Anomaly Detection from System Logs based on Spectral Method

Sagar Dhakal
Head of Technical Department
LogPoint Nepal
Steel Tower, Lalitpur, Nepal
Email: sagar.dhakal@logpoint.com

Umanga Bista
Research Engineer
LogPoint Nepal
Steel Tower, Lalitpur, Nepal
Email: umanga.bista@logpoint.com

Basanta Joshi
Assistant Professor
DOECE, Institute of Engineering
Pulchowk, Lalitpur, Nepal
Email: basanta@ioe.edu.np

***Abstract*—This research provides a statistical method for detecting anomaly using multivariate data from system metrics. The spectral anomaly detection technique has been implemented using PCA (Principal Component Analysis) as unsupervised method to detect anomaly from data consisting of both normal and abnormal events. A comparison has been done with another unsupervised algorithms namely K-means which shows PCA outperformed K-means by 30% when F-measure is used to evaluate for probe attacks for standard DARPA (Defense Advanced Research Projects Agency) dataset. The real data of the system has been collected and analyzed using unsupervised approaches for detecting anomalies. Also, the results of real data have been validated with visual plots. The result shows PCA as a better way to find the anomalous events and detect them precisely.**

Keywords*—*System, multivariate data, DARPA, Anomaly Detection, K-means, PCC, CART.

I. Introduction

The term anomaly-based intrusion detection in networks refers to the problem of finding exceptional patterns in network traffic that does not conform to the expected normal behavior. These non-conforming patterns are often referred to as anomalies, outliers, exceptions, aberrations, surprises, peculiarities or discordant observations in various application domains. In anomaly detection, the normal behavior of the system is modeled. Incoming patterns that deviate substantially from normal behavior are labeled as attacks.[1] The increase in number of networked machines has lead to an increase in unauthorized activity; not only from external attackers, but also from internal attackers, such as disgruntled employee and people abusing their privilege for personal gain.[2] Anomaly detection attempts to recognize behavior that does not conform to normal behavior. This technique is based on the detection of traffic anomalies. The anomaly detection systems are adaptive in nature; they can deal with new attack but they cannot identify the specific type of attack. An ideal IDS does not produce false or inappropriate alarms. In practice, signature based IDS is found to produce more false alarms than expected. This is due to the very general signatures and poor built in verification tool to authenticate the success of the attack. The large amount of false positives in the alert logs generates the course of taking corrective action for the true positives, i.e. delayed, successful attacks, and labor intensive.There are basically four types of remotely launched attacks: denial of service (DOS), U2R, R2L, and Probes.A probing is an attack in which the hacker scans a machine or a networking device in order to determine weaknesses or vulnerabilities that may later be exploited so as to negotiate the system. This practice is commonly used in data mining e.g. portsweep, nmap, imap, satan etc.[4]

Several research works have already been done and many research papers have been published regarding anomaly detection techniques. The research work done by Sami et. al [2] focuses on the analysis of server log data and the detection and potential prediction of anomalies related to the monitored servers . Chen et. al. [5] investigates a multivariate control technique to detect intrusions by building a long-term profile of normal activities in information systems and using the norm profile to detect anomalies. Syarif et. al.[11] described about misuse detection techniques and anomaly detection techniques. In case of misused detection techniques, the research implemented naive bayes, decision tree and nearest neighbor methods. For anomaly detection techniques, different clustering algorithms K-Means, K-Medoids, EM Clustering and distance-based outlier detection algorithm have been implemented [11]. Bentley [13] used nearest neighbors approach KDtree and mention about the structures attempt to reduce the required number of distance calculations by efficiently encoding aggregate distance information for the sample.

Even-though there are more works on anomaly detection and some research also talk about multivariate analysis with different approaches for process control, there is hardly any research, which properly use the statistical-based multivariate method for System. Survey has been done in different areas of log data but implementation is not done except basic attributes like packets in, packets out etc. Only attributes like CPU, memory and processes have been analyzed. Other attributes have not been considered, which could have significant impact on anomaly consideration. And, most of the researches done have used popular datasets like DARPA and KDDcup99 [14]. Research using real dataset is very less heard due to complexity attached. With these motivations the authors have implemented both supervised and unsupervised statistical methods for anomaly detection with large no of attributes as features and performance assessment has been done .

