Network Anomaly Detection Based on Statistical Approach and Time Series Analysis

Huang Kai
School of software engineering
Shanghai Jiao Tong University
Shanghai, China
Email: huangkaikiki@sjtu.edu.cn

Qi Zhengwei
School of software engineering
Shanghai Jiao Tong University
Shanghai, China
Email: qizhwei@sjtu.edu.cn

Liu Bo
School of software engineering
Shanghai Jiao Tong University
Shanghai, China
Email: poe.liu@gmail.com

Abstract

Network always suffers from the traffic anomaly such as router rate change, device restart or the worm attack. The early detection of unusual anomaly in the network is a key to fast recover and avoidance of future serious problem to provide a stable network transmission. In this paper we present a statistical approach to analysis the distribution of network traffic to identify the normal network traffic behavior. We adapt the EM algorism to estimate the distribution parameter of Gaussian mixture distribution model. If only there is a statistical signature of unusual fluctuation or change in the network traffic an alarm will be triggered. We adapt the time series analysis of the statistical analysis result. Up bound and low bound will be defined through the analysis. The exceeding of the bound will be the signal of traffic anomaly. Another time series analysis approach also can reflect the fluctuation of network with the crossover of two indicator lines called K line and D line. These two indicator lines are some think like the mean value of the historical data in a time slice with one more sensitive to the change of the new coming data and another not. The approach three-MACD indicator approach is like the K D approach but more blunt to the unusual fluctuation of network traffic which can submit an alarm more correctly.

Keywords: *EM algorism, Gaussian Mixture Model, K and D indicator approach, MACD(Moving Average Convergence and Divergence)*

1. Introduction

The network has been widely used in every field. Different applications, new protocols and new type of networks have made the network changes greatly every day. Also there are more and more kinds of viruses and network attacks in the nowadays network. So the signature based approach cannot be a good solution because the signature data base always should be updated for the new kinds of virus or attack. So the statistical approach would be a better solution to meet the needs.

Nevertheless traffic of different kinds of protocol has different kind of statistical distribution. Telnet connection and ftp control connection can be validly modeled by the Poisson process. The Self-similar process can be a better model for the WAN arrival processes. The telnet package can be properly described by the exponential process [1].

In our approach, we adapt the Gaussian Mixture Model to approximate the combined statistical model. The Gaussian Mixture Model can be a model with less residual in the network distribution of combined traffic of different type. Then we use the EM algorism to estimate the value of different Gaussian distribution. We use the sum of all the value to evaluate the fluctuation of the network traffic. If a great change happened in a short time slot, an alarm will be triggered.

We adapted two approaches to decide how great the change is and when to trigger an alarm. The first approach defines an up bound and low bound of the value. If the sum of all the values passes the bound, an alarm will be triggered. The second approach uses the historical mean value of the sum in a time slot to draw two lines with one more sensitive to the change of the new coming data. If the more sensitive one pass through the anther an alarm will be triggered.

2. Related works

In recent years, a lot of work has been done in the network anomaly detection. Machine learning techniques have been widely used on detecting network anomalies recently such as the n nearest-neighbor methods [2], also the famous Neural Network [3], support vector machines [4]. Some other approaches like the genetic computation [5], Bayesian networks [6], outlier detection [7], Y-means clustering algorithm [8], Probability Statistics [9-11] have been adapted in the anomaly detection and analysis work.

Nevertheless the machining learning approach cannot be proven secure [12]. Also most of these approaches should analysis large amount of source data. The great time cost is well known to us all such as the Neural Network is a good example. It cannot be used in the real time system. These approaches also have difficulty in determining the scope of

parameter standard, the lack of flexibility and high rate of false alarm, etc.

Other approaches are bases on the network anomaly signature [13]. Different applications, new protocols and new type of networks have made the network changes greatly every day. Also with the development of wireless network and ADHOC network, the signature based approach cannot be a good solution for the network anomaly because you should update the solution to keep up with new attack or new development of network [14].

Best approach should be the statistical analysis. A lot of statistical method has been adapted in the network traffic analysis and anomaly detection. Because the network traffic of different protocol has different characteristic and distinctive statistical distribution. So the approach to model the combined traffic with one distribution process will not work well except for special application. Some research has proved the network traffic has the pattern of self-similar. A lot of work has been done to analysis the network traffic with self similar process. However the analysis of our source data we find the residual is not normal if modeling the traffic with self similar approach. So we model the traffic with Gaussian Mixture Model which can describe the network traffic distribution well enough.

