International Journal of Security and Its Applications
Vol. 10, No. 9 (2016), pp.87-98
http://dx.doi.org/10.14257/ijsia.2016.10.9.10

Chi-Square Statistical based Technique for Intrusion Detection

Sheenam[1] and Abhinav Bhandari[2]

[1]Research scholar, [2]Assitant Professor, Punjabi university, Patiala
Dept. of Computer science and Engineering
sheenamgoyal24@gmail.com[1], bhandarinitj@gmail.com[2]

Abstract

Tools required for the security purposes are firewall, passwords, IDS, IPS for the detection of anomaly and prevent it from sending out the harmful traffic to the network. So, it is very necessary to examine the behavior of traffic that coming to the network and recognize the anomalous behavior. In this paper, statistical based chi-square method is used to detect the anomalous behavior and predict the intrusions by calculating the observed and expected frequencies. Setting of interval is difficult for the detection of anomaly but in our case we set interval according to the less variation in traffic. Traffic contains from the backscatter dataset. Chi-square method is good to detect the anomalous behaviors because it gives the Poisson's distribution for the whole traffic on network. Large difference shows anomaly occurs.

Keywords*: Chi-square(X^2), backscatter dataset, IDS, PCAP*

1. Introduction

Intrusion detection system (IDS) plays a major role to develop computer system and network security structures. Common attack like DOS attacks, DDOS attacks that prevent users to access the system. Links on network increases so resources used on network are costly and require more storage space. Now days, network speed has increased to 40 Gbps. Network intrusion detection system is productive or application software. NIDS is used to programmed the network intrusion detection process. From the various apex or points, NIDS can be inspecting like freight catch process, network position, and compute preference and amid alternatives. It has been clarified that network freight are allocate as normal or anomalous behavior in intrusion detection. There are following types for intrusion detection(IDS): Misuse detection and Anomaly detection.

Misuse detection compares observations with the well known patterns of attacks and if match found considers as intrusion and tools used for detection are IDS and IPS. The Problem occur here is that weakness emerging but not detect with unknown pattern. Example, a guard must handle the database of culprit's picture. A user will not allow entry whose picture matches with the database. There is problem in signature based is that a picture not present in the database consider as a valid user. It may be possible users are attackers. Anomaly based detection provides the information about the novelty attacks. Security guard matches the pictures with the database pictures. Database contains the valid user pictures. If picture found in the database, then user will allow entry. But if not found in the database then not entry will allow. Anomaly consists of following parts: *training* and *detection* part. Normal outline built under *training*. Observed outline is compared with the normal outline in *detection* part. If two outlines are far apart, alarm generated or we can say anomaly detected. Anomaly detection describes about the known or unknown attacks. In detection part, *(a) data gathering and storage (b) data preprocessing and analysis* intrusion detection system consists of above two phases. Data captured in the data gathering and storage phase and then data exported in storage. In further step, calculate the flow of network by statistic and features representing traffic

ISSN: 1738-9976 IJSIA

behaviors are extracted[1]. Now, find the difference between the non anomalous and anomalous by performing the analysis.

Number of attacks like Dos, DDos, scan, flood attacks affects the normal profile of the system. It is very necessary to identify these attacks and gain the detailed information about the attackers so to prevent system from further attacks. Attackers firstly know about the weakness about the system or the network. Once he knows about the vulnerabilities of system, they exploit the system resources and take the system's control. So, to prevent the system from attacks in future we have to detect the harmful traffic at initial stage in proper manner[2]. There are tools scan and flood are used to prevent from attacks for avoiding firewall canon and previous intrusion detection[2]. In these days, Anomaly detection becomes an important research area in network security.

On Section 2 we have mentioned the *technique used,* on 3 *Related Work,* 4 *Proposed Work,* 5 *Results and discussions* and 6 *Conclusion.* Every section has been discussed one by one as follows.