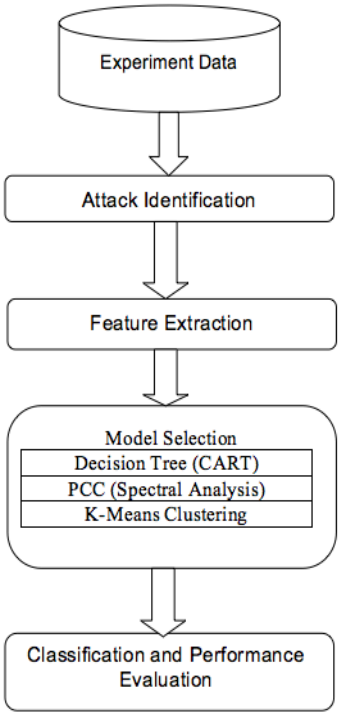


Fig. 1. Flowchart for Implementation Model

II. Methodology

Anomaly Detection method detect anomalous events that occurs across data across network and system. It takes data as input and with that processes it and generates anomalous events from data, which also consists of normal data. Unsupervised method of anomaly detection has been used. So, this research classified events as anomaly with proper model selection and its performance assessment has been done.

A. Implementation Model

Logs are collected from variety of sources and analyzed. The collection of data that has been considered for research is standard dataset that is DARPA and real dataset. In case of real dataset, a standard toolset, sar is used which serves to log and evaluate a variety of information regarding system activity. With performance problems, sar also permits retroactive analysis of the load values for various sub-systems (CPUs, memory, disks, interrupts, network interfaces,commits, faults etc). A linux server has been used to collect data where sar tool has been installed. In case of DARPA standard set, the first step i.e attack identification is to first determine the types of attacks e.g. Probe attacks(i.e.portsweep, ipsweep,nmap and satan). Model selection describes about model and the algorithm implemented for the research. The algorithm 1 has been described in detail in next section. For model, Spectral anomaly detection techniques using PCA and KMeans clustering has been used and analysed.

B. Feature Extraction

Feature Extraction is finding or processing the overall attributes and only extracting the important attributes out of them. E.g. for standard dataset; src_bytes, dst_bytes sport, dport, source ip, destination ip etc. Contents in the packets logged_in, su etc. Connection errors based on time same_source_ports probe attacks, rerror etc. E.g. For real dataset;memory used, CPU util, commit,faults etc.

C. Algorithm

For this research, using unsupervised model, two methods have been evaluated: K-means clustering and Spectral Anomaly detection.
Let $\mathbf{X} \in \mathbb{R}^{\mathbf{m*n}}$ denotes train dataset, where $\mathbf{m}$ is number of train-data and $\mathbf{n}$ is dimension of train data.
Notation $\mathbf{X_j^{(i)}}$ denotes $\mathbf{j^{th}}$ feature of the $\mathbf{i^{th}}$ example in our dataset.

K-Means Clustering: K-means clustering is very simple to implement yet effective in performance. Clustering is a form of unsupervised learning whereby a set of observations is partitioned into natural groupings or clusters of patterns in such a way that the measure of similarity between any pair of observations assigned to each cluster minimizes a specified cost function. K-means partition m observations into k clusters in which each observation belongs to the cluster with the nearest mean. Every learning problem is an optimization problem; there is the need to search for values of parameters in parameter space that minimizes some defined cost function. The distance used for K-means is Euclidean distance.
For given input dataset, encoder C is found that assigns observations in training set to the K clusters in such a way that, within each cluster, the average measure of dissimilarity of the assigned observations from the cluster centroidmean is minimized.

$$J(C) = \sum_{k=1}^{K} \sum_{C(i)==k} \|X^{(i)} - \mu_k\|_2$$

Algorithm 1 k -means

Input: X: train-dataset and K:Number of clusters
Initialize Randomly pick K datapoints and assign them to k-centroids $\mu_k \in \mathbb{R}^n$. i.e.

$$\mu_1, \mu_2, ..., \mu_K \leftarrow sample(X, K)$$

repeat
Cluster assignment

$$C(i) \leftarrow k : \arg\min_k \|X^{(i)} - \mu_k\|_2$$

Move Centroid

$$\mu_k \leftarrow \mu_k : \arg\min_{\mu_k} J(C),\ k = 1..K$$

i.e. $\mu_k \leftarrow \frac{1}{\sum_i C(i)==k} \sum_{C(i)==k} X^{(i)},\ k = 1..K$
until convergence

Spectral Anomaly detection: Spectral anomaly detection uses the principal component analysis as in algorithm 2 results to check outlierness of datapoints. The most naive version of spectral anomaly detection is to χ^2 test the sum of normalized projection error on all dimensions with significance α and degree of freedom equals to number of dimensions. The intuition behind this is that the normal data will conform with the correlation within the variables, and outlier will not. This method is described in algorithm 3.

