Detecting Outlier on Intrusion Detection System Using Data Set of System Calls

Dahliyusmanto Dahlan, Abdul Hanan Abdullah, Marina Md. Arshad

Network Security (NetSecure) Group

Department of Computer System & Communication

Faculty of Computer Science & Information System

Universiti Teknologi Malaysia

Skudai, Johor Bahru 81310, Malaysia

yoes@siswa.utm.my, hanan@fsksm.utm.my , marina@fsksm.utm.my

Abstract

An Intrusion Detection System (IDS) seeks to identify unauthorized access to computer systems' resources and data using statistical approach. We present a method for detecting unusual observations that do not seem to belong to the pattern of variability produced by the other observations. Outliers are best detected visually whenever this is possible. Usually, the original data sets are not normally distributed. If normality is not a viable assumption, one alternative is to make non-normal data looks normal. This paper explains the steps for detecting outliers' data and describes the Box-Cox power transformation method that transforms them to normality.

Keywords: Intrusion detection, Outlier, Box-Cox transformation, System call

1 Introduction

The methodology of intrusion detection can be divided into two-category [1]: anomaly intrusion detection and misuse intrusion detection. Anomaly intrusion detection refers to detecting intrusion based on anomalous behavior of the attackers. Therefore, the distinction by categorizing the good or acceptable behavior is very important. In the anomaly detection method, a statistical approach [2] and neural net approach [3] are usually taken to detect intrusion attempts. In statistical approach, data sets gained from detection results are used. Further, the data set should be calculated and analyzed. When analyzing data, we will sometimes find that one value which is far from the others. Such a value is called an "outlier". Given a mean and standard deviation, a statistical distribution expects data points to fall within a specific range. Many researchers have used statistical data analysis [4, 5]. Outliers typically are attributable to one of the following causes [6]; (1) the measurement is observed, recorded, or entered into the computer incorrectly, (2) the measurements come from a different population, and (3) the measurement is correct, but represents a rare event. Sometimes, when we encounter an outlier, we may be tempted to delete it from the analyses. One possibility is that the outlier happened by chance. In this case, we should keep the values in our analyses. The value came from the same population as the other values, so should be included [7].

The paper is organized as follow. To motivate our work, section 2 presents the counting of the number of system calls; section 3 describes the steps for detecting outlier data and Box-Cox power transformation. Section 4 mainly describes standardizes value and generalized square distances. Section 5 explains transformation to near normality. Lastly, section 6 is the conclusion.

1.1 Intrusion Detection

Research and applications of intrusion detection techniques has resulted in its classification into two-category [1]*: misuse intrusion detection* and *anomaly intrusion detection*.

Misuse Intrusion Detection seeks to discover intrusions by precisely defining them ahead of time and watching for their occurrence. For example, many well known attacks can be discovered by searching for distinguished patterns or events in the audit trails. The main shortcoming of misuse detection is that future attacks cannot be predicted or detected without hard-coding them into the IDS attack database.

Anomaly Intrusion Detection is based on the assumption that misuse or intrusive behavior deviates from normal system use. In general, most anomaly detection systems learn a normal system activity profile, and then flag all system events that statistically deviate from this established profile. The strengths of anomaly detection is the ability to abstract information about the normal behavior of a system and detect attacks regardless of whether or not the system has seen them before and the anomaly detection method can also detect unknown intrusions. Most behavior models are built using metrics that are derived from system measures such as CPU usage, memory usage, number and time of login, network activity, etc. However, it creates very large overhead for the host machine, which must have the capacity to record all users' activities to create users' profiles based on the define measure for intrusions. The main weakness of anomaly detection system is their vulnerability to an intruder who breaches the system during their learning phase. A savvy intruder can gradually train the anomaly detector to interpret intrusive events as normal system behavior. Recent, research projects have addressed a new type of anomaly detection method, which does not monitor all user activities. This approach is effective and lightens the load of monitoring.

