Application of Statistical Process Control Methods for IDS

M.Sadiq Ali Khan
Department of Computer Science University of Karachi
Karachi,Sindh-75230, Pakistan

Abstract

As technology improves, attackers are trying to get access to the network system resources by so many means. Open loop holes in the network allow them to penetrate in the network more easily; statistical methods have great importance in the area of computer and network security, in detecting the malfunctioning of the network system. Development of internet security solution needed to protect the system and to with stand prolonged and diverse attack. In this paper Statistical approach has been used, conventionally Statistical Control Charts has been used for quality characteristics however in IDS abnormal access can be easily detected and appropriate control limit can be established. Two different charts are investigated and Shewhart chart based on average has produced better accuracy. The approach used here for intrusion detection in such a way that if the data packet is drastically different from normal variation then it can be classified as attack. In other words a system variation may be due to some special reason. If these causes are investigated then natural variation and abnormal variation can be distinguished which can be used for distinction of behaviors of the system.

Keywords: *Intrusion Detection System; Denial of Service attack; Shewhart Chart; Cusum Chart*

1. Introduction

Today's networks entirely depend on the information sharing as it is the demand to cooperate and provide a simple and secure means of communication. However sometimes security enhancement increases the complexity of the system; this may result in software vulnerabilities and errors in installed applications of the network [1]. We need an expert based IDS in order to keep the network system more secure, despite of having technological advancement installed in our network [2]. For the secure architecture of the network computing, following are the requirements include high-assurance security, secured process for the submission and retrieval, high usability and scalability [3]. As infrastructure of the internet provides no supportive security services, this increase the demand for increasing the internet security mechanism and provides a framework to develop a new security protocol [4]. Statistical Control Charts is conventional methods to monitor process to ensure that systems are stable and minimize abnormality by checking abnormal variability. Although control chart are used for quality improvement. However, by observing the other side of the coin it can be used to determine how a particular data set is different from normal variation. The control chart works on basic idea that each process / data acquires controlled variation which is natural but others may have on uncontrolled variation which is not present in the system.

2. Background

Problems like tempering of data and unauthorized access, the attacks of different types occur in the network that forced the administrator to make a foolproof security measures. IDS are becoming one of the vital tools for the private networks. Research

about Intrusion Detection has become more vigorous with the current high penetration due to illicit accesses to computer systems. In order to defend against DoS attacks there are some methods available but these are not perfect and have difficulty with large distributed attacks [5]. Statistical based methods used for software security, wireless internet security, internet security, distributed embedded firewall based applications, increase the resiliency of commercial firewalls, detection of passive denial of service attacks, policy based authentication and authorization mechanism, in evolution of viruses and worms and for the development of probabilistic information system for the web[6]. Currently deployed networks and installed applications are not fully capable to detect the attacks reliably [7]. So attempt, have been made to identify some problems of Intrusion Detection in order to find the root causes of the unreliability of IDS.

2.1 Major Areas of IDS in Network Security

In the domain of network security IDS play a vital role. Security is usually implemented as a multi layer infrastructure and different security approaches can be categorized into the following major areas [8].

Avoidance of Attacks: In order to prevent the launching of an attack we should try to increase the amount of apparent danger of harmful consequences for the intruder. A strong ID system is required for the attack avoidance. However, it requires well-built verification against the attacker in case an attack is launched. Methods used in this area are discussed, which may effectively mark out the actual source of attack. Mechanism for attack avoidance by using the cryptography is discussed in [9].

Prevention of Attacks: Before reaching the targeted machine it aims to prevent an attack by blocking. Practically it is very hard to prevent all kinds of attacks due to incomplete knowledge for the attacks and allow normal activities.

Deflection of Attacks: It refers to trapping the intruder by the system and the attacker intentionally made to reveal the attack. For instance honey pots discussed in [10].

Detection of Attack: It refers to the process of detecting an attack when it is still in progress or to detect such kind of attacks which occurred in the past. It is more significant in order to system recovery and to take preventive measures for occurrence of similar kinds of attacks in future.