The more improved version of this approach is to break

Algorithm 2 Principal Component Analysis

$\mathbf{\Sigma} \leftarrow \mathbf{X}^\mathsf{T}\mathbf{X}$
$\mathbf{Q}, \mathbf{\Lambda}, \mathbf{Q}^{-1} \leftarrow \mathbf{EigenValDecompose}(\mathbf{\Sigma})$
$\mathbf{Y} \leftarrow \mathbf{XQ}$, i.e. PCA without reducing dimensions.

where $\mathbf{Y} = [\mathbf{Y_1}, \mathbf{Y_2}, ..., \mathbf{Y_n}], \forall_\mathbf{j} \mathbf{Y_j} \in \mathbb{R}^\mathbf{m}$ representing the projections of $\mathbf{X}$ on eigen vectors $\mathbf{Q}$
and $\mathbf{\Lambda}$ is diagonal matrix of eigenvalues $\lambda_\mathbf{1}, \lambda_\mathbf{2}, ..., \lambda_\mathbf{n}$

Algorithm 3 Chi-Squared Test Based Classifier

for all $\mathbf{Y^{(i)}} \in \mathbf{Y}$ **do**
  $\mathbf{NormalizedProjError(i)} = \sum_{\mathbf{j=1}}^{\mathbf{n}} \frac{[\mathbf{Y_j^{(i)}}]^\mathbf{2}}{\lambda_\mathbf{j}}$
  if $\mathbf{NormalizedProjError(i)} > \chi^\mathbf{2}_{\mathbf{n},\alpha}$ **then**
    $\mathbf{anomaly(i)} \leftarrow \mathbf{TRUE}$
  else
    $\mathbf{anomaly(i)} \leftarrow \mathbf{FALSE}$
  end if
end for

Algorithm 4 Principal Component Classifier

Let $\mathbf{q}$ be number of components such that 50% variance is retained, and $\mathbf{r}$ be number of components that retains last 20% variance.

for all $\mathbf{Y_{(i)}} \in \mathbf{Y}$ **do**
  $\mathbf{major(i)} = \sum_{\mathbf{j=1}}^{\mathbf{q}} \frac{[\mathbf{Y_j^{(i)}}]^\mathbf{2}}{\lambda_\mathbf{j}}$
  $\mathbf{minor(i)} = \sum_{\mathbf{j=n-r+1}}^{\mathbf{n}} \frac{[\mathbf{Y_j^{(i)}}]^\mathbf{2}}{\lambda_\mathbf{j}}$
end for
Let $\mathbf{c_1}$ & $\mathbf{c_2}$ be the $\mathbf{k}$ quantile of distributions of variables $major$ & $minor$.

for all $\mathbf{Y^{(i)}} \in \mathbf{Y}$ **do**
  if $\mathbf{major(i)} > \mathbf{c_1}$ & $\mathbf{minor(i)} > \mathbf{c_2}$ **then**
    $\mathbf{anomaly(i)} \leftarrow \mathbf{TRUE}$
  else
    $\mathbf{anomaly(i)} \leftarrow \mathbf{FALSE}$
  end if
end for

the dimensional spaces into major and minor dimensions subjective to dataset[15]. The major dimensional subspace will consist of axes which cumulatively retains first 50% variance, and minor retains last 20% variance. Even in this approach, we will check normalized projection error, but separately along major and minor dimensions. A data-point is an outlier if the normalized projection error along both dimensional sub-spaces are extreme, which is determined by quantile method. This method is formally given in algorithm 4. Both Principal Component Classifier and χ^2 test based anomaly detection were tested on same set of significance, i.e α for χ^2 test, and $\mathbf{k}$ quantile for PCC.