2. Technique Used

From the literature survey IDS can be categorize into following techniques: (*i)* Statistical technique *(ii)* Data mining technique *(iii)* Knowledge based technique *(iv)* Machine learning IDS.

2.1. Statistical Approach

In statistical method, observations are checked under statistical measures so to generate the behavior of normal profile by use of mean, variance, standard deviation and covariance. Profile may be checked and anomaly score detected by comparison of behaviors. Anomaly score represents the degree of uniqueness. Statistical based consist of following methods*:*

1) Univariate model: Characteristics of Gaussian random variables are contained in this model [3].

2) Multivariate model:This model contains related more than onecharacteristic [1].

3) Time series model:Time occurrence is considered and it identifies by abnormal activities[1].

2.2. Data Mining Approach

To remove the temporary elements when made the Intrusion detection system by different process, researchers must look at the data mining technique. Following are the data mining techniques *Clustering and Classification methods.*

2.2.1. Clustering Method: Clustering is mechanism in which data must be grouped based on its similarity. Clusters are made according to the given similar data and a central point is considered for each cluster. A new data provided to that central point. If data not matched with any cluster, then this data considered as anomalous data. Clustering is unsupervised learning approach.

2.2.2. Classification Method: Classification based on *training* and *testing.* In training phase, classifier builds. In testing phase, check whether is data is anomalous or non-anomalous.

2.3. Machine Learning

Based on the performance of prior results or observations, machine learning constructs a model. According to the new gathered information machine learning helps to change the execution plan. It may consist of following parts:

1) Bayesian network: Bayesian networks encrypt the chances of interrelationship between the variables.

2) Neural Network: Neural network is traditionally being used in IDS. By this, examine the behavior of different users. Neural network has advantage to tolerate the vague data without having the previous information.

3) Genetic algorithms: It is used to find the optimal solution in the search space and distinguish the normal and anomalous attack traffic in IDS.

4) Fuzzy logic: Fuzziness helps to separate the abnormal behavior with the normal behavior. Fuzzy logic contains the 0 and 1 value.

2.4. Knowledge Approach

Determine the normal behavior of system by the set of routines in knowledge based approach. This method takes the long time to establish the knowledge.

