Communications on Applied Electronics (CAE) – ISSN : 2394-4714
Foundation of Computer Science FCS, New York, USA
Volume 4– No.3, January 2016 – www.caeaccess.org

Feature or Attribute Extraction for Intrusion Detection System using Gain Ratio and Principal Component Analysis (PCA)

O. Isaiah Aladesote
Computer Science Dept.
The Federal Polytechnic
P. M. B. 727, Ile Oluji,
Ondo State, Nigeria

Agbelusi Olutola
Computer Science Dept
Rufus Giwa Polytechnic
P. M. B. 1019, Owo,
Ondo State, Nigeria

Olasehinde Olayemi
Computer Science Dept
The Federal Polytechnic
P. M. B. 727, Ile Oluji,
Ondo State, Nigeria

ABSTRACT
Intrusion detection systems (IDS) refer to a category of defense tools that is used to provide warnings indicating that a system is under attack or intrusion. The IDS monitors activities within a network and alerts security administrators of suspicious activities. This paper extracted significant or highly relevant features or attributes of the Knowledge Discovery and Data Mining 1999 (KDD '99) dataset, which is a standard benchmark dataset for all intrusion problems using two features extraction techniques: Gain Ratio for discrete attributes and Principal Component Analysis (PCA) for continuous attributes. C# Programming language was used for the implementation. Also, Microsoft Excel was used to depict the result of the extraction. The result shows that thirteen (13) attributes were highly relevant and significant.

Keywords
Gain Ratio, Principal Component Analysis (PCA), Microsoft Excel, KDD '99 dataset, Intrusion Detection System

1. INTRODUCTION
Information security is a matter of serious worldwide concern as the incredible development in connectivity and accessibility to the internet has generated a tremendous security threat to information systems worldwide. Security is turning out to be a serious issue as internet applications are developing continuously. The majority of current security system mostly concentrates on encryption, firewall and access control. However all these approaches cannot promise perfect security. And hence, the security of a system can be improved upon by the introduction of intrusion detection system (Jaiganesh et al, 2012). Securing a network against attacks has become of great importance because of the increase in the services on the network coupled with the sensitivity of information on it. Intrusion detection system is the solution to securing a network since intrusion prevention techniques have suffered a great setback (Mrutyunjaya P. et al, 2007).

An Intrusion is defined as any set of actions that attempt to compromise the integrity, confidentiality or availability of a resource. This includes a deliberate unauthorized attempt to access information, manipulate information, or render a system unreliable or unusable (Dharamraj et al, 2010). Integrity can be compromised when intruders are able to modify the data, thus making the information unreliable (Tsai et al, 2006). Confidentiality can be breached when a non-privileged user is able to access the information. Availability, such as Denial of Service (DoS) attacks on the Internet, thus this can disrupt the web service and force information to be unavailable (Flora, 2009).

Feature extraction is used to reduce the dimension of a large dataset having many features or attributes that are both significant and non-significant. It is worth noting that such dataset will contain both redundant and duplicate records that need to be transformed into a reduced form of features, also called features vector. If the process of extracting significant features of a large dataset is carefully done or carried out, this will lead to selection of only features that shows relevant information rather than full size dataset. The purposes of feature extraction are to reduce the dimension, that is reduction in the attribute of a large dataset and to improve the classification accuracy by elimination both redundant and irrelevant information (Rupati et al, 2012).

2. RELATED WORKS
Mrutyunjaya P. et al (2007) proposed a framework on Network Intrusion Detection using Naïve Bayes. The experiment carried out over KDD '99 dataset, was able to detect attacks. The result obtained was compared with other researcher's result using neural network algorithms, the comparison revealed that Naïve Bayes algorithm gives higher detection rates, low cost factor and also consumes less time. Applying feature extraction technique(s) on KDD '99 dataset would have helped in removing redundant and non-significant features, leading to robust and better performance results with the approach (Naïve Bayes) applied on it.

In the work of Jainganesh and Sumathi (2012), Support Vector Machine and Kernel Support Machine with Levenberg-Marquardt were applied on KDD dataset. The data was segmented in training and testing dataset. Support Vector Machine was applied on it and Kernel Support Machine was tuned with Levenberg-Marquardt on training dataset with the purpose of constructing and training the models. The trained models were evaluated on testing dataset. The accuracy of the proposed system was tested based on detection rate and false alarm rate. The detection rate for KSVM with LM was better than SVM in all other attacks. Applying feature extraction techniques(s) will give better classification result.