1.2 Approaches to Intrusion Detection

There are many approaches to detecting intrusion detection one of them is Statistical Anomaly Detection which uses statistical analysis to measure variation in the amount and type of audit data. There are two techniques in statistical anomaly detection:

Threshold detection: Each occurrence of a specific event is recorded. The idea is that a usually high number of occurrences within a trying period may indicate the presence of an intruder. But the difficulty with this technique is to identify the threshold number. Interval time analysis employs this technique in their system.

Top

Profile-Based: This technique uses statistical measures to identify expected behavior of users or user groups. Masquerader and misfeasor can be detected by monitoring a system's audit log for user activity that deviates from established patterns of usage.

Rule-Based Anomaly Detection: This technique shares the same advantage and disadvantages with the statistical anomaly detection techniques. The major difference between rule-base and statistical anomaly detection is that rule base anomaly detection uses a set of rules to represent and store the usage pattern, whereas statistical anomaly detection uses statistical formulas to identify usage patterns in audit data.

Rule-Based Penetration Identification: These are characterized by their expert system properties that fire rules when audit information indicates illegal activities. Most of the current intrusion detection supplements their anomaly detection components with rule-based expert system components.

2 Counting the Number of System Call

Computer immune system research's group has collected several data sets of system calls executed by active process. They have data for the same program from multiple locations and/or multiple versions of the program. Each of these is a distinct data set; normal traces from one set can be quite different from those of another. Intrusions collected at one location or with a certain version of the program should not be compared to normal data from a different set. Each trace file (*.int) lists pairs of numbers, one pair per line. The first number in a pair is the PID of the executing process, and the second is a number representing the system call. For example we show a portion of the system calls which have been extracted from an .int file (Table 1).

Table 1: Strace summary output on ftp normal activities.

PID	Occurrence	Syscall
2623	53	read
2827	73	open
3005	20	write
3225	6	getuid32

In table 1, *read* occurs 53 times; *open* occurs 73 times, *write* occurs 20 times, and *getuid32* occurs 6 times. In this manner, we count the number of system calls occurring in each sample.

In our research, to detect outlier data on unusual observations, we used the system calls data set as variables for analysis. We need to extract these data to obtain the number of system call, number of processes, and to identify its characteristic. These are included as the result of programs that run as daemons and console programs (Table 2).

Table 2: Kinds of program running on daemons and console programs.

No	Daemon Program	Description
1	ftp	File transfer protocol: A program allows a user to transfer file to and from a remote network site.
2	named	A Unix background process that converts hostname to Internet addresses for the TCP/IP protocol.
3	inetd	A program that listens for connection requests or messages for certain ports and starts server programs to perform the services associated with those ports.
No	**Console Program**	**Description**
1	ps	Report process status: gives a snapshot of the current process (daemon program)
2	xlock	A program that allows a users to lock an X terminal
3	login	To start a session with a system, usually by giving a user name and password as a means of user authentication

3 Detecting Outlier & Transformation

3.1 Outline of Detecting Outlier and Cleaning Data

Original data sets contain one or a few unusual observations that do not seem to belong to the pattern of variability produced by the other observations [8]. Outliers occur when the relative frequency distribution of the data set is extremely skewed, because such a distribution of the data set has a tendency to include extremely large or small observations. The situation can be more complicated with multivariate data. Before addressing the issue of identifying these outliers, it must be emphasized that not all outliers are wrong numbers. Furthermore, finding outliers is an important task for many knowledge discoveries in database applications [9]. Outliers can also be detected spatially. A spatial outlier is a spatially referenced object whose non-spatial attribute values are significantly different from the value of its neighborhood. The algorithm for spatial outlier detection has been proposed by Tien Lu *et al.* [10]. Their approaches not only can detect false spatial outliers but also find true spatial outliers ignored by existing methods.

The outliers' data should be represented visually whenever this is possible. When the number of *n* observations is large, dot plots are not feasible. If the number of characteristic *p* is large, the large number of scatter plots may prevent viewing them.

There are four steps in detecting outlier, that are;

i) Make a dot plot for each variable.
ii) Make a scatter plot for each pair of variables.
iii) Calculate the standardized values, and examine these standardized values for large or small values.
iv) Calculate the generalized squared distances, and examine these square distances for unusually large values. In a chi-square plot, these would be points farthest from the origin.