D. Perfomance Evaluation

For performance evaluation, following measures are taken into consideration:

$$Precision(P) = \frac{TP}{TP + FP} \tag{1}$$

$$Recall(R) = \frac{TP}{TP + FN} \tag{2}$$

$$F - 1Measure(F_1) = \frac{2 * P * R}{P + R} \tag{3}$$

where, TP = True Positive, FP = False Positive, TN = True Negative, FN = False Negative

III. Experiments and discussions

R language is used as an experimentation tool. R provides a wide variety of statistical and graphical techniques, including linear and nonlinear modeling, classical statistical tests, time-series analysis, clustering and others. Along with R, python language is also used. Experiments have been conducted for both standard dataset and real dataset.

A. Standard Dataset Results

Under the sponsorship of Defense Advanced Research Projects Agency (DARPA) and Air Force Research Laboratory (AFRL), the MIT Lincoln laboratory has established network and captured the packets of different attack types and distributed the data sets for the evaluation of researches in computer network intrusion detection systems. The KDDCup99 data set is a subset of the DARPA benchmark data set. Each KDDCup99 training connection record contains 41 features and is labeled as either normal or an attack, with exactly one specific attack type. This Dataset has been taken as testing data for validation of the research work with different classifiers and methods.

1) Attack Identification: From KDD dataset, following probe attacks are considered for research:

- ipsweep: An Ipsweep attack is a surveillance sweep to determine which hosts are listening on a network..
- portsweep: Surveillance sweep through many ports to determine which services are supported on a single host.
- satan: Network probing tool, which looks for well-known weaknesses. Operates at three different levels. Level 0 is light
- nmap: Network mapping using the nmap tool. Mode of exploring network will varyoptions include SYN.

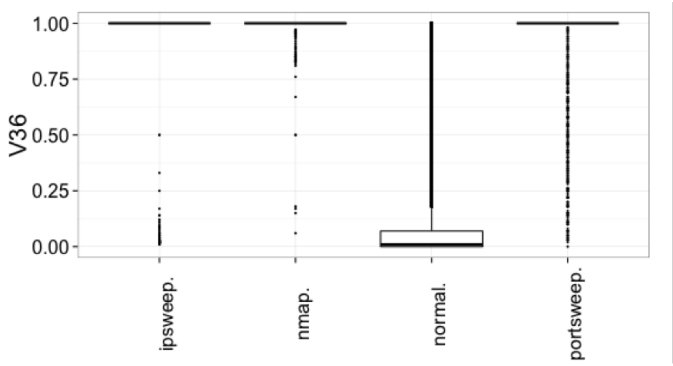


Fig. 2. Probe attacks with respect to dst_host_same_source_port

2) Feature Extraction (Attributes): Based on the attack requirement, following feature were extracted:

- Basic features: This group summarizes all the features that can be extracted from a TCP/IP connection. Feature used for research is src_bytes.
- Content features: These features are purely based on the contents in the data portion of the data packet. Feature used for research is logged_in.
- Traffic features: This group comprises features that are computed with respect to a 2 Sec. time window and it is divided into two groups: same host features and it service features. Some of the traffic features used in research are count, error_rate, diff_serv_rate,srv_count, and srv_rerror rate.

For different attacks, the attribute in consideration might be different. These variables are considered based on the distribution that were plotted using boxplot command provided by R. A plot for this is drawn in Figure 2 using boxplot was made with variable dst_host_same_source_port (V36) for probe attacks ipsweep,nmap,portsweep & satan.The outliers and the plot characteristics showed that all three probe attacks showed different characteristics than for normal. This means, this variable could be one of the attribute for anomaly in case of these attacks.

3) Classification and Performance Evaluation : Only probe attacks category has been analysed for research. A total of 14 attributes has been extracted out of 41 attributes for probe attacks. Results have been analysed using supervised approach Decision tree. Also unsupervised approaches-K-means and PCC have been used.

