ACTA TECHNICA CORVINIENSIS – Bulletin of Engineering
Tome VI (Year 2013) – FASCICULE 4 [October–December]
ISSN 2067–3809

[1.] Petar ČISAR, [2.] Sanja MARAVIĆ ČISAR

IMPROVEMENT OF EXPONENTIAL SMOOTHING USING SIMULATED NETWORK ENVIRONMENT

[1.] ACADEMY OF CRIMINALISTIC AND POLICE STUDIES, BELGRADE – ZEMUN, SERBIA
[2.] SUBOTICA TECH, SUBOTICA, SERBIA

ABSTRACT: This paper gives a general overview of the EWMA statistics. In addition to already known concepts, the paper presents a comparative analysis of different smoothing schemes. The optimization possibilities of this statistics are also discussed. With intention to improve the process of exponential smoothing, the behavior of the moving trimmed mean and moving median in computer network environment was examined in relation to the most commonly used moving average. For this purpose, several different types of distribution are used to model network traffic. Using the software package "Matlab", sequences of random numbers are generated for each type of distribution, as the way to simulate network traffic. It is shown that the moving average and moving trimmed mean better follow the curve of original traffic. At the same time, comparing the situation with and without outliers, the smallest relative jump of MSE was determined for moving average and moving median. This conclusion justifies their use for the elimination of the negative impact of outliers and positively affects the general level of traffic control in computer networks.

KEYWORDS: computer networks, mean square methods, optimization, smoothing methods, statistical distributions

INTRODUCTION

Inherent in the collection of data taken over specific time is some form of random value variation. There exist various methods for reducing the unwanted effect due to random variation. An usual statistical technique is smoothing. This technique, when properly applied, detects more clearly the underlying trend, seasonal and cyclic components. There are two different categories of smoothing methods: averaging and exponential smoothing. Taking averages is the simplest way to smooth data. Given a series of numbers and a fixed subset size, the moving average can be obtained by taking the average of the first subset. The fixed subset size is then shifted forward, forming a new subset of numbers, which is averaged. This process is repeated over the entire data series. The line connecting all the calculated partial averages is the graphical representation of moving average. A moving average is a set of numbers, each of which is the average of the corresponding subset of a larger set of data points.

The exponentially weighted moving average (EWMA) is a statistic for monitoring the process that averages the data in a way that gives adjustable weight to data as they are further removed in time. For the EWMA control technique, the decision regarding the state of control of the process depends on the EWMA statistic, which is an exponentially weighted average of all prior data, including the most recent measurements.

By the choice of weighting factor λ, the EWMA control procedure can be made sensitive to a small or gradual drift in the process.

The statistic that is calculated is the following:

$$EWMA = \lambda Y_t + (1-\lambda)\cdot EWMA_{t-1} \qquad (1)$$

for t=1, 2, …, n

where

- EWMA0 is the mean of historical data
- Yt is the observation at time t
- n is the number of observations to be monitored including EWMA0
- 0 < λ ≤ 1 is a constant that determines the depth of memory.

This equation has been established by Roberts as described in [1]. The parameter λ determines the rate at which "older" data enter into the calculation of the EWMA statistic. A value of λ = 1 implies that only the most recent measurement influences the EWMA. Thus, a large value of λ = 1 gives more weight to recent data and less weight to older data - a small value of λ gives more weight to older data. The value of λ is usually set between 0.2 and 0.3 [2], although this choice is somewhat arbitrary. Lucas and Saccucci [3] have shown that although the smoothing factor λ used in an EWMA chart is usually recommended to be in the interval between 0.05 to 0.25, in practice the optimally designed smoothing factor depends not only on the given size of the mean shift δ, but also on a given in-control Average Run Length (ARL). ARL represents the average number of determined process points before the first point indices the appearance of out-of-control state (exceeding one of the control limits).

