemset* if it is a subset of I. $T=\{t_i,t_{i+1},...,t_n\}$ is a set of transaction. A transaction t contains *itemset* X iff, for all items, where $i\in X, i$ is a $t-$ *itemset.* All *itemset* X in a transaction database T has a support, denote as $Sup\ (X)$ [12] [13], see as in

$$Sup\ (X)=\frac{\sigma(X)}{|T|} \qquad (10)$$

$\sigma(X)=|\{t_i|X\subseteq t_i,t_i\in T\}|$, Given an *itemset* X , we find all the rules $X\rightarrow Y$ with minimum support and confidence. Support is a probability that a transaction contains $X\cup Y$, denoted as $Sup(X\cup Y)$ .Confidence is a conditional probability that a transaction contains X as well as Y, denoted as $Sup(X\cup Y)/Sup(X)$ .Frequent *itemsets* is used to generate all frequent *itemsets* in a given database T. Dislike general association rules mining, the time of data point should be focused on when we using time series association rules mining. Each feature elements have different value at each time, so we apply association rule mining to different values of different network feature within the same time interval.

The abrupt changes of different time series at the same time bin usually have a certain motif association pattern. This rule lays the foundation for us to analyze anomalous behavior of large-scale network and predict the network security situation. The demo of Multiple Time Series Association Rules Data Mining is showed in the following Fig.3.

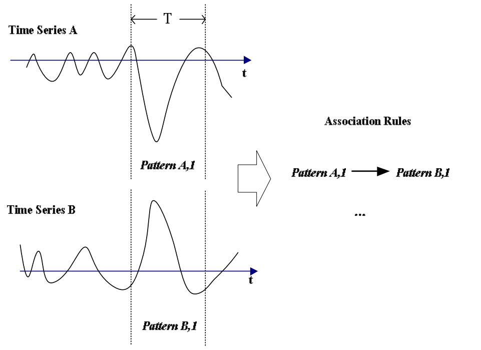
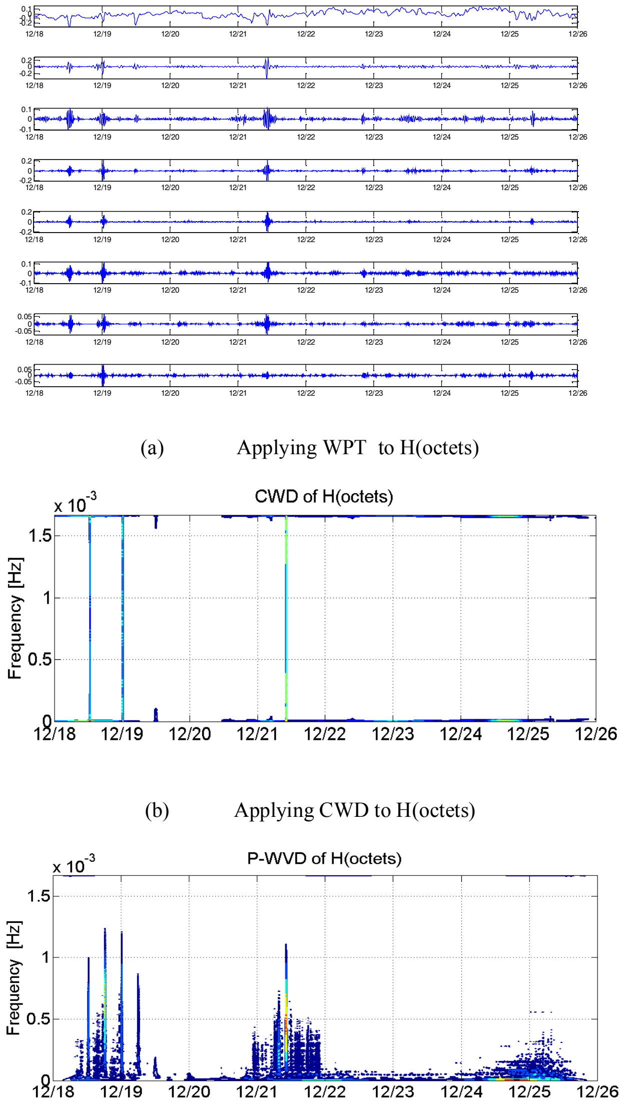