Using PCC,q(quantile value) and all evaluation measures were calculated. When the quantile value was increased,it was observed that precision was also increasing while recall was decreasing. The quantile value was tuned with values starting from 0.9 to 1 with quantile values varying to three decimal values. The optimal value was set where F-measure showed the best result. The choice of features was very dependent for unsupervised method; hence a lot of effort was required to tune the parameters. In this case there were 97278 normal data and 1040 anomalous data. There, α (significance value),

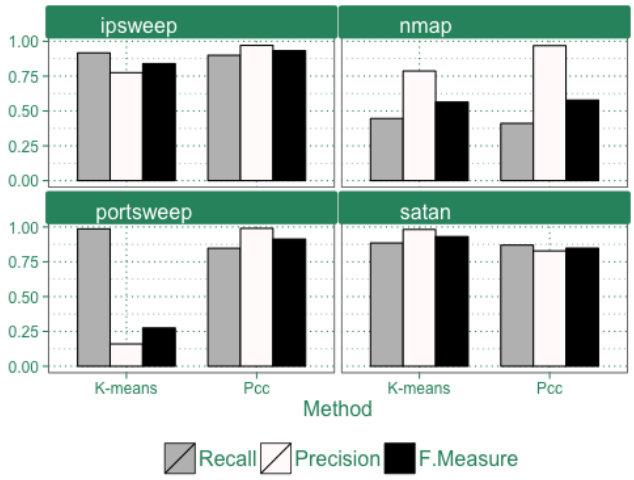


Fig. 3. Comparison Plot for probe attacks

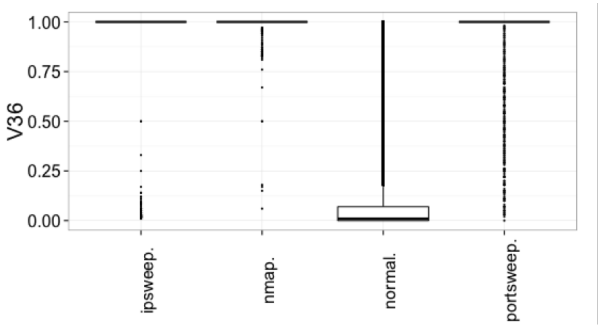


Fig. 4. Comparison Plot for combined probe attacks using different methods

q(number of major components) and r(number of minor components) were considered. When attacks were classified based on these parameters, we saw following result with the model. Significance value α=0.985 and q and r set as 0.5 and 0.8 respectively were used for the classification. After classification, results obtained:

Using K-means, the feature sets were extracted from KDD99 dataset. Based on number of principal components k, precision, recall and F-measure were computed. Using R language, subset of probe attacks from KDD99 dataset were defined. Subset were a collection of extracted features for research. From total dataset of KDD99, sampling was done and a total of 5000 data was analyzed. After classification, results obtained:

A comparison plot of probe attacks using K-means and PCC has been drawn as in Figure 3. For each attack type, all the evaluation measures i.e. Recall, Precision and F-measure were calculated. Only the optimum values of each attacks were considered.The obtained optimal values were 0.931, 0,933, 0,578 and 0.913 for satan, ipsweep, nmap and portsweep respectively. Along with individual probe attacks, a combined result of all probe attacks was considered for research as shown

in Figure 4. The statistical plots of probe attacks were created using K-means, chisq and PCC. The optimal values for all the three methods were 0.835, 0.575 and 0.569 for PCC, Chisq and K-means respectively.

B. Real Dataset Results

For real data, the collection was done for different attributes of the system. A development server with 24/7 operation and activities was monitored and data was collected. System Activity Report (SAR) was used as one of the tool for data collection.

1) Attack Identification: Identifying attack means finding anomaly in system, and confirming those anomaly as attacks based on the multivariables considered, and also the expected behavior of system. The target has been to find the anomaly from the data of system where there were normal and anomaly both events.

2) Feature Extraction (Attributes): This part was the most challenging part for a real data. With tools, they provide almost all the attributes of system. The task was to extract the useful attributes for further analysis. Total of 24 system attributes were considered. These values were extracted from SAR events and activity file with past one month data in live development server. Altogether around 4000 activities and events of system were collected. Attributes chosen covered paging faults, I/O read writes and utilization, network interface error, socket, queue length and load averages, memory utilization, cpu utilization, number of files handled and number of tasks created.