The estimated variance of the EWMA statistic is approximately [4]:

$$\sigma^2_{EWMA} = \frac{\lambda}{2-\lambda}\cdot\sigma^2 \qquad (2)$$

where σ is the standard deviation calculated from the historical data.

The center line for the control chart is the target value or EWMA$_0$. The upper and lower control limits are:

$$UCL = EWMA_0 + k\sigma_{EWMA} \qquad (3)$$

$$LCL = EWMA_0 - k\sigma_{EWMA} \qquad (4)$$

where the factor k is either set equal 3 (the 3-sigma control limits) or chosen using the Lucas and Saccucci tables (ARL = 370).

In addition to the aforementioned authors, the publications [5]-[12] have also dealt with the topic of EWMA statistics and the application of statistical anomaly detection in computer networks.

Control charts are specialized time series plots, which assist in determining whether a process is in statistical control. Some of the most widely used forms of control charts are X-R charts and Individuals charts. These are frequently referred to as "Shewhart" charts after the control charting pioneer Walter Shewhart who introduced such techniques. These charts are sensitive to detecting relatively large shifts in the process (i.e. of the order of 1.5σ or above). In computer network practice, shifts can be caused by intrusion or attack, for example. Two types of charts are usually used to detect smaller shifts (less than 1.5σ), namely cumulative sum (or CUSUM) charts and EWMA charts. A CUSUM chart plots the cumulative sums of the deviations of each sample value from a target value. An alternative technique to detect small shifts is to use the EWMA methodology. This type of chart has some very attractive properties, in particular:

1. *Unlike X-R and Individuals charts, all of the data collected over time may be used to determine the control status of a process.*
2. *Like the CUSUM, the EWMA utilizes all previous observations, but the weight attached to data exponentially decreases as the observations become older and older.*
3. *The EWMA is often superior to the CUSUM charting technique due to the fact that it detects larger shifts better.*
4. *EWMA schemes may be applied for monitoring standard deviations in addition to the process mean.*
5. *EWMA schemes can be used to forecast values of a process mean.*
6. *The EWMA methodology is not sensitive to normality assumptions.*

In real situations, the exact value of the shift size is often unknown and can only be reasonably assumed to vary within a certain range. Such a range of shifts deteriorates the performance of existing control charts.

COMPARISON OF SMOOTHING SCHEMES

Calculating the optimal value of parameter λ is based on the study of authentic samples of network traffic. Random variations of network traffic are normal phenomena in the observed sample. In order to decrease or eliminate the influence of individual random variations of network traffic on occurrence of false alarms, the procedure of exponential smoothing is applied, as an aspect of data preprocessing.

For any time period t, the smoothed value St is determined by computing:

$$S_t = \lambda\cdot y_{t-1} + (1-\lambda)\cdot S_{t-1} \qquad (5)$$

where 0 < λ ≤ 1 and t ≥ 3.

This is the basic equation of exponential smoothing. The formulation here is given by Hunter [2]. It should be noted that there is an alternative approach, in which, according to Roberts [1], y_t is used instead of y_{t-1}.

This smoothing scheme starts by setting S_2 to y_1, where S_i stands for smoothed observation or EWMA, and yi stands for the original observation. The subscripts refer to the time periods 1, 2, ..., n. For example, the third period is $S_3 = \lambda\ y_2 + (1 - \lambda)\ S_2$ and so on. There is no S_1. The optimal value for λ is the value which results in the smallest sum of the squared errors (SSE) or mean of the squared errors (MSE).

Comparative analysis of two different approaches (Roberts and Hunter) can be shown using the example of a process (y_t), with adopted values EWMA$_0$ = 50 and λ = 0.3. EWMA values in the table below correspond to Roberts's and S_t to Hunter's equation.