In intrusion detection system also abnormal behavior of network activity or attack can be distinguished from normal activity by estimating variability in quality characteristic. The most commonly used charts for quality control are Shewhart chart and Cusum Chart which explained in the next section.

2.2 Shewhart Chart

Shewhart chart is developed to investigate the variability of some quality characteristics of data under consideration. The sample statistics of quality characteristics is used to establish control limits. The control limits are used to compare normal or abnormal behaviors. Let ***w*** be a sample statistics that measure the varying quality of particular parameter ***u***. Let μ_w and δ_w are mean and standard deviation of ***u***. Then central line, upper control limit (UCL) and lower control limit (LCL) can be defined as follows:

$$\text{Centre line} = \mu_w$$
$$UCL = \mu_w + k\,\delta_w$$
$$LCL = \mu_w - k\,\delta_w$$

Where **k** is the distance of control limit from centre. However normally k =3 is used as standard so

$$UCL = \mu_w + 3\,\delta_w$$
$$UCL = \mu_w - 3\,\delta_w$$

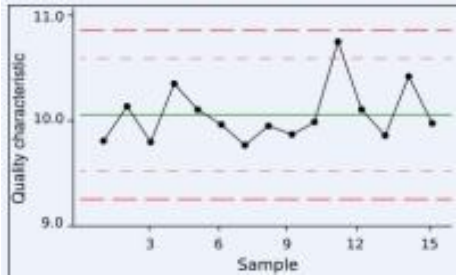

Fig. 1: Shewhart Chart

Shewhart may have different types depending on the nature of data. Following three types are commonly used *Shewhart Sample Mean (X-Chart), Shewhart Sample Mean (R-Chart), Shewhart Sample (¯X-Chart)*

2.3 Cusum Chart

Shewhart chart effectively monitor quality characteristics, however one limitation of the chart is it inability to predict small shift from the mean. Cusum chart is another option for monitoring small shift from mean. The Cusum chart was originally developed by Page and is developed by calculating and plotting cumulative sum against sample or time. Typical Cusum chart is depicted hereunder in Fig.2.

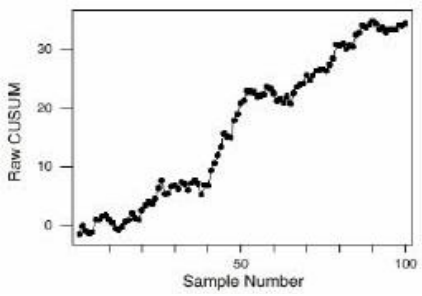

Fig. 2: Cusum Chart

In this section cumulative sum is calculated using sample of quality characteristics count. Let x_1,x_2,x_3,x_4 represent sample of size n with mean μ_0. Let Q_i is the target quality characteristics, and then Cusum chart may be defined mathematically as

$$Cusum_i = (Q_i - \mu_0) + Cusum_i\text{-}1$$

$Cusum_i$ is the i^{th} cumulative sum of the quality characteristic. In this paper tabular method is used for control. In the tabular form deviation from μ_0 is determined by using control limits, which are called high side Cusum or low side Cusum.

Upper limit/ Upper Cusum.

$$\mathbf{U_i = max\ [0,\ (\ x_i - (\mu_0 + Q\ ) + U_{i-1}\ ]}$$

Lower Limit/Lower Cusum

$$\mathbf{L_i = max\ [0,\ (\mu_0 + Q) - x_i + L_{i-1}\ ]}$$

2.4 Multivariate Statistical Process Control Methods for IDS

In Statistical control charts generally small number of attributes are used. However, the quality of services/product depends upon many variable which runs in hundred in some cases. It is practically difficult to use more than two or three quality chart to control Intrusion Detection System. As it is helpful only a few events are driving a system at any time; and useful combinations of measurement usually depict the same underlying events. Multivariate SPC techniques which can be utilized for large attribute/ multivariate data. SPC reduce the information of all attributes of a process to two or three composite metrics through statistical modeling [11]. These composite metrics may be used for controlling the overall process. The statistical control charts process is completed in two steps. In first step it is tested whether process was in control with first subgroup of attributes. In this step chart is used to bring the process within the statistical control. In the second step it is tested whether the process remain in control if another group is added [12].