3) Classification and Performance Evaluation : For different load, memory and cpu utilization and other network attributes, the collection model was implemented with the system tool. The server where data was collected was 247 and there we had the possibility to stress the machine, generate out of memory exceptions, disk space full issues and other file system issues. Also, it had applications running resource intensive jobs. The server machine was running in a production environment. When events were classified following results were obtained:

Total Events: 3885
Normal Events: 3854
Anomaly Events: 31

Since, unlabeled dataset was used, comparison between two unsupervised methods was done for validation of algorithm. The distance measurement used was Euclidean distance. It was based on the assumption that anomaly data must have higher distance. Nearest neighbor has been used to visualize events as anomaly or normal and for validation.

For validation, KD-tree algorithm has been used with nearest neighbor for detecting anomaly. Based on that, PCC and K-Means algorithm have been implemented to label them as anomaly or normal. A visual representation of that has been the validation of this research.

In the Figure 5, the distributions of events that are of same nature were always in cluster. Circles showed the anomalous

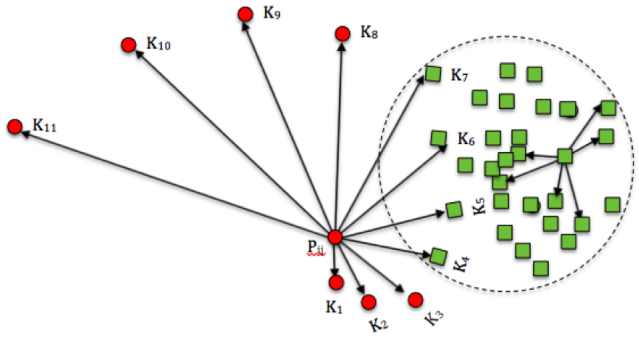
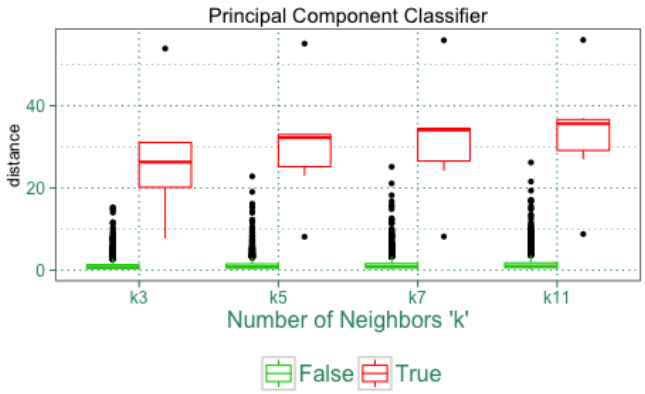


Fig. 5. Distribution of normal and anomaly events and neighbor distance



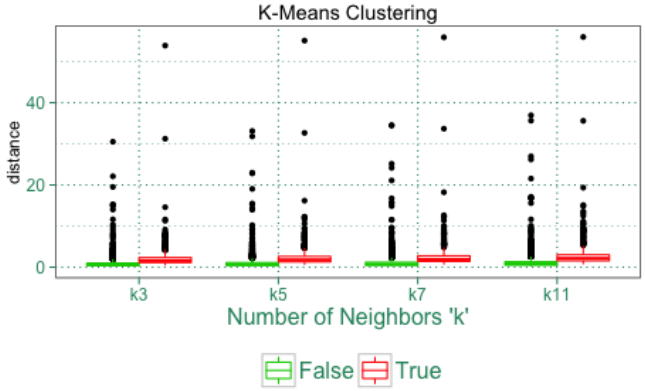


Fig. 6. Anomalies in real data using K-means and PCC

events while square represented the normal events and Pij as an anomalous event. It has been well verified with nearest neighbor. The square boxes termed as normal were in near distance with each other, whereas anomalous events were far from each other or normal event. Ki represented the distance from anomalous event to other anomalous or normal event.
As shown from the Figure 6, it can be clearly seen that PCC outperformed K-Means clustering in determining the anomalous events. Based on the number of neighbor k, it can be clearly seen that PCC anomalous events were distinct as opposed to K-Means, where distinction of anomalous events