In step 3, "large" must be interpreted relative to the sample size and number of variables and in step 4, "large" is measured by an appropriate percentile of the chi-square distribution with *p* degrees of freedom.

3.2 Box-Cox Power Transformation

If normality is not a viable assumption, one of the alternatives is to make non-normal data more *normal looking* by considering transformations of the data. Normal-theory analyses can then be carried out with the suitably transformed data.

Transformations are nothing more than a reexpression of the data in different units. Appropriate transformations are suggested by (1) theoretical considerations or (2) the data themselves (or both).

In many instances, the choice of a transformation to improve the approximation to normality is not obvious. For such cases, it is convenient to let the data suggest a transformation. A useful family of transformation is the family of *power transformations*.

Power transformations are defined only for positive variables. Box and Cox consider the slightly modified family of power transformations.

$$x^{(\lambda)} = \begin{cases} x^{\lambda} - 1 / \lambda & \lambda \neq 0 \\ \ln x & \lambda = 0 \end{cases} \qquad (1)$$

where x represents an arbitrary observation and λ as parameter of power family of transformation indexed.

To accommodate negative observations, Box and Cox further proposed the following shifted power transformations

Top

$$p\ \lambda_1, \lambda_2\ (X) = \begin{cases} \dfrac{(X + \lambda_1)^{\lambda_2 - 1}}{\lambda_2}, & \lambda_2 \neq 0 \\ \log\ (X + \lambda_1)\ , & \lambda_2 = 0 \end{cases} \quad for \quad X + \lambda_1 > 0 \qquad (2)$$

It is sufficient to substitute a convenient value for λ_1 and use the Box-Cox transformation on the shifted value of $X + \lambda_1$. If X has minimum, the $X + \lambda_1$ can be made positive by the proper choice of λ_1.

4 Calculating Standardizes Value and Generalized Square Distances

Dataset of system calls that are presented are executed by active processes collected by Stephanie Forrest research's group. We have selected 9 system calls as variables that are correlated to intrusions [11, 12] and the numerical data summaries as shown in Table 3.

Table 3: Numerical summaries data sets.

System Call	Variable	N	Sum	Mean 1.0e+003	Std Deviation 1.0e+004	Maximum	Minimum
chdir	x_1	49	2668	0.054	0.023	1517	1
geteuid	x_2	49	1604	0.033	0.021	1448	1
open	x_3	49	20033	0.409	0.116	7418	1
read	x_4	49	135335	2.762	1.031	71411	1
setgid	x_5	49	135	0.003	0.000	21	1
setuid	x_6	49	299	0.006	0.001	62	1
exit	x_7	49	4919	0.100	0.069	4814	1
getpgrp	x_8	49	2664	0.054	0.037	2599	1
unlink	x_9	49	107	0.002	0.006	29	1

The data in table 3 are the characteristic of system calls triggered from 49 normal and intrusive activities. Each of them is called as x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8 and x_9 variables, respectively. Also included in the table are sum of system call, mean, standard deviation, maximum and minimum system calls are resulted.

As a first step, it would be wise to carry out a preliminary analysis of the data. In this research we follow the steps for detecting outlier data. Firstly, we will make a dot plot for each variable (Figure 1).