3. Statistical Approach

Because the network traffic of different protocol has different characteristic and distinctive statistical distribution. We just test the distribution of the source data and finally proved that the residual of the statistic result is not normal. The statistical data with Gaussian distribution should be with the normal residual.

3.1. Gaussian Mixture Model

After analysis of the source data, we find the network traffic cannot be described as a Gaussian distribution. The distribution of a Gaussian should be like the shape of ellipse and its residual should be normal. We tested the statistical source data and finally find that it is not gaussian. So we adapt the Gaussian mixture model to approximate the unknown distribution [16-17].

The Gaussian mixture model probability density function is a weighted average of several Gaussian distribution. Here we take the Gaussian mixture model with three single Gaussian distribution as an example.

$$p(x) = \alpha_1\, g(x;\mu_1,\theta_1) + \alpha_2\, g(x;\mu_2,\theta_2) + \alpha_3\, g(x;\mu_3,\theta_3)$$

The parameter list $(\alpha_1, \alpha_2, \alpha_3)$ must meet with the listed condition:

$$\alpha_1 + \alpha_2 + \alpha_3 = 1$$

The single Gaussian mixture distribution can be described as:

$$g(x;\mu,\sigma^2) = (2\pi)^{-d/2}\,\sigma^{-d}\exp[-\frac{(x-\mu)^T(x-\mu)}{2\,\sigma^2}]$$

The more Gaussian models, the more accurate the Gaussian mixture model will be. In our approach we find the amount of Gaussian distribution will influence the time cost and performance of our approach.

3.2. EM Algorithm

EM is an iterative technique for estimating the value of some unknown quantity, given the values of some correlated, known quantity.

The approach is to first assume that the quantity is represented as a value in some parameterized probability distribution (the popular application is a mixture of Gaussians, hence the example below). The EM procedure, then, is [15]:

Initialize the distribution parameters Repeat until convergence: E-Step: estimate the Expected value of the unknown variables, given the current parameter estimate

$$Q(h'|h) = E[\ln p(Y|h')|h, X]$$

M-Step: re-estimate the distribution parameters to maximize the likelihood of the data, given the expected estimates of the unknown variables

$$h \leftarrow \arg\max Q(h'|h)$$

At here, we use the EM algorism to estimate the mean value of different Gaussian distribution which overlap with each other to form the Gaussian mixture distribution.

3.3. Time Slice Window

We adapt the approach of the mixture of Gaussian model to match the network traffic distribution. Then we adapt the EM algorism to estimate the mean value of each Gaussian distribution.

The data in our approach is the data of time series. So the data should be divided with time slot. We call it the "window".

The size of the windows should be determined. From the source data the network traffic demonstrate the circular fluctuation of date and night. So the time slot window should be the integral times of the 24 hours. Here in our source data 1440 is the circular length. The calculation time delay can be adjusted. the time cost of will change greatly with the time delay. The experiment result will be given is the experiment section. Here we assume the value is 100.

The calculation window consequence should be:

$$window_n : t_n \rightarrow t_n + 1440$$

$$window_{n+1} : t_n + 100 \rightarrow t_n + 100 + 1440$$

3.4. Iterative Algorism

Through the application of EM we can get the mean value of these Gaussian distribution.

Step 1:

Calculate the $E[z_{ij}]$ for each hidden $variablez_ij$. Assume the current $h =< \mu_1, \mu_2 \cdots \mu_j >$.

$$E[z_{ij}] = \frac{p(x = x_i \,|\mu = \mu_j)}{\sum_{n=1}^{2} p(x = x_i \,|\mu = \mu_n)}$$

$$= \frac{e^{-\frac{1}{2\sigma^2}(x_i - \mu_j)^2}}{\sum_{n=1}^{2} e^{-\frac{1}{2\sigma^2}(x_i - \mu_n)^2}}$$

Step 2:

We adapt the maximum likelihood method to calculate the $h' =< \mu'_1, \mu'_2, \cdots \mu'_j >$, we take the $E[z_ij]$ as an estimation of z_ij. then we replace the $h =< \mu_1, \mu_2 \cdots \mu_j >$ with $h' =< \mu'_1, \mu'_2, \cdots \mu'_j >$.

$$\mu_j \leftarrow \frac{\sum_{i=1}^{m} E[z_{ij}]\, x_i}{\sum_{i=1}^{m} E[z_{ij}]}$$

To repeat these two steps we can easily achieve the estimated mean value μ_j of distribution j.