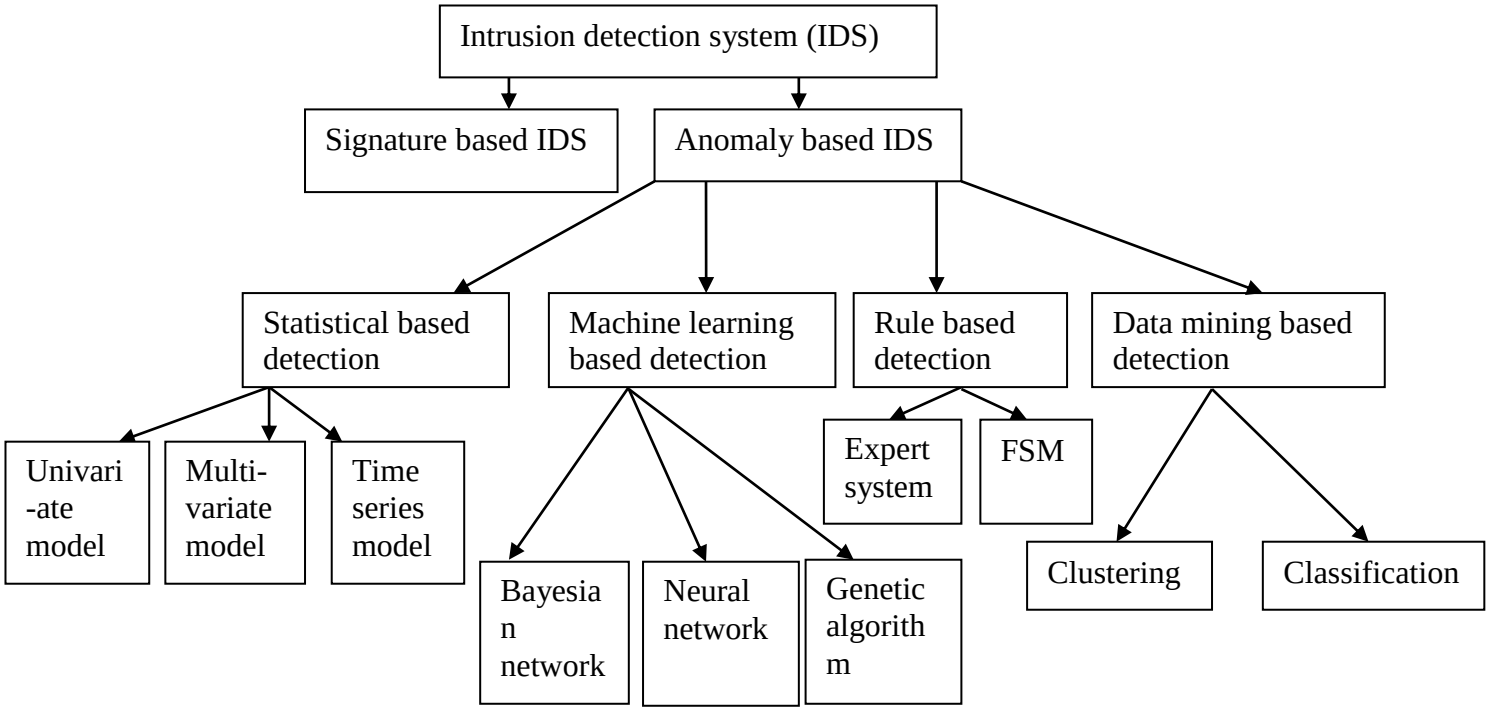

Figure 1. Taxonomy of IDS

3. Related Work

Survey by Varun, Arindam and Praveen [4] describes about the anomaly based detection and gives comprehensive outline of research on it and described about the complications of methods in IDS. Qayyum *et al.* [5] described the instruction in the different areas of application for statistical based anomaly detection technique and discussed about the advantages and disadvantages of it.

From research point of view, for anomaly based detection, chi- square is mainly used because of its good performance and clarity. Nong Ye *et al.* [6] it is based on the examination of packets. SM Masum and Nong Ye [7],In real time, described about the strength and flexibility of $X^2$with Canberra method. Nong Ye, CM Borror and D parmar [8] described about distinction between the Hotelling T^2 and Chi-square mechanism. Results showed that Chi-square gives more accurate results. In anomaly based detection examination of packets can be costly during high speed network traffic. For network controlling X^2, is widely used with flow based data to eliminate the expensiveness[2].

Santaigo-Paz *et al.* [9] proposed the entropy and mahalanobis distance technique in network traffic to detection of anomalies. By using the mahalanobis distance, check the given traffic is normal or abnormal. Apart from the proposed approach, researchers concentrate on the source and destination ip address of entropy [10]. Firstly, detection defines the individualistic number of variables like source and destination address. secondly, in packet header withdraw the sign. At last, calculate the entropy and chi-square according to features. To find the entropy, mathematical formula is here: With the probability ofPi, *n*represents the symbol then entropy can be calculated as [9]:

$$H = -\sum_{i=1}^{n} P_i \log_2 P_i$$

Traditionally, Snort is most widely used for record assayer [11]. It is open-source locator for detection of intrusions on network. The programs that are arranged in snort, it detects only these types of attacks from the set of rules. Denning was explained a real time model for intrusion detection in the field of security [11]. 7th USENIX Security Symposium San Antonio, Texas [12] gives the popular and standard method for detection of intrusions. Lee and Stoflo described the various data mining approaches like classification learning, frequent, association learning for intrusion detection [13, 14]. RIPPER learning algorithm used in classification developed by Cohen [11].

Chi-square approach is based under statistical anomaly detection because it gives Poisson distribution. Others test does not provide the Poisson distribution. P test and Z test use the normal distribution.