To utilize quality control chart in multivariate scenario, all kind of control charts like Shewhart, Cusum and EWMA can be defined for multivariate also. Multivariate Control Chart is usually based on Mahalanobis distance statistics. In the form of distance statistics can be defined as follows:

a)

$$X_i^2 = n\,(\overline{X_i} - \mu)\,\Sigma^{-1}\,(\overline{X_i} - \mu)'$$

for i=1, 2,.....m rational subgroups, where **n** is the sample size of each rational subgroup (with n=1 for individual observations), **μ** is the vector of known means, Σ is the known covariance matrix and finally $\overline{X_i}$ is the vector of samples means for the i^{th} rational subgroup

b) $T_i^2 = n (\overline{X}_i - \overline{\overline{X}}_i) \bar{S}^{-1} (\overline{X}_i - \overline{\overline{X}}_i)'$

, for i=1,2,.....m where $\overline{\overline{X}}_i$ is the pooled vector of sample means calculated using the **n** observed sample mean vectors, and $\bar{S}$ is the pooled sample covariance matrix. The T_i^2 , and χ_i^2 statistics represent the weighted distance of any point from the target (process mean under stable conditions). Under the assumption that the m samples are independent and the joint distribution of the **p** attributes is the multivariate normal, the χ_i^2 follows a chi-square distribution with **p** degrees of freedom and the T_i^2 follows **p(m-1)(n-1) / mn –m-p+1** times an **F** distribution with **p**, mn – m – p+1 degrees of freedom. Thus, the appropriate probability limits may be obtained using the known distributions of the corresponding statistic. In general Shewhart control charts is based on covariance matrix or generalized matrix which is the sum of the variances of the attributes. However, an inherent limitation of Shewhart chart is that it is insensitive to small shift in mean; therefore, multivariate Cusum chart (MCUSUM) and exponentially weighted moving average control chart can be used to overcome the problem.

3. Methodology

The basic idea to develop Shewhart chart is to determine how a particular sample is different from normal variation. In other words how to ring an alarm for abnormal behavior or to detect accurately the systematic change in data. This is determined through Average Run Length. By ARL it is determined (on the average) low long chart, be plotted before we detect a point beyond the control limit. Among these parameters 2 parameters count and source_byte are more deterministic. Count parameter can be studied reliable to network activity with respect to time. Count factor depend upon many factors as studied in Bayesian network [13]. The IDS based on single quality characteristics effectively detected anomalies behavior. However, in large dimension data variability of attributes may contribute factors which are invisible in single attribute. Multivariate approach is used to investigate and capture this dimension. The 41 features of KDD data set are reduced to 14 features using PCA [13]. The multivariate process is adopted using these 14 features. This method also considers the addition of deletion of subgroups of attributes. Multivariate Shewhart Chart or Cusum Chart can be used. The Shewhart Chart clearly indicate better result as compare to single attribute Shewhart Chart as in case of single attribute false alarm are 9 which are reduced to 5 in multivariate method. Therefore, count is selected as quality characteristic to determine uncontrolled variability. Data set was classified into normal and attack irrespective of the type of attack. Fifty random samples each of size 5 are collected from normal and attacks records. Both data sets were analyzed using Shewhart means (x) and S-chart. As the sample size was small so R-chart is also plotted to compare with other charts. Unbiased estimate of variance δ was calculated as discussed.

4. Result and Observations

4.1 S-chart

The control limits for normal as depicted in Figure 3a were established as 23.852 ± 25.975. It is interesting to note that false alarm is more common in normal than attack. As much as 9 false alarms is reported (Table 1) relative to variation from normal variability which suggests that system is more efficient in learning the attack as compared to normal activity. On the other hand control limit for attack was established as 2.462 ± 1.79(Figure 3b). In this case one sample crossed the control limit.