3. OBJECTIVE OF THE STUDY
The objective of the research is to extract relevant attributes or features of KDD '99 dataset.

Communications on Applied Electronics (CAE) – ISSN : 2394-4714
Foundation of Computer Science FCS, New York, USA
Volume 4– No.3, January 2016 – www.caeaccess.org

4. METHODOLOGY

The existing works of the authorities in the field of Intrusion Detection System were reviewed. Gain Ratio and Principal Component Analysis (PCA) techniques were employed to extract significant features of discrete and continuous attributes type of KDD '99 dataset respectively. Knowledge Discovery and Data Mining 1999 (KDD '99) dataset, which is an effective benchmark dataset to help researchers on intrusion detection problems. According to Chou et al. [2007], the DARPA (Defense Advanced Research Projects Agency) KDD '99 dataset is made up of a large number of network traffic connections and each connection is represented with 41 features. Further, each connection had a label of either normal or the attack type.

Seven (7) of these attributes are discrete in nature while the remaining thirty-four (34) attributes are continuous. The dataset has seven (7) attributes that are discrete in nature. Let D be set consisting of d data samples with n distinct classes. The expected information needed to classify a given sample is given by (Asha Gowda Karegowda et al, 2010)

$$I(D) = -\sum_{i=1}^{n} p_i \log_2 p_i \quad \ldots\ldots\ldots\ldots \quad 1$$

where pi is the probability that an arbitrary sample belongs to class Ci and is estimated by di/d.

Let attribute A has v distinct values. Let dik be number of samples of class Ci in a subset Dj. Dj contains those samples in D that have value aj of A. The expected information or entropy based on the partitioning into subsets by A, is given by

$$E(A) = -\sum_{i=1}^{n} I(D) \frac{d_{1i} + d_{2i} + \cdots d_{ni}}{d} \quad \ldots\ldots\ldots\ldots \quad 2$$

The information gained is given by

$$\text{Gain}(A) = I(D) - E(A) \quad \ldots\ldots\ldots\ldots \quad 3$$

where E(A) is the entropy of the A and I(D) is the expected information.

$$\text{The splitInfoA (D)} = -\sum_{i=1}^{v} (|Di|/|D|) log_2 (|Di|/|D|) \quad 4$$

Equation (1.3) represents the information generated by splitting the training data set D into v partitions corresponding to v outcomes of a test on the attribute A.

$$\text{Gain Ratio}(A) = \frac{Gain\,(A)}{\text{SplitInfo}\ A(D)} \quad 5$$

$$\text{The threshold Y1} > \frac{1}{d-1} \left(\sum a_i^2 - \frac{1}{d} \left(\left(\sum ai\right)^2\right)\right) \quad 6$$

The remaining thirty- four (34) attributes of the data set are continuous in nature. Let D be set consisting of M data samples. Suppose x1, x2, …..,xM are Nx1 vectors, the mean of the data samples is given by (Shilpa, L. et al, 2010)

$$x(D) = \frac{1}{M} \sum_{i}^{M} xi \quad 7$$

The mean vector of the data M data samples is given by

$$\phi_i = x_i - x \quad 8$$

where ϕ_i is the mean vector, xi is the data sample and x is the mean. The covariance matrix of the sample data from the matrix A = [ϕ1ϕ2…ϕM] (N*M) is given by

$$C = \frac{1}{M} \sum_{N=1}^{M} \phi_n \phi'_n = AA^T \quad \ldots\ldots\ldots\ldots \quad 9$$

The eigenvalues of the covariance matrix, C for the sample data is given

$$C: \lambda_1 > \lambda_2 > \ldots\ldots > \lambda_N \quad \ldots\ldots\ldots\ldots \quad 10$$

$$\text{The eigenvectors of C is: } u_1, u_2, \ldots u_N \ldots\ldots\ldots \quad 11$$

$$\text{The threshold Y2} \geq \frac{1}{M} \sum_{l}^{M} \lambda i \ldots\ldots\ldots\ldots \quad 12$$

These two extraction techniques were implement using C# Programming language

5. RESULTS

The highly relevant attributes from the dataset were selected using Gain Ratio and Principal Component Analysis (PCA) for discrete and continuous attributes respectively. Table 1 and table 2 show the results of attribute selection using Gain Ratio and PCA respectively.