4. Proposed Work

4.1. Chi-square (X^2)

Chi-square test provides the different characteristics to put on any univariate scattering so to calculate the acumulative frequency. Chi square is used to measure the instance data came from the population that is specified. It can be applied to the set of bins [11]. Histogram and frequency table calculated for the non-binned data before applying X^2 technique. X^2 based on assumptions i.e. Null hypothesis (H_0**)** and alternative hypothesis (H_a**)**.

H_0: Data that pursue particular dissemination.
H_a: Data that do not pursue particular dissemination.
Chi-square for test data can be calculated as

$$X^2 = \sum_{j=1}^{n} (O_j - E_j)^2 / E_j$$

where O_j observed frequency,
E_j Expected frequency.

Expected frequency defined as:

E_j = Relative frequency(f)*Observed frequency for each(n). Relative frequency can be calculated as individual occur divided by total number of possible results.

4.2. Material

CAIDA BACKSCATTER 2007 dataset is used for the detection of anomaly. From Literarture survey, we concluded that the statistical technique is good for the detection of intrusions. Dataset contains the 2463638 rows with different number of attributes. Attributes are Number, Time, Source, Destination, Protocol, Length, Info. Our research implemented in Database with queries.

4.3. Methodology

The approach is modeled for intrusion detection system. Firstly backscatter dataset is analyzed provided by CAIDA. Backscatter contains the attributes like source address, destination address, Time, Protocol, information *etc*. Then, dataset should open in the wireshark. It must contains in the PCAP (packet captured) format. In our dataset, it contains the TCP, ICMP and IPV4 protocols. The large PCAP file is split into smaller ones. The useful features are eradicated and converted into CSV format through the wireshark. Then, file run in the database. Firstly, make the 180sec or 3 minute interval for protocolslike tcp, icmp and ipv4. Find the average packets per second for the whole distribution. And find the relative frequency based on average packets per second for whole distribution. Further, find the observed frequency based on making the intervals. Then, calculate expected frequency and proceed with the following steps to calculate the values to detect the intrusions.

After splitting the files into smaller ones, following are shown in table. The relative frequency is defined as calculating individual or object divided by the total numbers of individual or object based on calculating the average number of packets per sec. Expected frequency can be calculated by multiplying the relative frequency with total observed frequency.

The limitation of our Research is time window analysis is the big tradeoff. Setting the interval is the big issue. Another is that We can do the comparative analysis of our technique with the another statistical technique. There are number of statistical techniques for the detection of anomaly or intrusions.