4. Judgment on Time Series Analysis

4.1. The Up and Low Bound Approach

After we calculate the mean value μ_j, we add them all into one and check whether the value changes greatly. If so we can assert that there may be some traffic anomaly in the network traffic.

$$z_{up}(t) = \overline{x(t)} + k * r(t)$$

$$z_{down}(t) = \overline{x(t)} - k * r(t)$$

$\overline{x(t)}$ is the mean value of mean value μ_j in the latest m samples;

$$\overline{x(t)} = \frac{x(t) + x(t-1) + x(t-2) + \cdots + x(t-m+1)}{m}$$

$r(t)$ is the standard deviation of mean value μ_j sum in the latest m samples;

$$A_i = \left(x(t-i) - \overline{x(t)}\right)^2$$

$$r(t) = \sqrt{\frac{\sum_{i=0}^{m-1} A_i}{m}}$$

k is a weighting factor of fluctuation.

The $z_{up}(t)$ defines the upper limit of mean value μ_j sum according to its tendency.

The $z_{down}(t)$ defines the down limit of mean value μ_j sum according to its tendency.

If the value exceeds the line a alert will be submitted.

The result is explained in experiment section. The k would be a configurable parameter which related with the fluctuation range of the normal network traffic behavior. Future research would be adapted for a adaptable approach which can learn the K value by itself.

4.2. The K and D Indicators Approach

This index calls the KD index. It is combined with the K line and the D line. This index reflects the relationship between highest value, lowest value of recent days and the value of the last day. This index can reflect the sudden increase or decrease of the network traffic.

The calculation approach is listed below:

$$k(n) = 100 * [\frac{(C(n) - L5)}{(H5 - L5)}]$$

$$D(n) = 100 * (\frac{H3}{L3})$$

In the formulation the C(n) is the value of time stamp n; L5 is the lowest value in the most recent 5 times. H5 is the highest value in the most recent 5 times. H3 is the sum of (C-L5) in three times. L3 is the sum of (H5 - L5) in three data points.

The K line is more sensitive to the change of the new coming data than the D line. So if the K line passes through the D line, a fluctuation of network traffic is indicated. So an alarm will be triggered. On the other side, the next cross would be the signal of normal which means the anomaly has passed away. However through the experiment you will find this approach is very sensitive to the small change of the network traffic. So if it is in the condition that the network remains stable, this approach would be very useful. Very small fluctuation would not be ignored by it.

4.3. The MACD Indicator Approach

Developed by Gerald Appel, Moving Average Convergence/Divergence (MACD) is one of the simplest and most reliable indicators available. MACD uses moving averages, which are lagging indicators, to include some trend-following characteristics. These lagging indicators are turned into a momentum oscillator by subtracting the longer moving average from the shorter moving average.

Calculate the EMA:

EMA of 12 data points:

$$EMA_{12}(n) = EMA_{12}(n-1) * \frac{11}{13} + data(n) * \frac{2}{13}$$

EMA of 26 data points:

$$EMA_{26}(n) = EMA_{26}(n-1) * \frac{25}{27} + data(n) * \frac{2}{27}$$

Calculation of DIF :

$$DIF(n) = EMA_{12}(n) - EMA_{26}(n)$$

Calculate the 9 data point EMA of DIF. The result will be the MACD. To distinguish it we rename it as the DEA or DEM.

$$DEA - MACD(n) = DEA(n-1) * \frac{8}{10} + DIF(n) * \frac{2}{10}$$

The result of DIF and DEA can be positive and negative.

In practice, the DIF and DEA(MACD) would be the sensitive short term index, and the blunt one with long term index.

The judgment approach is similar with the ¡K,D ¿index approach. If the DIF line pass through the DEA it demonstrate the start of anomaly, the next crossover of the line indicates the end of the anomaly.

5. Experiment Test

5.1. Source Data Achievement

We dump the source data from the CRDC (Cisco China Research & Development Center) Hong Kong and CRDC Beijing. We use TCL script get number of counters from MIB (Management Information Base). The data were collected from Aug 4 11:40:00 2008.

With drawing the figure of these data with MATLAB, we found so many counters remains zero. The result is shown in Figure 1.

