# Identification of Effective Network Features for Probing Attack Detection

Gholam Reza Zargar
*School of Computer Engineering*
*Iran's University of Science and Technology*
*Tehran, Iran*
*Zargar@vu.iust.ac.ir*

Peyman Kabiri
*School of Computer Engineering*
*Iran's University of Science and Technology*
*Tehran, Iran*
*Peyman.Kabiri@iust.ac.ir*

## Abstract

*Existing intrusion detection techniques emphasize on building intrusion detection model based on all features provided. But not all the features are relevant ones and some of them are redundant and useless. This paper proposes and investigates identification of effective network features for Probing attack detection using PCA method to determine an optimal feature set. An appropriate feature set helps to build efficient decision model as well as a reduced feature set. Feature reduction will speed up the training and the testing process considerably. DARPA 1998 dataset was used in the experiments as the test data. Experimental results show a reduction in training and testing time while maintaining the detection accuracy within acceptable range.*

## 1. Introduction

Intrusion Detection System (IDS) deals with huge amount of data which contains irrelevant and redundant features. These two types of features will slow down training and testing processes, will cause higher resource consumption and reduced detection rate.

All feature selection methods need to use an evaluation function together with a search procedure to obtain the optimal feature set. The evaluation function measures how good a specific subset can be in discriminating between classes. These measures can be divided into two main groups: filters and wrappers. Filters measure the relevance of feature subsets independent of classifier type, whereas wrappers use the classifier's performance as the evaluation measure.

IDSs are typically classified as host-based or network-based. A host-based IDS will monitor resources such as system logs, file systems and disk resources; whereas a network-based IDS monitors the data passing through the network. Different detection techniques can be employed to search for attack patterns in the data monitored. Signature-based detection systems try to find attack signatures in the monitored resource. Anomaly detection systems typically rely on knowledge of normal behavior and flag any deviation from it [1]. Signature-based IDSs typically require human input to create attack signatures or to determine effective models for the normal behavior. Implementing the learning algorithms provide a potential alternative to the expensive human input.

Inefficient feature selection method not only reduces speed of its operation but also declines its accuracy [2].

The main goal in feature selection is to reduce the amount of data which are less important to the classification and can be eliminated. This has the benefit of decreasing storage requirements, reducing processing time and improving the detection rate. IDS has to examine a very large audit data in a short period of time. Therefore, any reduction in the volume of data may save the processing time [3].

This paper proposes a method based on TCP/IP Header parameter. In the proposed approach, Principal Component Analysis (PCA) is used as a dimension reduction technique

PCA is a statistical method applicable to feature selection. It is employed to reduce the high dimensionality in data and consequently improve the efficiency of the process and reduce usage of the system resources. In the proposed approach a reduced number of dimensions in the feature space is used for the detection of different types of attacks and normal activities. The low computational cost of this approach improves the real-time performance of the IDS.

## 2. Related Works

Data reduction can be achieved by filtering, data clustering and feature selection [4]. Generally, the capability of anomaly-based IDS is often hindered by its inability to accurately classify variation of normal behavior as an intrusion. Additionally, network

978-1-4244-4615-5/09/$25.00 ©2009 IEEE 

traffic data is huge and it causes a prohibitively high overhead that often becomes a major problem for IDS [5]. Chakraborty in a published paper [6] mentions that, "the existence of these irrelevant and redundant features generally affects the performance of machine learning or pattern classification algorithms". Hassan et al [7], proved that proper selection of feature set has resulted in better classification performance. Sung and Mukkamala [4], have exploited SVM [8] and Neural Networks to identify and categorize features with respect to their importance in regard to detection of specific kinds of attacks such as probe, DoS, Remote to Local (R2L), and User to Root (U2R). They have also demonstrated that the elimination of these unimportant and irrelevant features did not significantly reduce the performance of the IDS. Chebrolu et al [3], tackled the issue of effectiveness of an IDS in terms of real time performance and detection accuracy from the feature reduction perspective.

## 3. Problem of Irrelevant and Redundant Features

Basically, there are two approaches for designing IDS: Signature-based and Anomaly-based intrusion detection [9]. Signature-based intrusion detection is also called misuse detection. In principle, it is typically realized by modeling known attack behavior with prior understanding about specific attacks and system vulnerabilities. In this method, the IDS compares network traffic data versus well defined attack patterns to identify any possible penetrations into the system. Once an input pattern is detected to be similar to one of explicitly defined attack patterns in IDS, an alarm is raised. Defined attack patterns are frequently referred to as the signatures of intrusions. The signature could be a static string or a sequence of events.