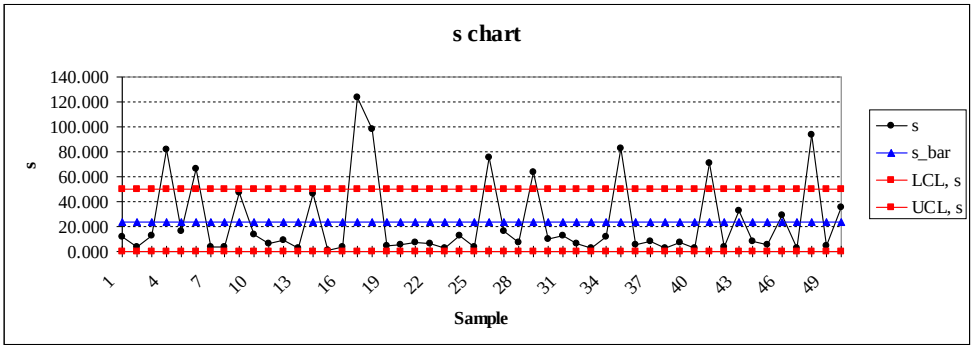

Fig. 3a: Shewhart S-chart for Normal

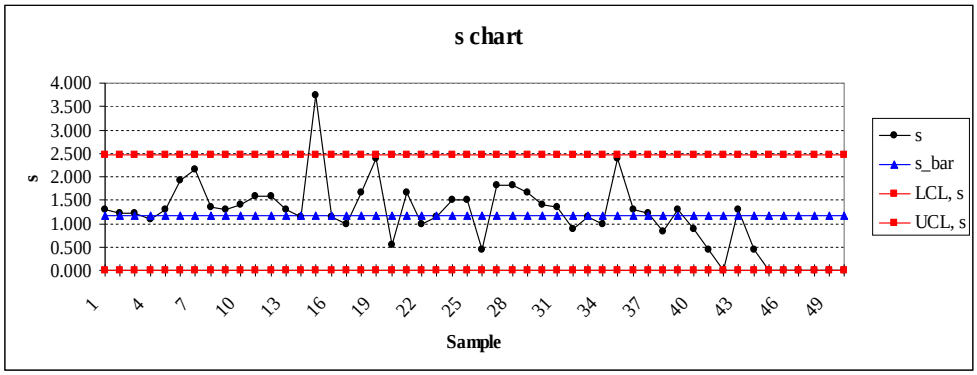

Fig. 3b: Shewhart S-chart for attack

4.2 X- chart:

Sample mean is used to draw x-chart. The control limits for normal were established as 19.0 ± 41.5(Figure 4a). As much as 3 false alarms are reported relative to variation from normal variability which suggests that system is more efficient in learning the attack as compare to normal activity. On the other hand control limit for attack was established as 1.9 ± 2.167(Figure 4b). In this case two samples crossed the control limit.