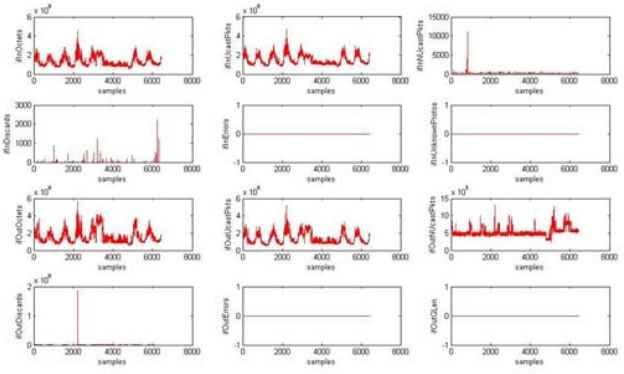

Figure 1. All Counters of gw2

	1 3	1 4	1 8	1 9	1 33	1 36	1 38
ifInOctets			AN				
ifInDiscards	X	X	X	X	X	X	X
ifInNUcastPkts	AN	AN	X	X	AN		
ifInUcastPkts			AN				
ifOutOctets			AN	AN		AN	
ifOutDiscards	X	X	X	X		X	X
ifOutNUcastPkts	AN	AN	X	X	X	X	X
ifOutUcastPkts			AN	AN		AN	

Table 1. source data analysis result (1)

1 39	2 3	2 4	2 11	2 12	2 25	2 31	2 32	2 34
	AN	AN	AN					
X		X	X	X	X	X	X	AN
X	AN	AN	X	X	X		AN	AN
	AN		AN					
X	AN		AN	AN	X			AN
X	X	X	X	X	X	X	AN	X
X	AN	AN	X	X	X	X	X	X
X	AN		AN	AN	X			AN

Table 2. source data analysis result (2)

5.2. Data Analysis

After analyzing the data 8-20, we found several useful issues:

Some interfaces are dead. The interface is dead if all of its statistics are zero. The dead interfaces here are gw1-32, gw1-35, gw2-5, gw2-13, gw2-24, gw2-26, gw2-27, gw2-33

Some of the 12 statistics are useless. These statistics are: ifInErrors, ifOutErrors, ifOutQLen, ifInUnknownProtos. In all of the interfaces, these statistics remain zero.

Based on these issues, we could get the table 1 and 2:

In this table, we have removed the dead interfaces and the useless statistics. The head line is the interface id (1_3 means gw1-Interface3). The leftmost column is about the 8 statistics (eliminate 4 out of 12).

There are three kinds of cells in this table:

1) "X": All of the observed values are zeros, it means we cannot use this value to detect anomaly. We call these data invalidate;

2) "AN": There are abrupt changes in the plotted figure. We figured out these changes by our eyes;

3) blank: The observed values are valid, and there are no significant changes in the figure.

With the help of qq-plot, we analysed these source data and found the distribution of the network traffic can't be described by Single Gaussian Distribution. Traditional Gaussian model is inappropriate in this scenario. The result is shown in Figure 2. So here we adapt the multi Gaussian model to describe the network traffic distribution.

5.3. EM Algorithm for Gaussian Mixture Model

The anomaly network traffic can be detected throw our approach. The source data reflects the network traffic's fluctuation. The fluctuation involves mainly two parts. First, the traditional circular fluctuation of day and night, second, the unusual traffic rate change. Figure 3 was the network traffic diagram of GW1.

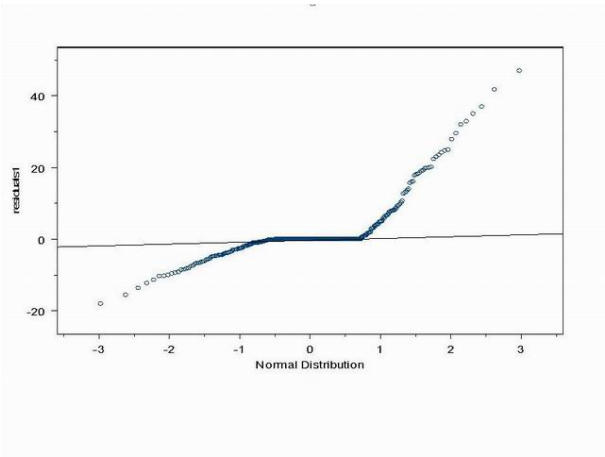

Figure 2. Gaussian test result

The Gaussian distribution amount means how many different Gaussian distribution these source data is combined with data conform to. Here we use 15 as the Gaussian distribution amount. It means that data of 15 different Gaussian distribution overlap with each other to form the final transaction. The estimated mean value of multi Gaussian model of GW1 are show in Figure 4.