While signature-based intrusion detection is achieved by modeling known attack behavior, on the contrary, anomaly-based intrusion detection models normal or expected behavior of computer users. It looks for malicious activities by comparing the observed data with these acceptable behaviors. If the data collected from the network traffic differs from the known normal behavior, an alarm is raised. In other words, anything will be suspected to be an attack if its behavior is deviated from the previously learned behaviors. However, there is one major problem in the collected network traffic database: problem of irrelevant and redundant features [2] Details of this problem are described in the following

In general, the quantity of data processed by the IDS is large; it includes thousands of traffic records with a number of various features such as the length of the connection, type of the protocol, type of the network service and many other information. Theoretically and ideally, the ability to discriminate attacks from normal behavior should be improved if more features are used for the analysis. However, this assumption is not always true, since not every feature in the traffic data is relevant to the intrusion detection. Among the large amount of features, some of them may be irrelevant or confusing with respect to the target patterns. Some of the features might be redundant due to their high inter-correlation with one or more of the other features in the dataset [9]. To achieve a better overall detection performance, any irrelevant and redundant features should be discarded from the original feature space. Selecting a meaningful subset of features from network traffic data stream is therefore a very important and indispensable task at the early stages of an intrusion detection process.

## 4. Data Reduction and feature selection using PCA

PCA is one of the most widely used dimensionality reduction techniques for data analysis and compression. It is based on transforming a relatively large number of variables into a smaller number of uncorrelated variables. This transformation is carried out by finding those orthogonal linear combinations of the original variables with the largest variance. The first principal component of the transformation is the linear combination of the original variables with the largest variance. The second principal component is the linear combination of the original variables with the second largest variance and orthogonal to the first principal component and so on. In many datasets, the first several principal components have the highest contribution to the variance in the original dataset. Therefore, the rest can be ignored with minimal loss of the information value during the dimension reduction process [10]. The transformation works as follows:
Given a set of observations $x_1$, $x_2$, …, $x_n$, where each observation is represented by a vector of length *m*, the dataset is thus represented by a matrix $X_{n\times m}$

$$X_{n\times m}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}=[x_1,x_2,\ldots x_n] \quad (1)$$

The average for each observation is defined by the equation (2).

$$\mu=\frac{1}{n}\sum_{i=1}^{n}x_i \quad (2)$$

The deviation from the mean is defined in (3).

$$\phi_i = x_i - \mu \tag{3}$$

The sample covariance matrix of the dataset is defined in (4)

$$C = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T = \frac{1}{n}\sum_{i=1}^{n}\phi_i\phi_i^T = \frac{1}{n}AA^T \tag{4}$$

Where $A = [\phi_1, \phi_2, ..., \phi_n]$

In applying PCA to reduce high dimensional data, eigenvalues and corresponding eigenvectors of the sample covariance matrix *C* have to be calculated [11]. Let $(\lambda_1, u_1), (\lambda_2, u_2), ....(\lambda_m, u_m)$ present *m* eigenvalue-eigenvector pairs of the sample covariance matrix *C*. The *k* eigenvectors associated with the largest eigenvalues are selected. The dimensionality of the subspace *k* can be determined by the following equation [12]:

The resulted $m \times k$ matrix *U,* with *k* eigenvectors as its columns is called eigenvectors matrix or coefficient matrix. Data transformation using principal components into the *k*-dimensional subspace is carried-out using equation (5).

$$y_i = U^T(x_i - \mu) = U^T\phi_i \tag{5}$$

## 5. The Dataset and Pre-processing

In the following subsections the dataset used in the reported work and the pre-processing applied on the dataset are presented.

## 6. The Dataset used in this work

In this work, basic features are extracted from TCP/IP packets [13]. These features can be derived from packet headers without inspecting the payload. In the reported work, TCP dump from the DARPA'98 dataset is used as the input dataset. Extracting the basic features, packet information in the TCP dump file is summarized into connections. Specifically, "a connection is a sequence of TCP packets starting and ending at some well defined times, between which data flows from a source IP address to a target IP address under some well defined protocol" [14].

DARPA'98 dataset provides around 4 gigabytes of compressed TCP dump data [15] for 7 weeks of network traffic [16]. This dataset can be processed into about 5 millions of connection records each about 100 bytes in size. The resulted dataset contains the payload of the packets transmitted between hosts inside and outside a simulated military base. BSM audit data from one UNIX Solaris host for some network sessions were also provided. DARPA 1998 TCP dump Dataset [16] was preprocessed and labeled with two class labels, e.g. normal and attack. The dataset contains different types of attacks. Satan and Portsweep from Probing attack category are extracted. Probing is a class of attacks in which an attacker scans a network of computers to collect information or find known vulnerabilities. An intruder with a map of machines and services that are available on a network can use this information to look for exploits. Examples for this type of attack include Ipsweep, Mscan, Nmap, Saint, Satan, ping-sweep and Port sweep attacks.