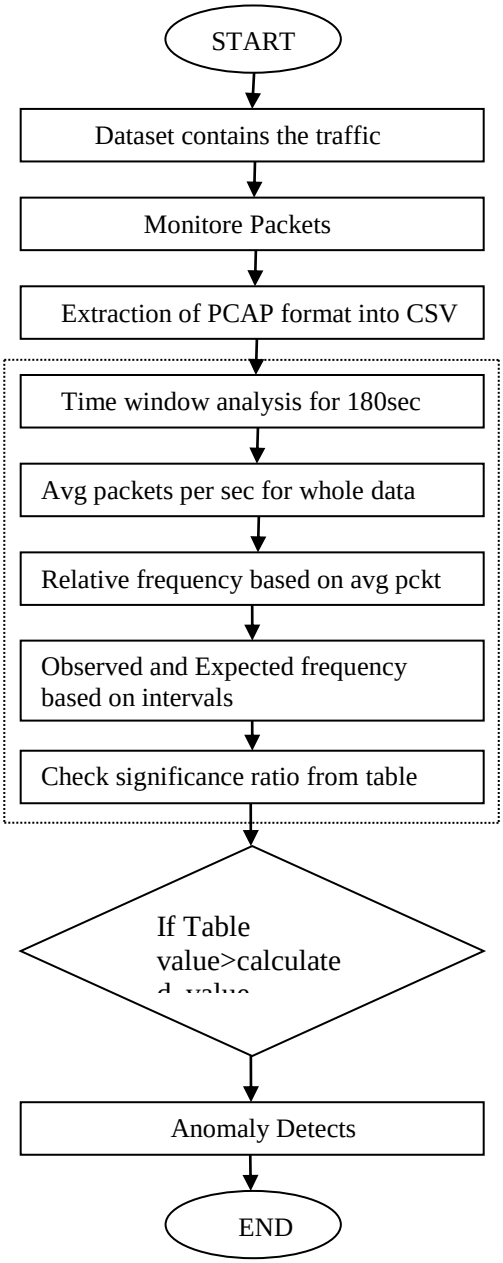

Figure 2. Flow Chart of Model

Now, we want to check anomaly in dataset, we choose the first interval T1 between 0 to 180 sec for anomaly detection. In this work, we make the intervals according to the records in our dataset. We make seven different intervals according to time window. We choose interval 0 to 180 for anomaly detection.

H_0: Interval T1 has particular dissemination or no anomaly
H_1: Interval T1 does not have particular dissemination or anomaly occurs.

5. Results and Discussion

Table 1. Packets Average per sec

Categories	No. of average packets per sec	Relative frequency
ICMP	167.768333333333	0.23719879163191
IPV4	0.4275	6.04419686408687E04
TCP	539.09416666666	0.762196788681682
Total	707.29	1

Table 2. Packets Average per sec

Categories	No. of average packets per sec	Relative frequency
ICMP	166.8175	0.262010842676481
IPV4	0.37166666 6666667	5.83755880097066E04
TCP	469.4925	0.737405401443422
Total	636.68166666667	1

Table 3. Packets Average per sec

Categories	No. of average packets per sec	Relative frequency
ICMP	172.5083333 33333	0.243292159789392
IPV4	0.530833333 333333	7.48645504013539E04
TCP	536.0191666 66667	0.755959194706594
Total	709.0583333 33333	1

The following Tables 4, 5, 7 have shown the different values of intervals 0 to 180. Firstly we, split the large PCAP file into smaller one because work done on large file is difficult. After calculating the different values, add all total values of Tables 4, 5, 6. It will give the one value of interval between 0to180. Like this, calculating all intervals values. In our work, I made 7 different intervals of 3 minutes. The calculating values of table are given below:

Table 4. Frequency Table

Categories	Relative Frequency(f)	Observed Frequency(O)	Expected Frequency(E)	(O-E)	$(O-E)^2$	$\frac{(O-E)^2}{E}$
ICMP	0.237198 79163191	152.272222 222222	152.612384 764906	-0.340162 5426838	0.115710555 445108	7.5819898 6427974E04
IPV4	6.044196 86408687E04	0.37777777 7777778	0.388880268 348202	-1.110249 05704245	1.232652968 66365E04	3.16974932 644291E04
TCP	0.762196 788681682	490.744444 444444	490.393179 41119	0.3512650 33254265	0.123387123 58712	2.51608563 83702E04
TOTAL	**1**	**643.394**				**Total= 132678.2482**

Table 5. Frequency Table

Categories	Relative Frequency(f)	Observed Frequency(O)	Expected Frequency(E)	(O-E)	$(O-E)^2$	$\frac{(O-E)^2}{E}$
ICMP	0.243292159 789392	155.83333333 3333	181.939283582947	26.1059 502496138	681.520 63843531	3.745868 53929542
IPV4	7.486455040 13539E04	0.4666666666 66667	0.55985374446808	-9.318707 78014138E02	8.6838314691 6674E03	1.55108928 983181E02
TCP	0.755959194 706594	591.52222222 222	565.323084894807	26.19913732 74152	686.394796 70076	1.21416374 997048
TOTAL	**1**	**747.822222222 2222**				**Total= 160.0689612**