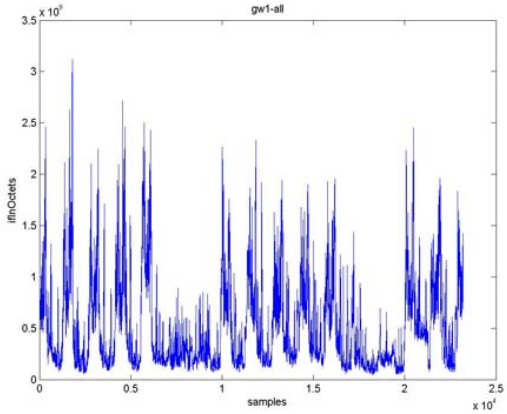

Figure 3. network traffic of GW1

You can find through the experiment result that the periodic fluctuation of network traffic with day and night is not reflected by the value of estimated mean value sum in Figure 4. But the unusual network traffic change has been reflected in the result.

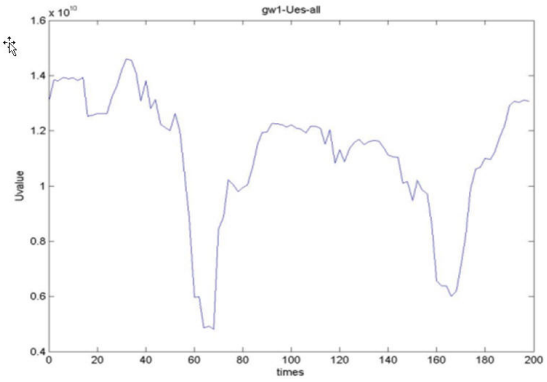

Figure 4. sum of estimated through EM algorism for GW1

Interval (data point)	50	100	500	1000	5000
Time cost (second)	1033.335654	541.583613	195.339136	146.205685	107.890785

Table 3. interval and time cost

5.4. The Time Cost and Calculation Interval Comparison

The calculation interval directly influence the time cost of the algorism. The corresponding relation is listed in Table 3. the unit in the picture is second. You can find the time cost changes stiffly.

5.5. Gaussian Amount Related Effect and Time Cost Analysis

The Gaussian amount as a parameter of multi-Gaussian model can greatly influence the effect of our algorism. We find that too little Gaussian amount will greatly influence the effect of EM algorism.

Through the comparison of Figure 4 and Figure 5. You can find that the amount of 5 is so bad in the performance. But with the amount more than 10 the performance is good enough and even more will not improve the performance.

5.6. The Time Cost and Gaussian Amount Analysis

The Gaussian amount will greatly influence the time cost. Sufficient Gaussian amount is enough. It is not necessarily the more the better. There relation ship between time cost and Gaussian amount is listed in Table 4. The unit of time cost is second.

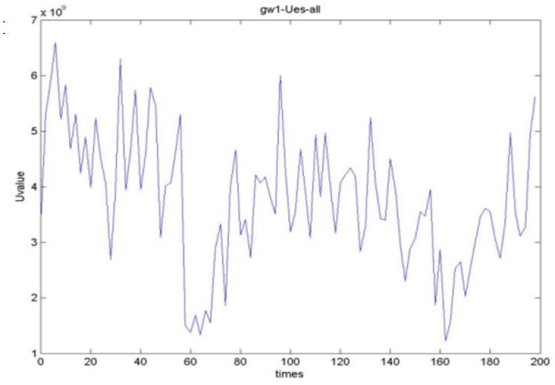

Figure 5. Gaussian amount 5 with GW1

Gaussian amount	5	10	15	25
Time cost	396.326250	411.428399	541.583613	934.317939

Table 4. Gaussian amount and time cost

5.7. Judge Approach on Time Series analysis

5.7.1. Up Bound Low Bound Approach. Here we tested the accurate of our judge approach with the k as 2 and m as 10. If the data line crosses with the up bound or the down bound the alarm will be submitted. The red line is the up bound, and the blue line is the low bound. The result is shown in Figure 6.