## 7. Pre-processing

As displayed in Table 1, 32 basic features are extracted from TCP, IP, UDP and ICMP protocols. Wireshark and Editcap softwares are used to analyze and minimize TCP dump files [13][17]. Finally, a Visual Basic program extracts the 32 basic features.

In this work, intention is to reduce the processing and data transfer time needed for the intrusion detection. To do so, an accurate feature selection scheme is proposed to select important features with minimum loss of information. Work also aims to select features in such a way that their discrimination set to be categorical. This means that the selection criteria will be the same or with a low variance for the attacks in the same category. This property will increase the adaptability of the IDS that is using this feature set to the variation of the attack patterns that fall in the same category.

## 8. Experiments

Training data from the DARPA dataset includes "list files" that identify the timestamp, source host and port, destination host and port, and the name of each attack [18]. This information is used to select intrusion data for the purpose of pattern mining and feature construction, and to label each connection record with "normal" or "attack" label types. The final labeled training data is used for training the classifiers. Due to the large volume of audit data, connection records are stored in several data files.

**Table 1. Basic feature extracted from the header of TCP/IP**

| No. | Feature | Description |
|---|---|---|
| 1 | Protocol | Type of Protocol |
| 2 | Frame_lenght | Length of Frame |
| 3 | Capture_lenght | Length of Capture |
| 4 | Frame_IS_marked | Frame IS Marked |
| 5 | Coloring_rule_name | Coloring Rule name |
| 6 | Ethernet_type | Type of Ethernet Protocol |
| 7 | Ver_IP | IP Version |
| 8 | Header_lenght_IP | IP Header length |
| 9 | Differentiated_S | Differentiated Service |
| 10 | IP_Total_Lenght | IP total length |
| 11 | Identification_IP | Identification IP |
| 12 | MF_Flag_IP | More Fragment flag |
| 13 | DF_Flag_IP | Don't Fragment flag |
| 14 | Fragmentation_offset_IP | Fragmentation offset IP |
| 15 | Time_to_live_IP | Time to live IP |
| 16 | Protocol_no | Protocol number |
| 17 | Src_port | Source Port |
| 18 | Dst_port | Destination port |
| 19 | Stream_index | Stream Index number |
| 20 | Sequence_number | Sequence number |
| 21 | Ack_number | Acknowledgment number |
| 22 | Cwr_flag | Cwr Flag(status flag of the connection) |
| 23 | Ecn_echo_flag | Ecn Echo flag (status flag of the connection) |
| 24 | Urgent_flag | Urgent flag(status flag of the connection) |
| 25 | Ack_flag | Acknowledgment flag(status flag of the connection) |
| 26 | Psh_flag | push flag(status flag of the connection) |
| 27 | Rst_flag | Reset flag(status flag of the connection) |
| 28 | Syn_flag | Syn flag (status flag of the connection) |
| 29 | Fin_flag | Finish flag(status flag of the connection) |
| 30 | ICMP_Type | specifies the format of the ICMP message such as: (8=echo request and 0=echo reply) |
| 31 | ICMP_code | Further qualifies the ICMP message |
| 32 | ICMP_data | ICMP data |

The traffic data on Thursday of the 6th week was selected as the sampled dataset, since it includes not only normal behaviors, but also a large number of Satan and Port sweep attacks that are types of probing attack activities. Table 2 shows number of the connection records that are randomly extracted from the sequences of normal and probing attack connection records to create the normal dataset and attack dataset. Together they make the sampled dataset for the data analysis using PCA.

Dictionary table is used to convert text data into numeric data. Matlab software is used in this work for the calculations mentioned in section 4.

**Table 2. Number of records selected for the sampled dataset.**

| Category | Number of Records |
|---|---|
| PROBING ATTACK | 10137 |
| NORMAL | 88860 |

## 9. Experimental Results

Figure 1, shows correlation between features, the correlation is mainly between features 29 and 26. As the amount of correlation between the two parameters is increased, the relation between parameters and their behavioral similarity is more justifiable. Analyzing their correlation requires comparing observations on different parameters, such as source and destination IP address, an identifiable network route, commands entered by a suspected attacker and the time when activity began or ended.