Table 6. Frequency Table

Categories	Relative Frequency(f)	Observed Frequency(O)	Expected Frequency(E)	(O-E)	$(O-E)^2$	$\frac{(O-E)^2}{E}$
ICMP	0.262010842676481	164.194444444444	176.245960 173713	-12.05151 57292687	145.2390313 72812	0.824071303 43187
IPV4	5.83755880097066E04	0.5444444444 44444	0.39267312 201196	0.1517713 22432485	2.303453431 29052E02	5.866083778 508
TCP	0.737405401443422	507.927777777778	496.028033370942	11.899744 4068362	141.6039169 48029	0.285475633 273603
TOTAL	**1**	**672.666666666 667**				**Total= 6.975630715**

Table 7. Total

$\sum \frac{(O-E)^2}{E}$
132678.2482
160.0689612
6.975630715
Total value=132845.2929

Table 7 shows the goodness of chi-square value is **132845.2929.**

Assumptions are performed on 5% level of significance. *i.e.*, α=0.05. In our work, there are 3 different categories are TCP, ICMP and IPV4. So, Degree of freedom is m-1, here in this case Degree of freedom is 2. Now, check the tabular values with the calculated chi-square values. Compare these values with each other.

Chi-squareX^2 table value for 0.05= **5.991.**

Chi-square calculated value= **132845.2929.**

Therefore, after finding the tabulated value and calculated value of chi-square approach, tabulated value is less than calculated X^2value, so we accept the hypothesis H_a. Anomaly occurs in interval 0 to 180(T1). From Literature survey, it has been seen that, if observed and expected difference is large, then intrusion occurs" [15]. Further, check other intervals and find in which interval, we get anomalous behavior.