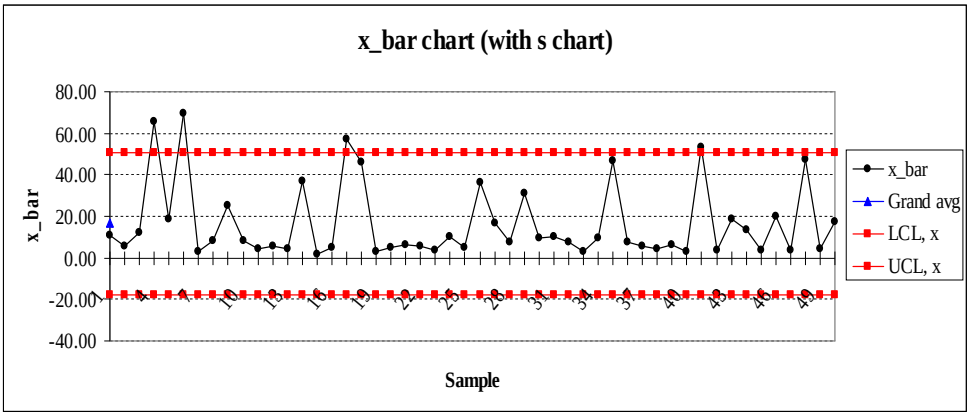

Fig. 4a: Shewhart X-chart for Normal

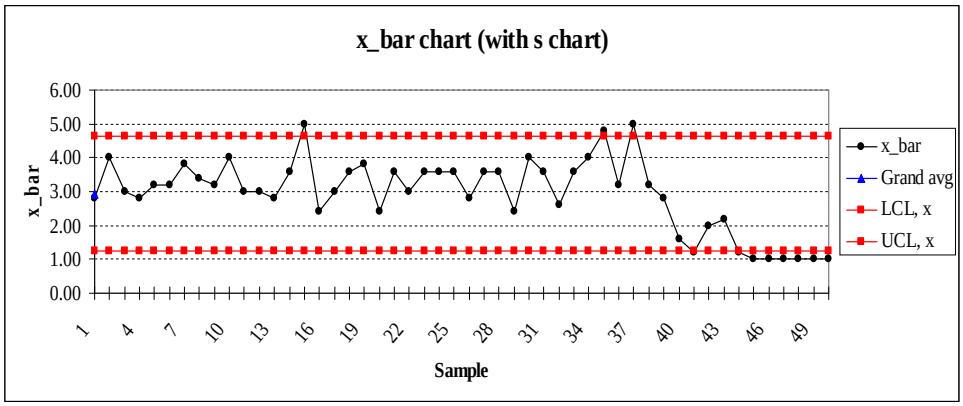

Fig. 4b: Shewhart X-chart for Attack

4.3 R-chart

Data Range R was used to draw Shewhart R-chart. The control limits for normal were established as 50.8 ± 60.0(Figure 5a). As much as 9 false alarms are reported relative to variation from normal variability which suggests that system is more efficient in learning the attack as compared to normal activity.

On the other hand control limit for attack was established as 2.282 ±3.143 (Figure 5b). In this case one sample crossed the control limit.