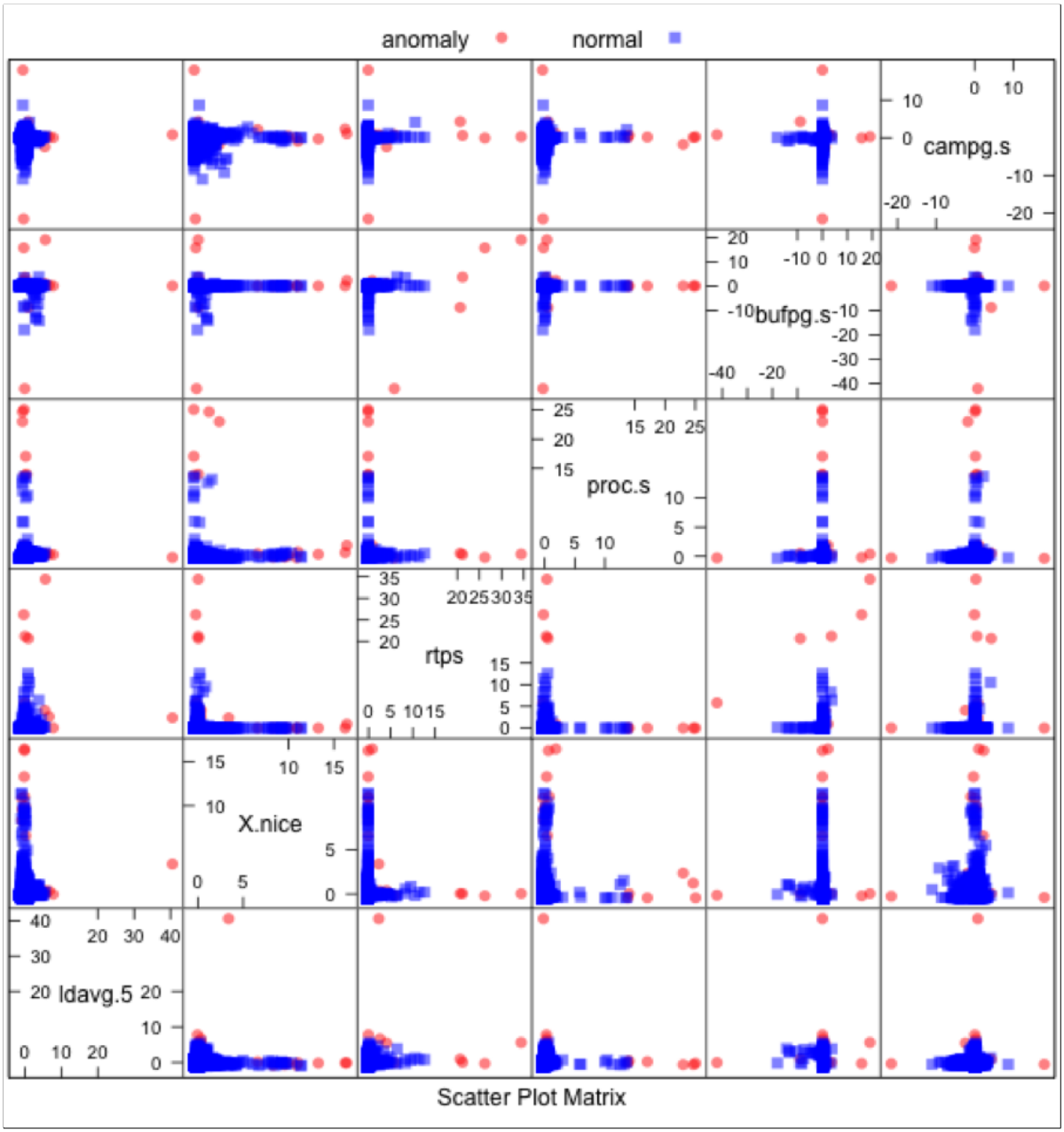


Fig. 7. Features Plot for real data showing normal/abnormal events

were rather in lower side. The plot had two components in X-axis, False and True. False represented the normal events whereas true represented the anomalous events. Y- axis represented the distance. The validation of unlabeled real data of system had been done visually. A feature plot using R where only important features were drawn. The reason for using only selected features or attributes was due to difficult it generated in visualization.Following features; ldavg.5, %nice,rtps,procs , bufpgs, camps were chosen as important features.