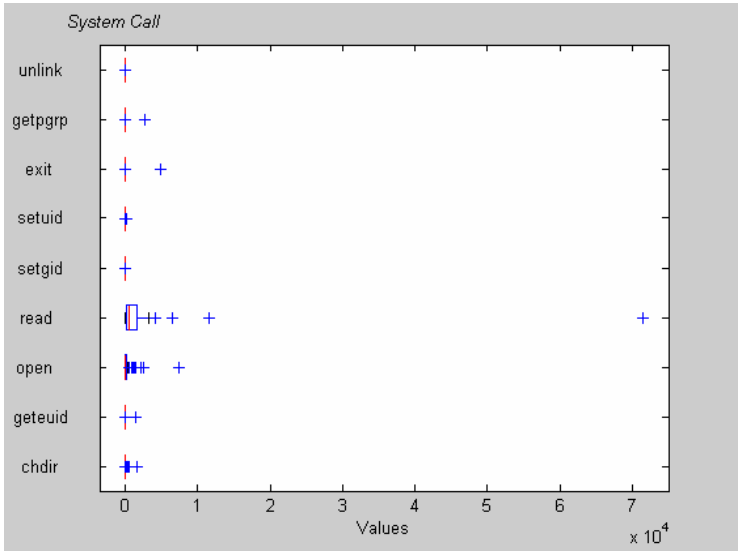


Figure 1: Box plot of system call variables.

The second step, we construct the scatter plot matrix for each pair of variables (Figure 2). We must order the distance of each observations from smallest to largest as $d^2_{(1)} < d^2_{(2)} < ... < d^2_{(n)}$ as shown in table 4, where as;

$$d^2 = (x - \bar{x}) \, . \, S^{-1} \, (x - \bar{x})' \qquad (3)$$

where $\bar{x}$ is the mean of the group, and S^{-1} is the inverse matrix of the covariance matrix of the group.

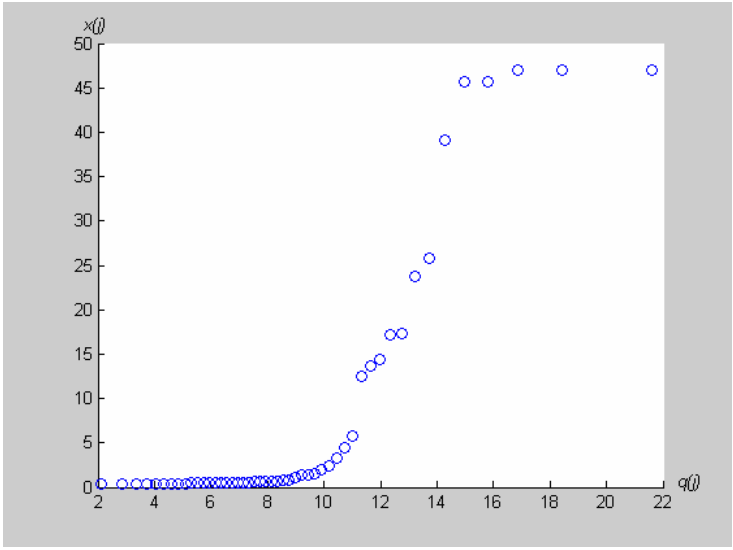


Figure 2: A scatter plot for the system calls data.

Top

The next step for detecting outlier data is to calculate the standardized values z on each column and examines these standardized values for large or small values. Let s be the standard deviation:

$$z_{jk} = x_{jk} - \bar{x}_k \,/\, \sqrt{s_{kk}} \qquad (4)$$
$$k = 1, 2, ..., 9 \quad j = 1, 2, \ldots, 49$$

In this step, "large" must be interpreted relative to the sample size and number of variables. There are n x p standardized values. When $n = 49$ and $p = 9$, there are 441 values.

The last step is to calculate the generalized square distances (see equation 3) and examine these distances for unusually large values (Table 4). In a chi-square plot, the dataset with outliers would be the points farthest from the origin.

Table 4: The dataset with outliers.

No.	Activity	Squared Distance
1	login1	39.086
2	ftp	46.944
3	xlock17	23.763
4	intend-i	46.940
5	login1-i	45.639
6	ps2	25.748
7	xlock1-i	47.020
8	ftp-i	45.655
9	login2	14.468
10	ps1	13.675
11	xlock13	17.257
12	xlock16	17.178
13	login2-i	12.560

For this reasons, "large" is measured by an appropriate percentile of the chi-square distribution with p degrees of freedom. If the sample size is n = 100, we would expect 5 observations to have values of the square distances that exceed the upper fifth percentile of the chi-square distribution. A more extreme percentile must be used to determine observations that do not fit the pattern of the remaining data.