Figure 3. Multiple time series association rules data mining at the same time interval T



(a) Applying WPT to H(octets)

(b) Applying CWD to H(octets)

(c) Applying PWVD to H(octets)

Figure 4. Apply WT,CWD and PWVD methods to anomaly entropy time series

By such method, we can obtain different motif pattern about various anomalous behaviors of large-scale communication network, which can be applied to network traffic analysis.

V. Experiment Result

A. Data Sets

The sampled flow data collected from the backbone networks: Abilene [1]. Abilene is the Internet2 backbone network, connecting over 200 US universities and peering with research networks in Europe and Asia. It consists of 11 Points of Presence (PoPs), spanning the continental US. Sampling is periodic, at a rate of 1 out of 100 packets.

B. Simulation Test

We collect data from 08:05 a.m. on December 18th, 2006 to 08:00 a.m. on December 26th, 2006.We firstly compute entropy value of each feature within each time bin. We analyze six time series of entropy value by PCA method, and obtain the time series of the anomalies' entropy value.

In order to locate the anomalies, we apply the wavelet packet transformation, Choi-Williams distribution and Pseudo Wigner-Ville distribution methods to anomaly entropy time series. The results are shown in Fig. 4.

From the Fig.4, we can make sure some anomalies （worms）certainly exist in 20:40 Dec.18.

In order to find out other anomalous time point, we should obtain the rule of the existing anomaly time point. Firstly, we must extract an anomalous time series segment between 146 point and 155 point (from 20:10 to 20:55 on Dec. 18). Then we apply PAA and SAX to the symbolic series segment. The outcome is shown by Fig. 5.

Let $H(srcIP) = 1, H(srcPort) = 2, H(dstPort) = 3, H(Octets) = 4,$ $H(prot) = 5, H(dstIP) = 6$, 'A' denotes the lowest entropy value, 'B' denotes the lower entropy value, 'C' denotes the medium entropy value, 'D' denotes larger entropy value, 'E' denotes the largest entropy value. Then we apply association rules mining to alphabet sequence of eight days (from Dec. 18 to Dec.25) to get the association rules of the anomaly pattern in the backbone IPLSng router:

$$1EE, 2AA, 4AA, 5AA \Rightarrow 3EE \ (\sup = 11, conf = 100)$$

……

The rule $1EE, 2AA, 4AA, 5AA \Rightarrow 3EE$ means that the distribution of source address is dispersive ($E \to E$, the curve doesn't increase and doesn't decrease), the distribution of source port is concentrated, the distribution of octets is concentrated, and the distribution of protocol is dispersive is also concentrated, from which we can infer that the distribution of destination port is dispersive. Hence this rule can be used to identify whether the anomaly is worm or not. The experimental result is shown in Table III.