Table 1. Comparison of smoothing schemes.

time	y_t	EWMA	S_t
1	52.00	50.60	
2	47.00	49.52	52.00
3	53.00	50.56	50.50
4	49.30	50.18	51.25
5	50.10	50.16	50.67
6	47.00	49.21	50.50
7	51.00	49.75	49.45
8	50.10	49.85	49.91
9	51.20	50.26	49.97
10	50.50	50.33	50.34
11	49.60	50.11	50.39
12	47.60	49.36	50.15
13	49.90	49.52	49.39
14	51.30	50.05	49.54
15	47.80	49.38	50.07
16	51.20	49.92	49.39
17	52.60	50.73	49.93
18	52.40	51.23	50.73
19	53.60	51.94	51.23
20	52.10	51.99	51.94
21			51.99

In Table 1 the fields with approximately equal values are marked with lighter colour, while fields with equal values are marked with darker colour. From this analysis it can be concluded that after a certain number of samples (in this case about the 16th sample) both schemes give the same smoothed values.

The behaviour of both smoothing schemes will be examined also with SSE values. After calculating SSE for different λ, results were as follows.

Analysis of the obtained results has shown that approximately similar values were obtained, with greater coincidence at higher values of smoothing factor.

Table 2. Comparison of values for SSE according to Roberts and Hunter

λ	SSE (Roberts)	SSE (Hunter)
0.1	62.81	75.01
0.2	49.95	55.86
0.3	39.28	42.16
0.4	30.25	31.62
0.5	22.40	23.01
0.6	15.50	15.71
0.7	9.55	9.57
0.8	4.70	4.66
0.9	1.31	1.29

The initial EWMA plays an important role in computing all the subsequent EWMA's. There are several approaches to define this value:

1. *Setting S_2 to y_1*
2. *Setting S_2 to the target of the process*
3. *Setting S_2 to average of the first four or five observations*

It can also be shown that the smaller the value of λ, the more important is the selection of the initial EWMA.

ARL CURVES

Using a graphical method, the EWMA chart can be designed to have minimal ARL for the out-of-control situation, for the known shift of the mean δ and given ARL for the in-control situation. This chart has two parameters - λ and k (derives from the definition of control limits).

Figures below show the dependence of λ and k of the mean shift δ, for ARL as parameter. Using appropriate curves, values k = 2.7878 and λ = 0.1417 were determined as the optimal choice for the earliest detection of shift δ = 1σ.

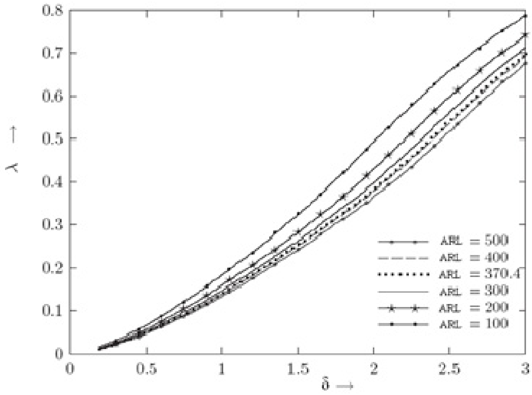

Figure 1. Optimal choice of λ in function of the mean shift [13]

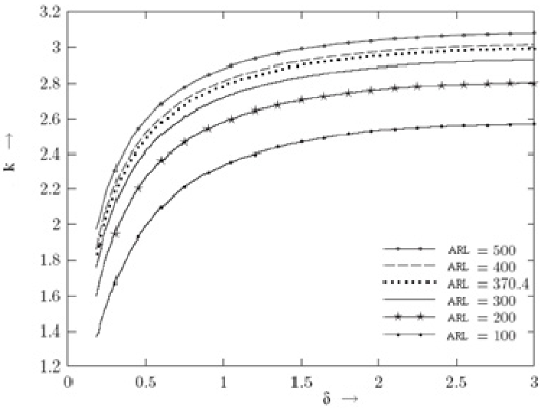

Figure 2. Optimal choice of k in function of the mean shift [13]

MOVING AVERAGE, MOVING MEDIAN AND MOVING TRIMMED MEAN IN SIMULATED NETWORK ENVIRONMENT

The calculation of moving average, based on time series network data, represents the basis of the application of EWMA statistics in network environment. This chapter will focus on the difference between moving average, moving trimmed mean and moving median, with the ambition to examine their behavior in a simulated network environment.