5 Transformations to Near Normality

To select a power transformation, an investigator looks at the marginal dot diagram or histogram and decides whether large values have to be "pulled" or "pushed out" to improve the symmetry about the mean. The final choice should always be examined by a Q-Q plot or other to see whether the tentative normal assumption is satisfactory.

Top

A convenient analytical method is available for choosing a Box-Cox power transformation (equation 1). Which is continuous in λ for $x > 0$? Given the observations $x_{1,}\ x_{2,}\ . . .,\ x_{n,}$ the Box-Cox solution for the choice of an appropriate power λ is the solution that maximizes the expression. With multivariate observations, a power transformation must be selected for each of the variables. Let $\lambda_{1,}\ \lambda_{2,}\ \ldots\ ,\ \lambda_p$ be the power transformations for the p measured characteristics. Each λ_k can be selected maximizing:

$$\ell_k\ (\lambda) = -\ n/2\ \ln\left[1/n \sum_{j=1}^{n} (x_{jk}^{(\lambda_k)} - x_k^{(\bar{\lambda}_k)})^2\right] + (\ \lambda_k\ -1\ ) \sum_{j=1}^{n} \ln x_{jk} \tag{5}$$

where $x_{1k,}\ x_{1k,}\ x_{nk}$ are the n observations on the kth variable, $k = 1, 2, p$. Here

$$x_k^{(\bar{\lambda}_k)} = 1/n \sum_{j=1}^{n} x_{jk}^{(\lambda_k)} = 1/n \sum (x_{jk}^{\lambda_k} - 1)/\lambda_k \tag{6}$$

is the arithmetic average of the transformed observations. Next, the jth transformed multivariate observation is

$$x_j^{(\hat{\lambda})} = \begin{bmatrix} x_{j1}^{(\hat{\lambda}_1)} - 1\ /\ \hat{\lambda}_1 \\ x_{j2}^{(\hat{\lambda}_2)} - 1\ /\ \hat{\lambda}_2 \\ \vdots \\ x_{jp}^{(\hat{\lambda}_p)} - 1\ /\ \hat{\lambda}_p \end{bmatrix} \tag{7}$$

where $\hat{\lambda}_1,\ \hat{\lambda}_2, \ldots,\ \hat{\lambda}_p$ are the value that individually maximize equation 5.

In our case, we gave readings of the data set of system calls through *chdir*, *geteuid*, *open*, *read*, *setgid*, *setuid exit, getpgrp* and *unlink* system call of $n = 49$ observations. The Scatter plot of these data in figure 2 shows that the observations deviate from what would be expected if they were normally distributed. Since all the observations are positive, let us perform a power transformation of the data which, we hope will produce results that are more nearly normal. Restricting our attention to the family of transformation in (see equation 1), we must find the value of λ that maximizes the function $\ell_k\ (\lambda)$ in (see equation 5). The pairs $(\ \lambda, \ell_k\ (\lambda))$ are listed in the following Figure 3 for several value of λ.