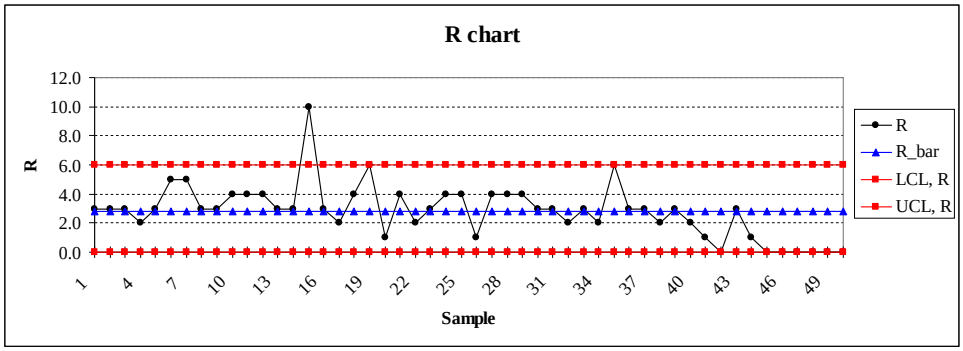

Fig. 5a: Shewhart R-chart for Normal

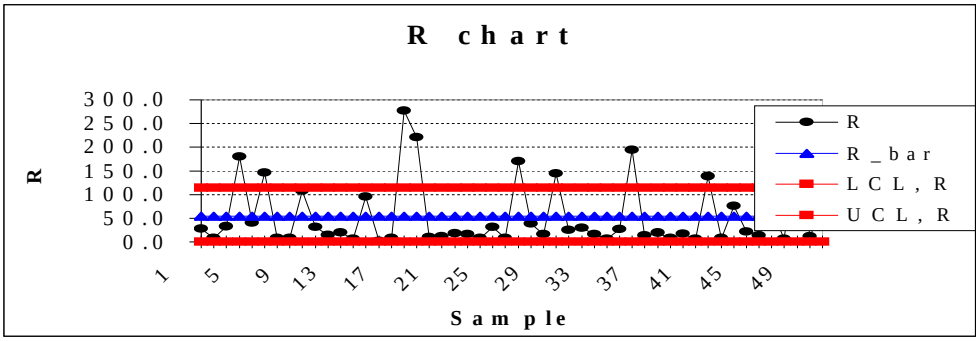

Fig. 5b: Shewhart R-chart for Attack

Table 1: False Alarm by Shewhart chart types

Chart Type	False Alarm	
	Normal	Attack
S-chart	9	1
X-chart	3	4
R-chart	9	9

It is evident from Table that Shewhart x-chart is more efficient as compared to others as it is more successful in detecting false alarm both in normal and attack.

4.4 Cusum Chart Results

In this section tabular form is used to determine out of control process i.e. deviation from normal activity of network. Sample of size 10 for parameter count are taken and cumulative sum of deviation from mean is calculated. As appeared in Figure 6 the upper and lower Cusum limit lies between +199.5 and -199.5. The data set only have normal activity parameter. Similarly sample of size 10 are taken from attack type data set and Cusum chart is plotted in Figure 7. In attack type upper and lower cusum limits
lies between +61.4 and – 61.4.

Decision interval for classifying normal or attack can be based on Cusum limit obtained in case of normal activity. Therefore, for decision interval lies between 60 and 200. It is interesting to note that target value is set at 120 as compared to target value in case of attack. However, if the target value is changed then other Cusum limits may be obtained. The normal data clearly indicate that sample are normally distributed along average and in case of attack data it is evident that data have steep slope beyond 3rd sample predicting abnormal behavior.

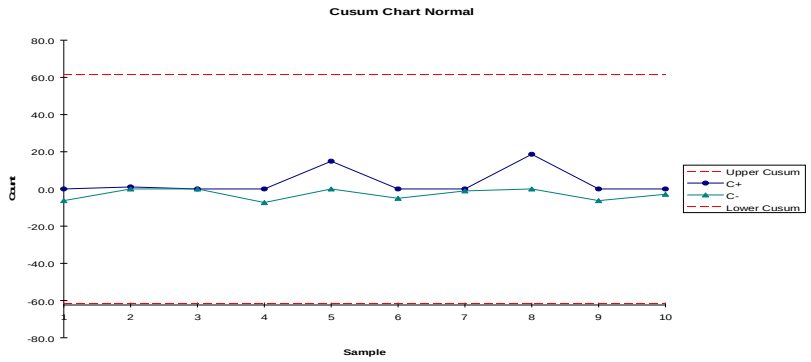

Fig. 6: Cusum Chart in normal data

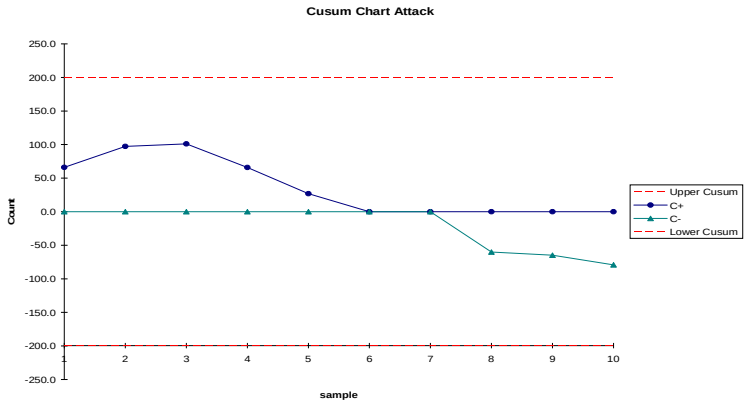

Fig. 7: Cusum chart in Attack

Table 2: Shewhart chart data sample (Normal)