Table 1: Result of attribute selection using Gain Ratio

S/N	Attribute	Information Gain	Gain Ratio
1	Land	0.0053	1.0392
2	Flag	1.0448	0.6748
3	Logged_in	0.5078	0.5099
4	Protocol	0.3154	0.4227
5	Service	1.089	0.3775
6	Guest_login	0.0023	0.0885
7	Is_host_login	0	0

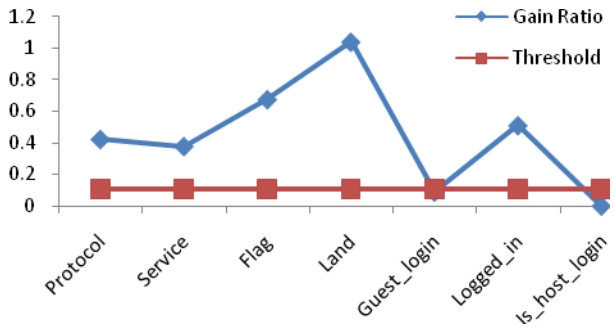

Figure 1: Graphical Representation of Gain Ratio Result

Figure 1 and Table 1 show that the Gain ratio of guest_login and is_host_login attributes are below the threshold value (see equation 1.5), which is 0.105752 and therefore considered weakly relevant and insignificant. Thus, they were removed. The Gain ratio of the following attributes: protocol, service, flag, land and logged_in are above the threshold value, and therefore considered to be highly relevant or significant.

Table 2: Result of Attribute Selection using PCA

S/N	Attribute	PCA
1	Duration	0.765
2	Src_bytes	0.738
3	Wrong_fragment	0.683
4	Srv_rerror_rate	0.678
5	Num_compromised	0.623

6	Dst_bytes	0.597
7	Dst_host_srv_diff_host_rate	0.581
8	Serror_rate	0.503
9	Dst_host_srv_rerror_rate	0.321
10	Su_attempted	0.318
11	Num_outbound_cmds	0.317
12	Count	0.316
13	Diff_srv_rate	0.316
14	Dst_host_srv_serror_rate	0.316
15	Dst_host_rerror_rate	0.313
16	Num_file_creations	0.305
17	Num_failed_logins	0.303
S/N	**Attribute**	**PCA**
18	Num_access_files	0.302
19	Urgent	0.301
20	Srv_diff_host_rate	0.301
21	Srv_count	0.286
22	Dst_host_serror_rate	0.276
23	Dst_host_srv_count	0.264
24	Root_shell	0.248
25	Dst_host_diff_srv_rate	0.234
26	Num_root	0.217
27	Num_shells	0.197
28	Dst_host_same_src_port_rate	0.186
29	Srv_serror_rate	0.157
30	Dst_host_same_srv_rate	0.119
31	Dst_host_count	0.116
32	Rerror_ rate	0.091
33	Hot	0.056
34	Same_srv_rate	0.017

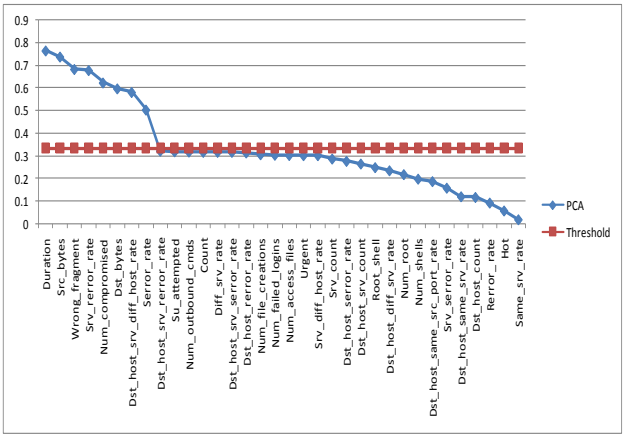

Figure 2: Graphical Representation of PCA Result

Figure 2 and Table 2 show that the PCA of the attributes duration, dst_bytes, wrong_fragment, num_compromised serror_rate, srv_rerror_rate, src_bytes and dst_host_srv_diff_host_rate are above the threshold value (see equation 2.1), which is 0.334147 and therefore considered highly relevant and significant.