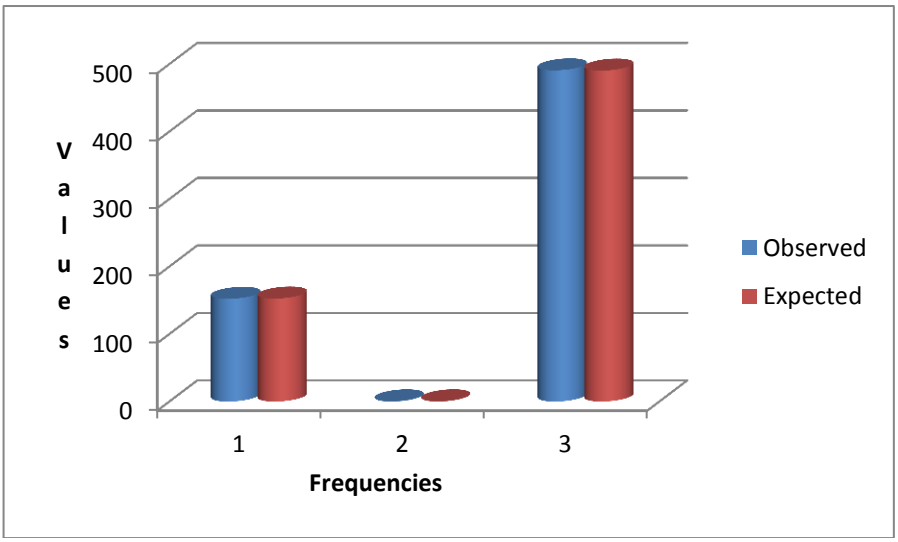

Figure 1. Graph of Frequency Table IV

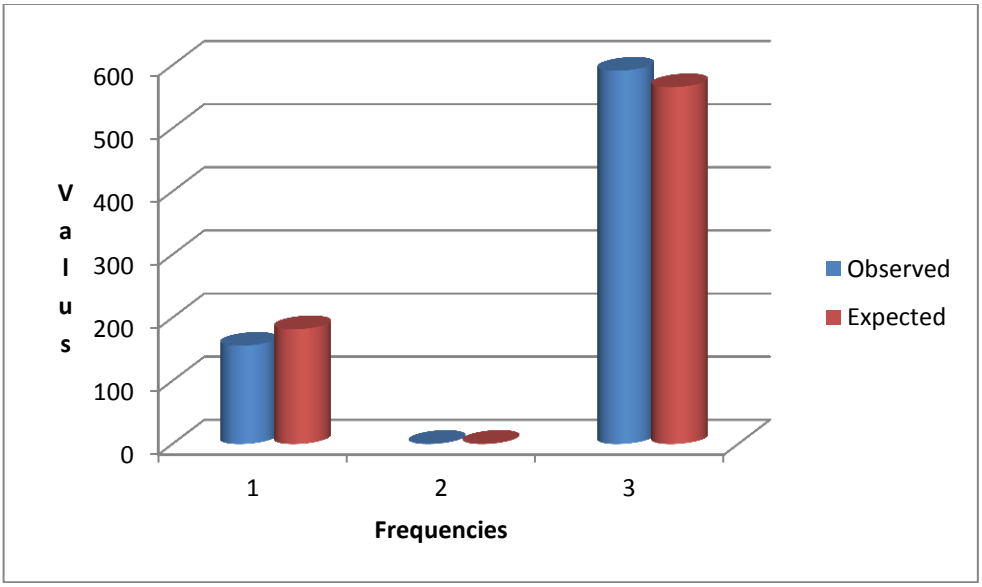

Figure 2. Graph of Frequency Table V

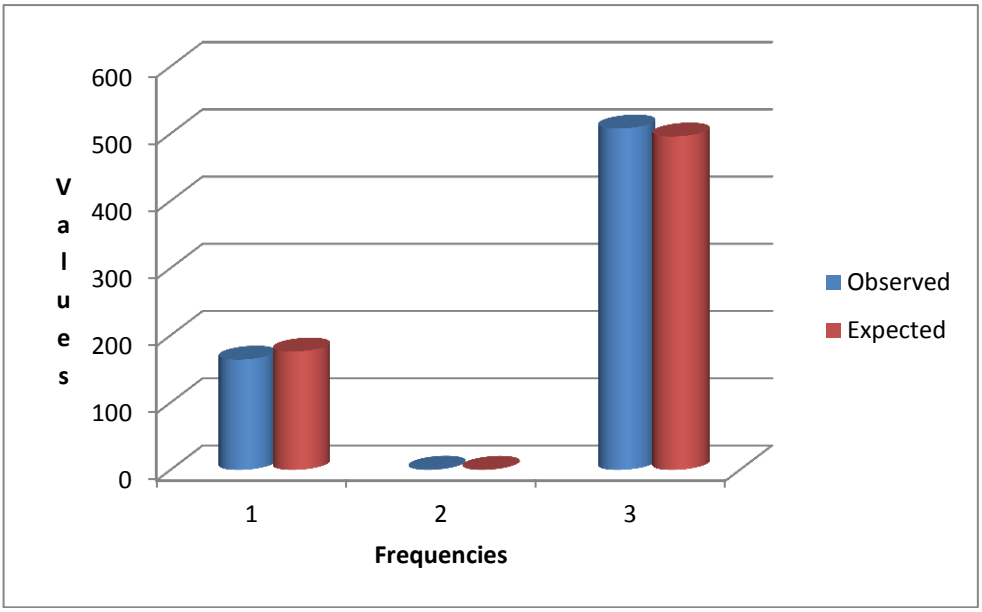

Figure 3. Graph of Frequency Table VI

6. Conclusion

In this paper, we showed the Statistical technique for the detection of intrusions. In proposed approach, we presented the Chi-square test to identify the detection of anomalies based on certain assumptions. Our Experiment shows the effectives of Chi-square approach based on Dataset CAIDA backscatter2007. Results and analysis shows that large difference between the observed and expected frequencies describes there is anomaly. In our results, shows that calculated value is much larger than the tabulated value. Chi-square is used to detect the intrusions. If calculated value is greater than table

value, then anomaly detect. The future work on this research is that to work on another statistical technique like annova test, P and Z test and compare with chi-square test.