Process characterization based on:				
	std. dev		Range	
x_bar	s	s/c4	R	R/d2
2.80	1.304	1.387	3.0	1.290
4.00	1.225	1.303	3.0	1.290
3.00	1.225	1.303	3.0	1.290
2.80	1.095	1.165	2.0	0.860
3.20	1.304	1.387	3.0	1.290
3.20	1.924	2.046	5.0	2.150
3.80	2.168	2.306	5.0	2.150
3.40	1.342	1.427	3.0	1.290
3.20	1.304	1.387	3.0	1.290

Table 3: Shewhart chart data sample (Attacks)

Process characterization based on:				
	std. dev		Range	
x_bar	s	s/c4	R	R/d2
10.60	11.675	12.420	28.0	12.038
5.80	3.564	3.791	8.0	3.439
12.40	12.482	13.279	32.0	13.758
65.60	81.574	86.782	180.0	77.388
19.00	16.155	17.187	40.0	17.197
69.40	66.278	70.509	146.0	62.771
3.00	3.464	3.685	8.0	3.439
8.20	3.493	3.716	8.0	3.439
25.40	46.891	49.885	108.0	46.433

The IDS based on single quality characteristics effectively detected anomalies behavior. However, in large dimension data variability of attributes may contribute factors which are invisible in single attribute. Multivariate approach is used to investigate and capture this dimension. The 41 features of KDD data set are reduced to 14 features using PCA[13].

The multivariate process is adopted using these 14 features. This method also considers the addition of deletion of subgroups of attributes. Multivariate Shewhart Chart or Cusum Chart can be used. The Shewhart Chart (Figure 8) clearly indicate better result as compare to single attribute Shewhart Chart as in case of single attribute false alarm are 9 which are reduced to 5 in multivariate method.

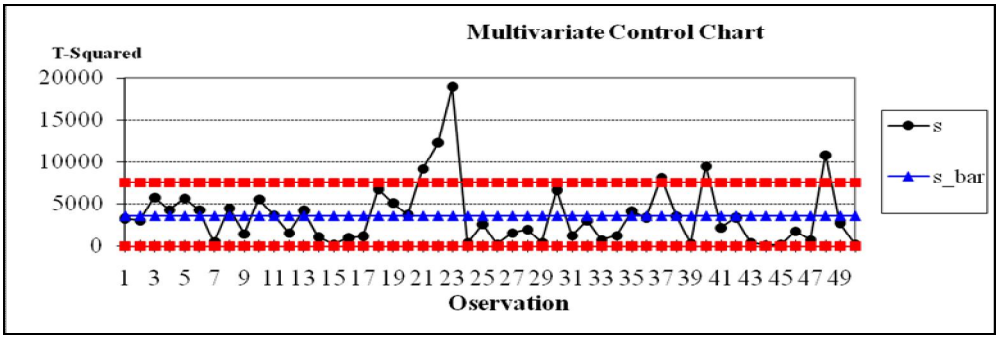

Fig. 8: A multivariate Shewhart Type Control Chart

5. Conclusion

In network security policy formulation, network models centralized or distributed systems are the good way to connect the user with a single platform. However, when such model and systems are implemented, different types of threat from system surrounding and from the boundary of a system are faced. Therefore we need to introduce new methods and techniques to avoid the threat faced by the systems. Intrusion Detection System needs to accommodate prevailing abnormal use of network beside with new variants. Secondly, some of the features or parameters of data packet are highly interdependent and contribute to particular attack and on the other hand in some cases only