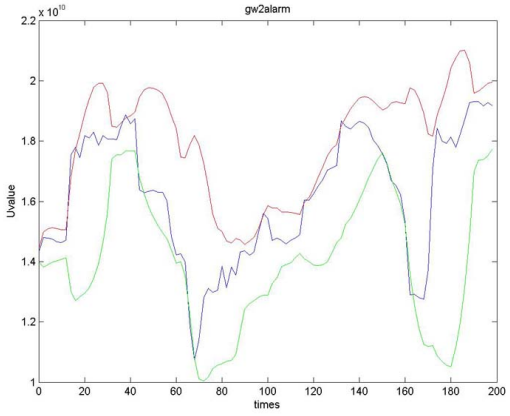

Figure 6. judge result with GW1

5.7.2. The K and D indicator Approach. Through the experiment you can find that every small fluctuation in network traffic will bring out a crossover of the K and D line. By comparing the Figure 7 and Figure 9, you will find every small fluctuation of estimated mean value in Figure 9 will bring out a crossover in Figure 7. So this approach well perform well in a network that its traffic is very stable. The red line is the K indicator line. The green line is the D indicator line.

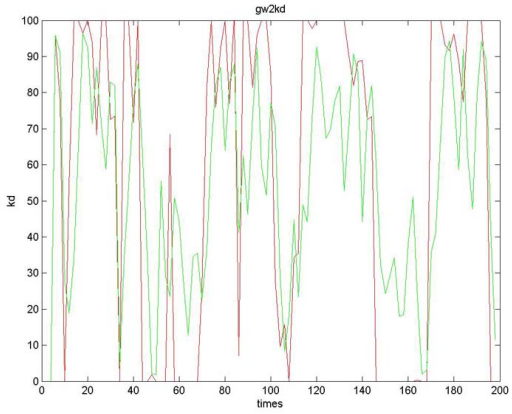

Figure 7. k and d indicator judge result with GW1

5.7.3. The MACD Approach. The approach with MACD index is not so sensitive like the K and D indicator approach. It works extremely well in our experiment in our network condition. You can compare the Figure 8 and Figure 9 . The judgment is very accurate and correct. The corresponding 3 anomaly fluctuation s of network traffic triggers the alarm correctly in the experiment result.

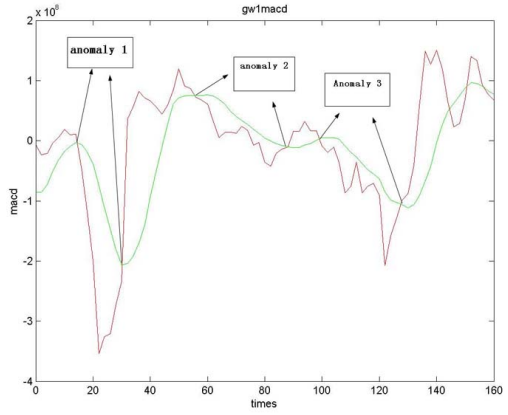

Figure 8. MACD indicator judge result with GW1

Of course, due to the unstable data at the beginning of the figure, the beginning part of our approach result needs to be neglected. It will not affect the remaining part of our approach and the final result.

6. Conclusion

This paper has presented our idea about the statistical anomaly identification of network traffic. By analysis of the

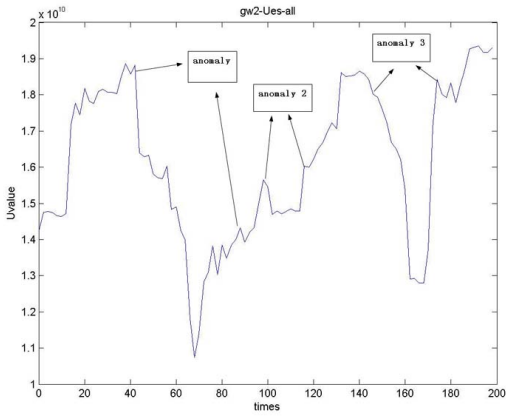

Figure 9. sum of estimated of GW1 for compare

source data achieved from the CRDC, we find the source data cannot be modeled by the single Gaussian model. So we adapt the Gaussian mixture model to monitor the unusual fluctuation of network traffic to submit an alarm at the proper time.