References

[1] A. Derhab and A. Bouras, "Multivariate correlation analysis and geometric linear similarity for real-time intrusion detection systems", *Security and Communication Networks*, vol. 8, no. 7, **(2015)**, pp. 1193-1212.
[2] N. Muraleedharan, A. Parmar and M. Kumar, "A flow based anomaly detection system using chi-square technique", In *Advance Computing Conference (IACC), IEEE 2nd International*, **(2010)**, pp. 285-289.
[3] P. Garcia-Teodoro, J. Diaz-Verdejo, G. Maciá-Fernández and E. Vázquez, "Anomaly-based network intrusion detection: Techniques, systems and challenges", *computers & security*, vol. 28, no. 1, **(2009)**, pp. 18-28.
[4] V. Chandola, "Anomaly Detection for Discrete Sequences: A Survey Varun Chandola, Arindam Banerjee and Vipin Kumar", **(2009)**.
[5] A. Qayyum, M. H. Islam and M. Jamil, "Taxonomy of statistical based anomaly detection techniques for intrusion detection", *Emerging Technologies, 2005. Proceedings of the IEEE*, **(2005)**.
[6] N. Ye, Q. Chen, S. M. Emran and K. Noh, "Chi-square statistical profiling for anomaly detection", In *IEEE Systems, Man, and Cybernetics Information Assurance and Security Workshop June 6-7, 2000 at West Point, New York*, **(2000)**, pp. 187-193.
[7] S. M. Emran and N. Ye, "Robustness of Chi-square and Canberra distance metrics for computer intrusion detection", *Quality and Reliability Engineering International*, vol. 18, no. 1, **(2002)**, pp. 19-28.
[8] N. Ye, C. M. Borror and D. Parmar, "Scalable chi-square distance versus conventional statistical distance for process monitoring with uncorrelated data variables", *Quality and Reliability Engineering International*, vol. 19, no. 6, **(2003)**, pp. 505-515.
[9] J. Santiago-Paz, D. Torres-Roman and P. Velarde-Alvarado, "Detecting anomalies in network traffic using Entropy and Mahalanobis distance", In *Electrical Communications and Computers (CONIELECOMP)*, IEEE *2nd International Conference on*, **(2012)**, pp. 86-91.
[10] D. Bayarjargal, and G. Cho. "Detecting an Anomalous Traffic Attack Area based on Entropy Distribution and Mahalanobis Distance", *International Journal of Security and Its Applications*, vol. 8, no. 2, **(2014)**, pp. 87-94.
[11] Khan, Rahul Rastogi1 Zubair Khan2 MH, "Network Anomalies Detection Using Statistical Technique: A Chi-Square approach", **(2012)**.
[12] W. Lee and S. J. Stolfo, "Data mining approaches for intrusion detection", In *Usenix security*, **(1998)**.
[13] M. Markou and S. Singh, "Novelty detection: a review—part 1: statistical approaches", *Signal processing*, vol. 83, no. 12, **(2003)**, pp. 2481-2497.
[14] W. Fan, M. Miller, S. Stolfo, W. Lee and P. Chan, "Using artificial anomalies to detect unknown and known network intrusions", *Knowledge and Information Systems*, vol. 6, no. 5, **(2004)**, pp. 507-527.
[15] R. Jensen and Q. Shen, "A rough set-aided system for sorting WWW bookmarks", In *Web Intelligence: Research and Development*, Springer Berlin Heidelberg, **(2001)**, pp. 95-105.
[16] A. M. Bahaa-Eldin, "Time series analysis based models for network abnormal traffic detection", In *Computer Engineering & Systems (ICCES), 2011 International Conference on* IEEE, **(2011)**, pp. 64-70.
[17] C. Gu, S. Zhang and H. Lu, "Online internet intrusion detection based on flow statistical characteristics", In *International Conference on Knowledge Science, Engineering and Management*, Springer Berlin Heidelberg, **(2011)**, pp. 160-170.