Top

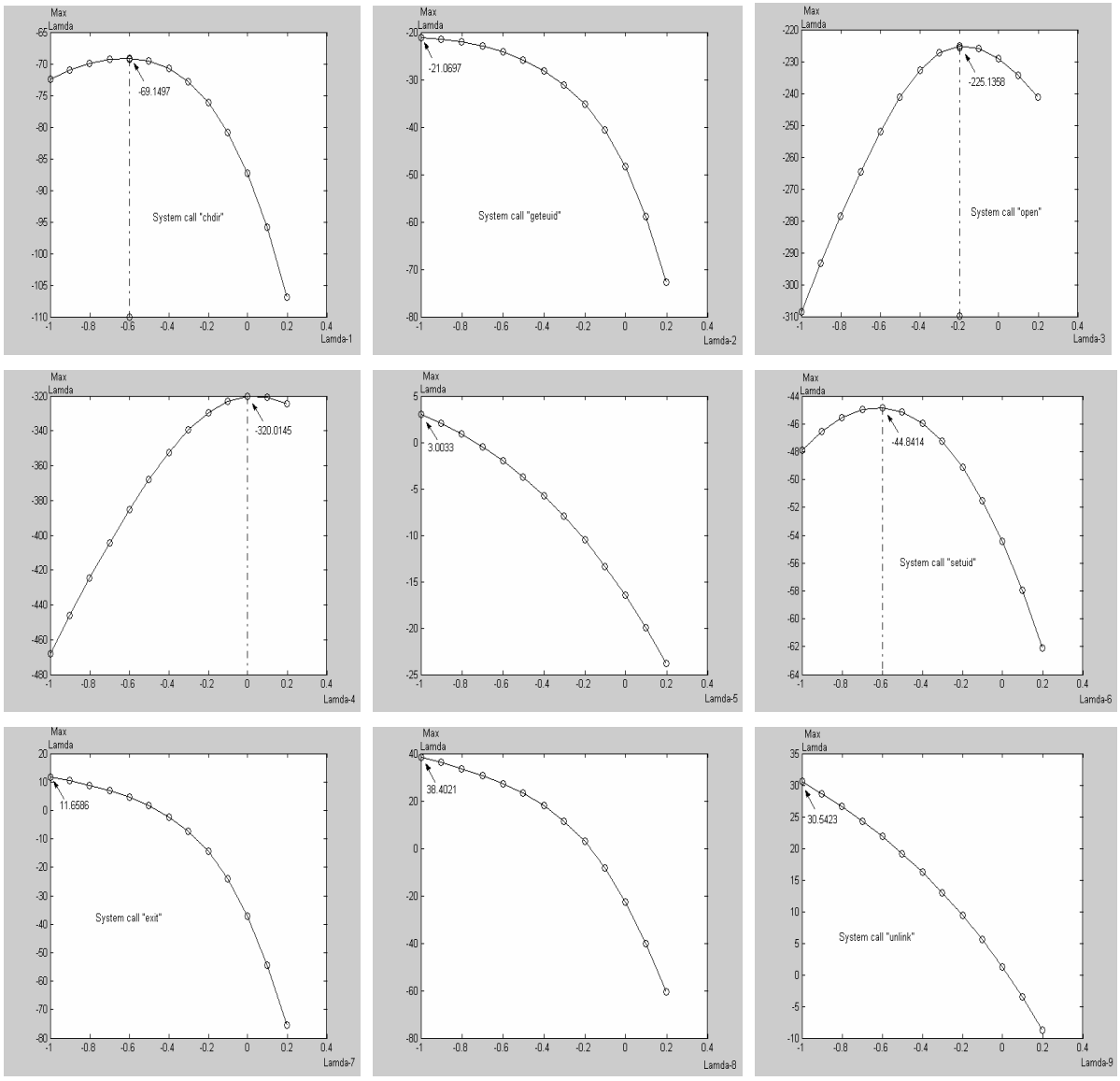
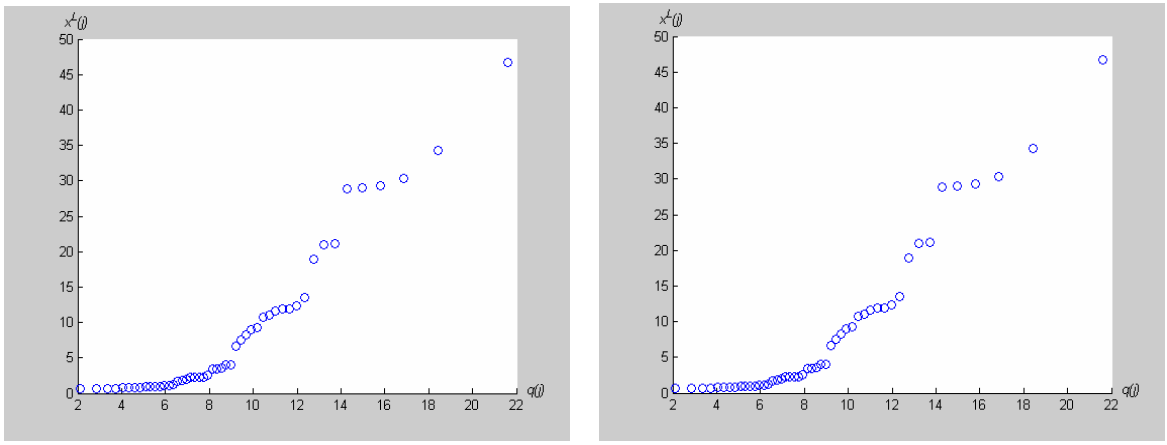


Figure 3: Plot of (λ, $\ell_k(\lambda)$) of system call variables.