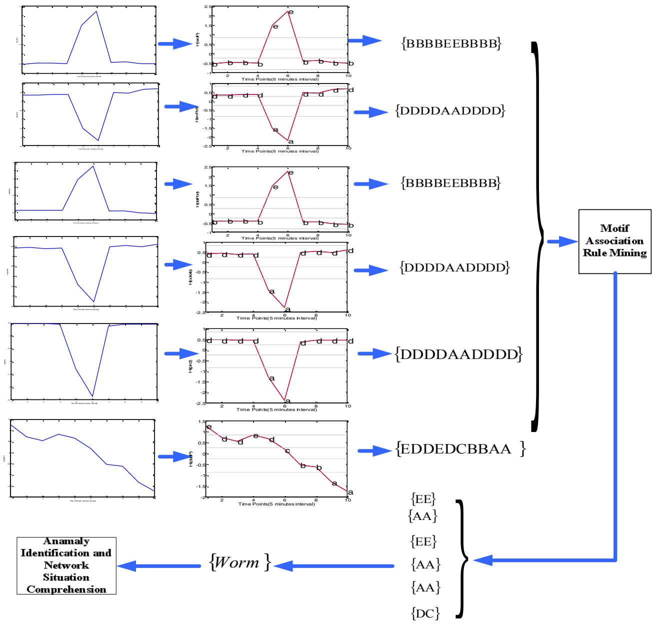

{BBBBEEBBBB}
{DDDDAADDDD}
{BBBBEEBBBB}
{DDDDAADDDD}
{DDDDAADDDD}
{EDDEDCBBAA }
Time Points(5 minutes interval)
Motif Association Rule Mining
{EE}
{AA}
{EE}
{AA}
{AA}
{DC}
{Worm }
Anamaly Identification and Network Situation Comprehension

Figure 5. Apply SAX on 6 time series

TABLE III. THE TIME DISTRIBUTION OF "$1EE,2AA,4AA,5AA \Rightarrow 3EE$"

Day	18	19	19	19	19	19	19	19	19	21	25
Hour	20	05	05	06	08	20	20	20	20	18	16
Minutes	40	30	40	40	40	00	10	20	30	20	00

ACKNOWLEDGMENT

The authors would like to thank Abilene for the providing Netflow data and will thank reviewers for their helpful comments. This research supported by Chinese National Science Foundation under grant No.60572092, Program for New Century Excellent Talents in University and National Basic Research Program of China ("973 program") under grant No.2007CB307100.

REFERENCES

[1] http://abilene.internet2.edu.

[2] http://www.snort.org

[3] Ke Wang, Salvatore J.Stolfo, "Anomalous Payload-based Network Intrusion Detection", the *7th International Symposium on Recent Advances in Intrusion Detection*,2004

[4] Masaki Ishiguro, Hironobu Suzuki, Ichiro Murase, Hiroyuki, "Internet Threat Detection System Using Bayesian Estimation," the *16th Annul FIRST Conference on Computer Security Incident Handling*,2004

[5] Lakhina,A.,Crovella,M.,and Diot,C,"Mining anomalies using traffic feature distributions". In *ACM SIGCOMM*, Philadelphia, Pennsylvania, USA, 2005),pp.217–228.

[6] Hotelling,H,"Analysis of a complex of statistical variables into principal components,"*J. Educ. Psy*.(1933), 417–441.

[7] Lin,J.,Keogh,E., Lonardi, S. & Chiu, B, "A Symbolic Representation of Time Series, with Implications for Streaming Algorithms," In *proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, San Diego, CA. June 13, 2003

[8] Lin,J.,Keogh,E., Patel, P. & Lonardi, S, "Finding Motifs in Time Series". In *proceedings of the 2nd Workshop on Temporal Data Mining, at the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.Edmonton, Alberta, Canada. July 23-26, 2002

[9] Eamonn J. Keogh ,Michael J. Pazzani, "An Enhanced Representation of Time Series Which Allows Fast and Accurate Classification,Clustering and Relevance Feedback. In *International Conference on Knowledge Discovery and Data Mining*, pages 239–243, New York,NY, USA, August 1998.

[10] R.J.Larsen and M.L.Marx. *An Introduction to Mathematical Statistics and Its Applications*.2nd ed. Englewood,Cliffs,NJ:Prentice Hall.1986

[11] L.Cohen,*Time-Frequency Analysis:Theory and Applications*. Prentice Hall.1998

[12] R.Agrawal, T. Imielinski, A. Swami, "Mining association rules between sets of items in large databases," in *Proc. ACM SIGMOD Int. Conf. Management Data,* 1993, pp. 207–216.

[13] Chengqi Zhang, Shichao Zhang. *Association Rule Mining:models and algorithms*.2002