Features plot as in Figure 7 implemented the classified anomalous and normal events and plotted accordingly. With the defined attributes in X-labels and Y-labels, features plot was drawn where circles represented the anomalous events and square represented the normal events. Events at the extremes were anomalous as they were far from the normal events and were rarely in clusters. Some of the circles were seen also near normal events. The reason for this was due to limited attributes considered for plot and possible effect of other attributes. And, comparison was done based on any two attributes of interest. So, this also provided research with another validation of real data set that algorithm used to extract anomaly from normal and abnormal events.

IV. Conclusion

In this research, unsupervised method for anomaly detection has been used. For unsupervised method, a spectral anomaly detection technique has been used using PCC. In

case of PCC, parameters (significance value), q (number of major components) and r (number of minor components) has been tuned to provide the best result. Our experiment was done for unsupervised approach and it showed that algorithm P outperformed other algorithms with an F-measure of 84% for probe attacks, which has been tested with standard DARPA dataset. In case of K-means result has been around 57%.
The Model has been tested with real dataset. For this, system events with SAR as one of the tool from live server have been collected for month. For real dataset, there is no labeling of anomaly or attack. PCC algorithm and K-means have been used to find the outlier or classify the events as anomalous or normal. At first, these real data are analyzed to detect anomaly using nearest neighbor approach and then visualization is done. The result shows that this method has potential to be applied to real-time logs.

ACKNOWLEDGMENT

The authors are highly indebted to faculty members of Nepal College of Information Technology for supporting directly or indirectly throughout the research work. The authors also would like to acknowledge LogPoint A/S for providing them with relevant data and environment to carry out the research.

REFERENCES

[1] Varun Chandola, Arindam Banerjee and Vipin Kumar,Anomaly Detection : A survey,Technical report ACM Computing Surveys,September 2009

[2] Sami Nousiainen, Jorma Kilpi,Paula silvonen & Mikko Hiirsalmi, ”Anomaly detection from server log data”,ISBN978-951-38-7289-2(URL: http://www.vtt.fi/publications/index.jp)

[3] Rahul Jain, Tejpal Singh, Amit Sinhai, A Survey on Network Attacks, Classification and models for Anomaly-based network intrusion detection systems”, ISSN 2319-5991

[4] Mostaque Md. Morshedur Hassan, ”Current studies on intrusion,detection system, genetic algorithms and fuzzy logic,” International Journal of Distributed and Parallel Systems (IJDPS) Vol.4, No.2, p. 13, 2013

[5] Nong ye, Syed Masum Emran, Qiang Chen, and Sean Vilbert, ”Multivariate Statistical Analysis of Audit Trails for Host-Based Intrusion Detection,” IEEE Transactions on Computers, Vol 51, No 7, Jul 2002.

[6] Mohammad Said Asem Khalidi,”Multivariate Quality Control: Statistical performance and economic feasibility”, p.6, p11-32

[7] Khaled Labib, V Rao Vemuri, ”An application of principal component analysis to the detection and visualization of computer network attacks,” Annales of Telecommunications, 2006,225-234.

[8] Cha, Sung-Hyuk; Tappert, Charles C (2009). A Genetic Algorithm for Constructing Compact Binary Decision Trees.Journal of Pattern Recognition Research 4 (1): 113.

[9] Mei-ling Shyu, Shu-Ching Chen,Kanoksri Sarinnapakorn and LiWu Chang, A novel anomaly detection scheme Based on Principal Component Classifier, In Proceedings of 3rd IEEE International Conference on Data Mining. 353365,2003

[10] Simon Haykin, Neural Networks and Learning Machines,third edition, kernel methods and Radial-Basis Function Networks, Chapter 5, pg 270.

[11] Iwan Syarif, Adam Prugel-Bennett and Gary Wills, Unsupervised Clustering approach for network anomaly detection,Network Digital Technologies,2012.

[12] Breiman, L., Freidman, J.H., Olshen, R.A. and Stone, C.J. (1984). Classification and regression trees Wadsworth, Belmont CA

[13] Bentley, J.L., Multidimensional binary search trees used for associative searching, Communications of the ACM,1975

[14] Cup, K. D. D. ”Intrusion detection data set.” (1999).

[15] Shyu, Mei-Ling, et al. ”A novel anomaly detection scheme based on principal component classifier”. MIAMI UNIV CORAL GABLES FL DEPT OF ELECTRICAL AND COMPUTER ENGINEERING, (2003).