The changes in the computer network traffic are identified by analyzing time series information for one or more variables which indicates how the monitored variable changes over time. Network operators sometimes visually inspect such time series information to detect and characterize operational problems. However, it can be advantageous to inspect such time series information in an automated manner.

Unfortunately, outliers, data entry errors, or glitches exist in almost all real data. An outlier is an observation that lies an abnormal distance from other values in a random sample. In a sense, this definition leaves it up to the analyst to decide what will be considered abnormal. Before abnormal observations can be singled out, it is necessary to characterize normal observations. The sample mean is sensitive to these problems. One extreme data value can move the average away from the center of the rest of the data by an arbitrarily large distance, causing the situation of statistical anomaly or false alarm. The median and trimmed mean are two measures that are resistant (robust) to outliers. The median is the 50th percentile of the sample, which will only change slightly if you add a large perturbation to any value. The idea behind the trimmed mean is to ignore a small percentage of the highest and lowest values of a sample when determining the center of the sample. The geometric mean and harmonic mean, like the average, are not robust to outliers.

From a statistical point of view, the moving average, when used to estimate the underlying trend in a time series, is susceptible to rare events such as rapid shocks or other anomalies. A more robust estimate of the trend is the simple moving median. Statistically, the moving average is optimal for recovering the underlying trend of the time series when the fluctuations about the trend are normally distributed. However, the normal distribution does not place high probability on very large deviations from the trend which explains why such deviations will have a disproportionately large effect on the trend estimate.

There is no one single model that can be used effectively for modeling traffic in all kinds of computer networks. The type of network and the traffic characteristics dominantly influence the choice of the traffic model used for analysis. In the available literature the following types of distributions are used to model traffic: Poisson, normal, lognormal, Pareto, chi-square, Rayleigh, Weibull and gamma distribution. Using the software *package "Matlab" [15], sequences of 500 random numbers are generated for each type of mentioned distributions, as the way to simulate network traffic. For purpose of this research, the occurrence of outliers was simulated with 5 equal values, two times greater than the maximum value of all generated samples and distributed on every hundredth sample locations. Calculating the moving average, moving trimmed mean (10%, i.e. 5% of the highest and 5% of the lowest observations are excluded) and moving median, with width of the interval of 40 values (statistically large sample), the following values for the mean of the squared error (MSE) are obtained (values in brackets are MSE for the case without outliers):*

Table 3. MSE for different distributions.

Distribution	MSE (moving average)	MSE (moving trimmed mean)	MSE (moving median)
Pareto	2.92 (0.84)	2.99 (0.85)	3.16 (0.95)
Normal	2.1 (0.976)	2.112 (0.975)	2.119 (0.986)
Poisson	8.41 (4.72)	8.57 (4.75)	8.63 (4.86)
Lognormal	0.473 (0.196)	0.481 (0.196)	0.487 (0.199)
Rayleigh	3.15 (1.715)	3.19 (1.717)	3.26 (1.749)
Chi-square	10.5 (4.32)	10.8 (4.41)	11.5 (4.86)
Weibull	0.396 (0.209)	0.401 (0.209)	0.412 (0.214)
Gamma	20.19 (7.65)	20.67 (7.67)	21.49 (8.08)

Analyzing the above table, it can be concluded that for all the observed types of distribution, moving average and moving trimmed mean generates lower MSE than moving median, which implies that the moving average and moving trimmed mean better follow the curve of original network traffic. Besides, in the context of absolute values, the smallest MSE is obtained using Weibull, lognormal and normal distribution. Comparing the situation with and without outliers, the smallest relative jump of MSE (corresponds to the best robustness) was determined for moving average and moving median. The smallest values for relative jump are calculated for Poisson (about 80%), Rayleigh (about 85%) and Weibull distribution (about 90%).