Through the experiment, we find that our statistical approach has good performance in the monitor of the unusual fluctuation of network traffic with the Gaussian mixture model. The performance is greatly related with parameters of our approach which includes the Gaussian distribution amount in the Gaussian mixture model, the length of the time slice window, the time delay during two calculation steps. Also these parameters are also related with the time cost of our calculation. We should find the balance point between the performance and time cost. At last we find the Gaussian amount 10 would be a best one with low time cost.

In the future work, we would implement some network attack to the network to analysis the different conditions like the router rate change, device restart, burst traffic, network attack such as the ICMP to find an adaptive algorism for these different conditions.

Acknowledgment

This work was supported by the National Nature Science Foundation of China under Grant No 60773093 and 60873209, the Microsoft Research Young Faculty Fund and the 863 Research and Development Program of China under Grunt No. 2006AA01Z169.

Thanks CRDC(Cisco China Research & Development Center) for the supply of the source data and the support in the research process. Especially thanks for the aids of Fred Baker who was Fellow at Cisco Systems.

References

[1] Vern Paxson and Sally Floyd, *Wide-AreaTraffic: The Failure of Poisson Modeling*. SIGCOMM, 1994.

[2] E. Eskin,A. Arnold,M. Prerau,L. Portnoy,and S. Stolfo, *A geometric framework for unsupervised anomaly detection: Detectingintrusions in unlabeled data*. Kluwer, 2002.

[3] Manikopoulos C, Papavassiliou S, *ANetwork intrusion and fault detection: A statistical anomaly approach*. IEEE Communications Magazine, 2002, 40 (10):7682.

[4] S. Peddabachigari, A. Abraham, C. Grosan, and J. Thomas, *Modeling intrusion detection system using hybrid intelligent systems*. Journal of Network and Computer Applications, 30(1):114-132, January 2007.

[5] W. Lu and I. Traore, *Detecting new forms of network intrusions using genetic programming*. Computational Intelligence,20(3):475-494, Aug.2004.

[6] D. Barbara, N. Wu, and S. Jajodia, *Detecting novel net work intrusions using bayes estimators*. In Proceedings of the First SIAM International Conference on Data Mining (SDM 2001), Chicago, USA,April 2001.

[7] W. Lu and I. Traore, *A novel unsupervised anomaly detection framework for detecting network attacks in real-time*. In 4th International Conference on Cryptology and Network Security (CANS), Xiamen, Fujian Province, China, December2005.

[8] Y. Guan, A. A. Ghorbani, and N. Belacel, *An unsuper vised clustering algorithm for intrusion detection*. In Proc. of the Sixteenth Canadian Conference on Artificial Intel ligence (AI 2003), pages 616-617, Halifax, Canada, May 2003. Springer.

[9] Staniford S, Hoagland JA, McAlerney JM, *Practical automated detection of stealthy portscans*. Journal of Computer Security, 2002, 10 (1/2):105136.

[10] Mahoney VM, *A machine learning approach to detecting attacks by identifying anomalies in network traffic*. Melbourne: Florida Institute of Technology, 2003.

[11] Wang K, Stolfo SJ, *Anomalous payload-based network intrusion detection*. In: Jonsson E, Valdes A, Almgren M, eds. Proc. of the 7th Int'l Symp. On Recent Advances in Intrusion Detection (RAID 2004). LNCS 3224, Heidelberg: Springer-Verlag, 2004. 203222.

[12] M. Barreno, B. Nelson, R. Sears, A. D.Joseph, and J. D.Tygar, *Can machine learning be secure?*. In ASIACCS '06: Proceedings of the 2006 ACM Symposium on Information, computer and communications security, pages 16-25, New York,NY, USA, 2006. ACMPress.

[13] F. Feather, D. Siewiorek, R. Maxion, *Fault detection in an Ethernet network using anomaly signature matching*. ACM SIGCOMM Computer Communication Review, 1993.

[14] Ilker Onat, Ali Miri, *AReal-TimeNode-Based Traffic Anomaly Detection Algorithm for Wireless Sensor Networks*. Proceedings of the 2005 Systems Communications (ICW'05) 2005.

[15] Tom M.Mitchell, *Machine Learning*. McGraw-Hill Science/Engineering/Math, 1 edition. (March 1, 1997).

[16] N Shental, A Bar-Hillel, T Hertz, D Weinshall, *Computing Gaussian mixture models with EM using equivalence constraints*. Advances in Neural Information Processing Systems, 2004.

[17] CE Rasmussen, *The infinite Gaussian mixture model*. Advances in Neural Information Processing Systems, 2000.