single attribute can predict abnormal access of the network. The use of statistical control chart is common in many fields for quality control. However its use in Intrusion Detection is new and not very common, although this simple approach may be very effective in many cases. In Intrusion Detection some attacks which are common can be detected by investigating single quality characteristics. For instance probing attacks can easily be detected using frequency of use of particular service. Therefore single quality characteristics *count* is used to determine the attack. Quality control chart can present real time situation through graphical representation of system behavior. In Statistical Quality control chart different statistics can be used, however mean chart found to be most effective and control boundaries facilitate the range for determining the attack. To further improve the accuracy Cusum chart is also used. The methods studied in this paper effectively detect the intrusion like other machine learning method. However statistical methods may perform better as they not only simple in implementation but also support a graphical presentation which is more user's friendly. Multivariate application of quality control for intrusion detection can be further investigated and applied in future for identification of various groups of attributes which are more relevant in intrusion.

References

[1] Daniel Bilar; "Known Knowns, Known Unknowns and Unknown Unknowns: Anti Virus issues malicious software and Internet attacks for non-technical audiences", 2010.

[2] S.M.Aqil Burney and M.Sadiq Ali Khan; "Network Usage Security Policies for Academic Institutions", International Journal of Computer Applications, October Issue, Published By Foundation of Computer Science, 2010.

[3] Paul C.Clark and at.el; "Secure Compartmented Data Access over an Untrusted Network using a COTS-Based Architecture", Published in Statistical Methods in Computer and Network Security by Dekker, 2005.

[4] Matt Bishop and at.el; "Internet Security", Published in Statistical Methods in Computer and Network Security by Dekker, 2005.

[5] Davis J. Marchette; "Passive Detection of Denial of Service Attacks on the Internet", Published in Statistical Methods in Computer and Network Security by Dekker, 2005.

[6] W. W. S Chen; "Statistical Methods in Computer Security", Marcel Dekker Publications, ISBN: 0-8247-5939-7, 2005.

[7] James Cannady, Jay Harrell; " A Comparative Analysis of Current Intrusion Detection Technologies", 2009.

[8] Christopher Kruegel, Fredrik Valeur, and Giovanni Vigna. Intrusion Detection and Correlation:Challenges and Solutions. Springer, 2005.

[9] Rodica Tirtea, Geert Deconinck; Fault Detection Mechanisms for Fault Analysis Attacks Resistant Cryptographic Architecture; Third International Conference on Systems, Signals & Devices; March 21-24, Sousse, Tunisia, 2005.

[10] Lokesh D. Pathak and Ben Soh; Incorporating Data Mining Tools into a New Hybrid-IDS to Detect Known and Unknown Attacks;Ubiquitous Intelligence and Computing;Lecture Notes in Computer Science, Volume 4159/2006, 826-834, DOI: 10.1007/11833529, 2006.

[11] Emad Shihab and at.el; "Understanding the impact of code and process metrics on post-elease defects: a case study on the Eclipse project", Proceedings of the ACM-IEEE International Symposium on Empirical Software Engineering and Measurement, ISBN: 978-1-4503-0039-1, 2010.

[12] S.Bersimis, J.Panaretos and S.Psarakis; "Multivariate Statistical Process Control Charts and the Problem of Interpretation: A Short Overview and Some Applications in Industry", Proceedings of the 7th Hellenic European Conference on Computer Mathematics and its Applications, Athens, 2005.

[13] Burney S. M. Aqil, Sadiq Ali Khan, Jawed Naseem, “Efficient Probabilistic Classification Methods for NIDS” , (IJCSIS) International Journal of Computer Science and Information Security, Vol. 8, No. pp168-172, , 2010.

M.Sadiq Ali Khan holding Ph.D in Computer Science with specialization in network security (2011), MS in Computer Networks (2003) and BS in Computer Engineering (1998). He is currently working as an Assistant Professor at Department of Computer Science University of Karachi since 2003. He has 14 years of teaching and research experience and his research interest includes Data Communication & Networks, Network Security & Cryptography & Wireless Network Security. He is a member review committee of some reputable international and national level journals and a member of several computer societies. Recently he received a 10^{th} Teradata National IT Excellence Award. He is member of CSI, PEC and NSP.