CONCLUSIONS

In addition to already known concepts about EWMA statistics, this paper presents a comparative analysis of different smoothing schemes. It was shown that after a certain number of samples (approximately after the 15th sample), both known schemes provide the same smoothed value. With aim to improve the process of exponential smoothing, the behavior of the moving trimmed mean and moving median in computer network environment was examined in relation to the most commonly used moving average. For this purpose, network traffic is simulated using different types of distribution. It is shown that the moving average and moving trimmed mean curves better follow the curve of original traffic. At the same time, comparing the situation with and without outliers, the smallest relative jump of MSE (corresponds to the best robustness) was determined for moving average and moving median. This conclusion justifies their use for the elimination of the negative impact of outliers in the field of statistical anomaly detection and positively affects the general level of traffic control in computer networks.

REFERENCES

[1.] S.W. Roberts, "Control Chart Tests Based on Geometric Moving Averages", Technometrics, 1959., Vol. 42, No. 1, Special 40th Anniversary Issue, pp. 97-101, 2000.

[2.] J.S. Hunter, "The exponentially weighted moving average", Journal of Quality Technology 18: 203-210, 1986.

[3.] J.M. Lucas and M.S. Saccucci, "Exponentially weighted moving average control schemes: Properties and enhancements", Technometrics 32, pp. 1-29., 1990.

[4.] Engineering Statistics Handbook-EWMA Control Charts, http://www.itl.nist.gov/div898/handbook/pmc/section3/pmc324.htm

[5.] G. Fengmin, "Deciphering Detection Techniques: Part II Anomaly-Based Intrusion Detection", White Paper, McAfee Security, 2003.

[6.] S. Sorensen, "Competitive Overview of Statistical Anomaly Detection", White Paper, Juniper Networks, 2004.

[7.] X. Wu, V.A. Mahadik and D.S. Reeves, "A summary of detection of denial-of-QoS attacks on DiffServ networks", DARPA Information Survivability Conference and Exposition, 2003., Proceedings, Vol. 2, pp. 277-282.

[8.] A.S. Neubauer, "The EWMA Control Chart: Properties and Comparison with other Quality-Control Procedures by Computer Simulation", Clinical Chemistry, Vol. 43, pp. 594-601, 1997.

[9.] D. Seibold, "Enterprise Campus Security-Addressing the Imploding Perimeter", http://www.itsa.ufl.edu/2003/presentations/IntSec.ppt

[10.] S. Vasilios and F. Papagalou, "Application of Anomaly Detection Algorithms for Detecting SYN Flooding Attacks", Global Telecommunications Conference, 2004. GLOBECOM 04 IEEE, Vol. 4, pp. 2050-2054.

[11.] J. Viinikka and H. Debar, "Monitoring IDS Background Noise Using EWMA Control Charts and Alert Information", Recent Advances in Intrusion Detection, Lecture Notes in Computer Science, 2004, Volume 3224/2004, pp. 166-187.

[12.] Y. Zhao, F. Tsung and Z. Wang, "Dual CUSUM Control Schemes for Detecting a Range of Mean Shifts", IEEE Transactions 2005 (37), pp. 1047-1057.

[13.] S.B. Vardeman and J.M. Jobe, "Statistical Quality Assurance Methods for Engineers", John Wiley & Sons, New York 1999.

[14.] V. Ivanova, MIT Academic Computing, "MATLAB Tutorials", http://web.mit.edu/acmath/matlab/course16/16.62x/16.62x_Matlab.pdf

ACTA TECHNICA CORVINIENSIS - BULLETIN of ENGINEERING

ISSN: 2067-3809 [CD-Rom, online]

copyright © UNIVERSITY POLITEHNICA TIMISOARA,
FACULTY OF ENGINEERING HUNEDOARA,
5, REVOLUTIEI, 331128, HUNEDOARA, ROMANIA
http://acta.fih.upt.ro