6. CRITICAL ANALYSIS OF RESULT/ DISCUSSION

All attributes whose results are above the set threshold value for each extraction technique were selected (see tables 1 & 2). This shows that thirteen (13) out of the forty -one (41) attributes of the dataset are relevant and highly significant and it will lead to drastic reduction in time of detection process and also upgrading the performance of intrusion detection system.

7. CONCLUSION AND RECOMMENDATION

With these features extraction techniques (Gain Ratio & PCA), the 41 attributes of the KDD dataset were reduced to 13. This reveals that the 13 attributes are considered to be highly relevant while the remaining 28 are considered irrelevant. The highly relevant attributes are Protocol, Service, Flag, Land, Logged_in, Duration, Dst_bytes, Wrong_fragment, Num_compromised, Serror_rate, Srv_rerror_rate, Src_bytes and Dst_host_srv_diff_host_rate.

It is therefore recommended that all researchers on intrusion detection should adopt the thirteen attributes extracted to be highly significant in order to reduce the time consumption in detection process and also bring about efficient and effective system performance. The limitation of this work is the setting of the threshold.

8. AREA OF FURTHER RESEARCH

The research work focused on the feature extraction of significant attributes of KDD '99 dataset. Further research work is hereby recommended that a method such as hypothesis testing and other techniques should be applied on the selected attributes. This will help to ascertain the effectiveness and efficiency of the system performance.

9. REFERENCES

[1] Ankita G., and Richariya V. (2007): "A Layered Approach for Intrusion Detection using Meta- modelling with Classification Techniques," International Journal of Computer Technology & Electronics Engineering (IJCTEE) Vol. 1, Issues 2.

[2] Asha Gowda Karegowda, A. S. Manjunati & M. A. Jayaram (2010): "Comparative Study of Attributes Selection using Gain Ratio and Correlation Based Feature Selection", International Journal of Information Technology and Knowledge Management. Volume 2, No. 2, pp. 271 – 277, July – December 2010.

[3] Dharamraj R. Patil and V. P. Kshirsagar (2010): "An overview of adaboost-based NISD and Performance evaluation on NSL –KDD dataset", International Journal of Computer Engineering and Computer Application, vol. 1, 2010.

[4] Flora S. T. (2009), "Network Intrusion Detection using Associative Rules," International Journal of Recent Trends in Engineering, Vol. 2, No. 2, November 2009.

[5] Jaiganesh, V. and Sumathi, P. (2012): "Intrusion Detection using Kernelized Support Vector Machine with Levenberg – Marquardt Learning", International

Journal of Engineering Science and Technology (IJEST), vol. 4 No. 03, pp. 1153 – 1160, March 2012.

[6] Kayacik, H. G., Zincir-Heywood, A. N, and Heywood M. I. (1999): "Selecting Features for Intrusion Detection: A Feature Relevance Analysis on KDD '99 Intrusion Detection Datasets," http://www.cs.dal.ca/projectx/

[7] KDD Cup 1999 Data: Available on http://kdd.ics.uci.edu/database/kddcup99/Database/kddcup99/kddcup99.html, October 2007

[8] Mrutyunjaya Panda and Manas Ranjan Patra (2007): "Network Intrusion Detection using Naive Bayes", International Journal of Computer Science and Network Security (IJCSNS), Vol. 7, No. 12, December 2007.

[9] Rupati D, and Bhupendra V. (2010): "Feature Reduction for Intrusion Detection using Linear Discriminant Analysis", International Journal on Computer Science and Engineering (IJCSE) Vol. 02, No. 04, 2010, 1072 – 1078.

[10] Shilpa, L., Joseph S., and Bhupendra V. (2010): "Feature Reduction using Principal Component Analysis for Effective Anomaly-Based Intrusion Detection on NSL-KDD," International Journal of Engineering Science and Technology, vol. 2(6), 2010, 1970 – 1977.

[11] Tsai F. S., Chan C. K. (eds), Cyber Security, Pearson Education, Singapore, 2006.