For convenient, we chose $\lambda_1 = -0.6$, $\lambda_2 = -1.0$, $\lambda_3 = -0.2$, $\lambda_4 = 0.0$, $\lambda_5 = -1.0$, $\lambda_6 = -0.6$, $\lambda_7 = -1.0$, $\lambda_8 = -1.0$, and $\lambda_9 = -1.0$. Using equation (7) with n = 49, the data x_j were reexpressed as a Scatter plot as shown in Figure 4(b).

Top



(a) (b)

Figure 4: Scatter plot of (a) the original and (b) the transformed system call data set.

Pairs fall very close to a straight line, and we would conclude from this evidence that the $x_j^{-0.2}$, $x_j^{-0.4}$, x_j^{0}, $x_j^{-0.1}$, $x_j^{-0.5}$, $x_j^{-0.1}$, $x_j^{-0.1}$, $x_j^{-0.1}$ and $x_j^{-0.2}$ are approximately normal.

6 Conclusions

There are several methods for detecting outliers. All the methods first quantify how far the outlier is from the other values. This can be the difference between the outlier and the mean of all points, the difference between the outlier and the mean of the remaining values, or the difference between the outlier and the next closest value. Next, standardize this value by dividing it with some measure of scatter, such as the SD of all values, the SD of the remaining values, or the range of the data.

In identifying the outliers' data, we must emphasize that not all outliers are wrong numbers. They may, justifiably, be part of the group and may lead to a better understanding of the phenomena being studied. The outliers can be transformed to near normality by using Box-Cox power transformation.

7 References

[1] Kumar, S. and Spafford, E., (1994), An Application of Pattern Matching in Intrusion Detection, Technical Report 94-013, Purdue University Department of Computer Science.

[2] Alexis, C, (2004), Algorithm-Based Approach to Intrusion Detection & Response. SANS Institute.

[3] Hussam, O. M., (2002), A Survey and Analysis of Neural Network Approaches to Intrusion Detection. SANS Institute.

Top

[4] Billof, N., Hadi, A. S., and Vallemen, P. F., (2000), Bacon: Blocked adaptive computationally efficient outlier nominators, *Computational Statistic & Data Analysis*, 34(3):279-298.

[5] Rousseeuw, P. J., and Drissen, K. V., (1999), A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*, 41:212-223.

[6] Read, R. J., (1999), Detecting Outliers in Non-Redundant Diffraction Data. *Acta. Cryst.* D55, 1759-1764.

[7] Motulsky, H., (2002), Detecting Outliers. GraphPad Software. http://www.graphpad.com/articles/outlier.htm

[8] Johnson, R. A., and Wichern, D.W., (1998), Applied multivariate statistical analysis, fourth ed., Prentice Hall, New Jersey.

[9] Breuning, M. M., Kriegel, H. P., and Raymond, T. Ng. (1999). OPTICS-OF: Identifying Local Outliers. Proceedings of the 3rd European Conference on Principles & Practice of Knowledge, Discovery in Database (PKDD'99), Progue.

[10] Tien Lu, C., Chen, D., and Kou, Y., (2003), Algorithm for Spatial Outlier Detection. Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM'03). Melbourne, Florida. USA.

[11] Asaka, M., Onabuta, T., Inoue, T., Okazawa, S., and Goto, S., (2001), A New Intrusion Detection Method Based on Discriminant Analysis, Journal of IEICE Trans. On Information & Ssytems, Vol.E-84-D NO.5, pp. 570-577.

[12] Intersect Alliance. (2002), System Intrusion Analysis and Reporting Environment, http://www.intersectalliance.com/projects/Snare/Documentation/index.html

Top