Each behavior is constructed with some parameters. For example, in every TCP session, the "protocol" field of IP packet is equal to "6". As another example, identification field in IP layer for each massage have the same value. In a TCP Flood attack, and probing attacks, in addition to some other features, there are lots of TCP packets with "SYN=1". But an ICMP flood attack never has these specifications. In fact, each attack has its own properties that are different from the properties of other attacks and even with the normal behavior. These features are used to compare each session versus the normal or a known attack behavior.

In Probing attack, the attacker scans a network of computers to collect information for detecting which port is open with TCP/IP vulnerabilities. For example, some attackers use the three way handshaking method in TCP connection. In this method, TCP flags such as finish flag (fin_flag), push flag (psh_flag), acknowledge flag (ack_flag) and Syn flag (syn_flag) are more effective.

The result is reported in Table 3. Figures 2, shows the information value of the features with respect to their variances. Table 3 shows classes of the relevant features with their associated information value. In this table, probing attack is compared versus normal state of operation. As it is shown in Table 3, different

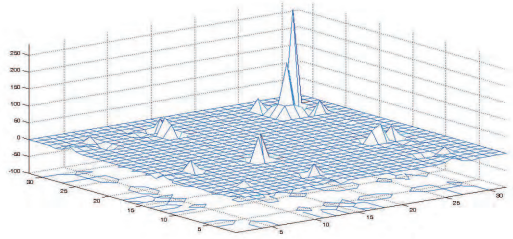

**Figure 1. Covariance matrix**

feature sets are selected for the probing and the normal state. Component's 29, 26, 25 and 28, all TCP flags, are more effective and have more information value than the rest of the features. These components are fin flag, psh flag, Ack flag and syn flag respectively as they are presented in Table 1. As calculated by the PCA, their total information value is %70.97 of the total information in the feature set. If components 12, 13, 5, 27, 1 and 10 are added to them, the total information value will climb to %98.01. In the normal dataset, component number 27 have the maximum information value with %98.22 of information. Once the component number 25 is considered as well, their total information value will rise to %98.84 of the total information value. Therefore, it can be said that the component number 25 does not have a significant effect in detecting the normal state. Figure 2, compares components of normal state and probing attack. The vertical axis in this figure shows absolute value of the eigenvalues.
Figure 3 explains the method for the selection of the component 27 to detect the normal state and components 29, 26, 28 for detecting the probing attack. The SCREE graph for the PCA coefficient for probing attack is depicted in Figure 4.
In order to use these features, one should first calculate a threshold value (or range) for the features. Once the weights of the features are calculated, feature values should be multiplied by these weight values. As for example, the feature number 27 multiplied by its corresponding weight should be within the selected range or the state of the network is anomalous.

## 10. Conclusions

Results presented in this paper show that normal state of the network and the probing attack features can be selected to be discriminate. On the other hand, it is proven that certain features have no contribution to intrusion detection. This also indicates that there

**Table 3. List of the relevant features for the Probing attack and the normal state of the network**

| Class name | Important features in descending order | Total accumulated information value |
|---|---|---|
| Probing | 29 | %39.03 |
| Probing | 29,26 | %50.30 |
| Probing | 29,26,25 | %60.89 |
| Probing | 29,26,25,28 | %70.97 |
| Probing | 29,26,25,28,12 | %79.87 |
| Probing | 29,26,25,28,12,13 | %88.76 |
| Probing | 29,26,25,28,12,13,5 | %92.84 |
| Probing | 29,26,25,28,12,13,5,27 | %95.68 |
| Probing | 29,26,25,28,12,13,5,27,1 | %97.04 |
| Probing | 29,26,25,28,12,13,5,27,1,10 | %98.01 |
| Normal | 27 | %98.22 |
| Normal | 27,25 | %98.84 |

are analytical solutions for the feature selection that are not based on the trial and error.

## 11. Future work

Plan for the future work is to calculate PCA for other attack categories and find feature sets for different attack categories. Later on, intention is to use classification methods to detect intrusions. Using the results derived from the intrusion detection and comparing it for the full feature set and the reduced feature set, one can analyze the different in accuracy and speed between them.

## 12. References


[1] I. Guyon, A. Elisseeff, "An Introduction to Variable and Feature Selection", *Journal of Machine Learning Research*, vol.3, pp. 1157-1182, 2003.

[2] T.S. Chou, K.K. Yen, and J. Luo, "Network Intrusion Detection Design Using Feature Selection of Soft Computing Paradigms", *International Journal of Computational Intelligence*, vol.4, no.3, pp. 196-208, 2008.


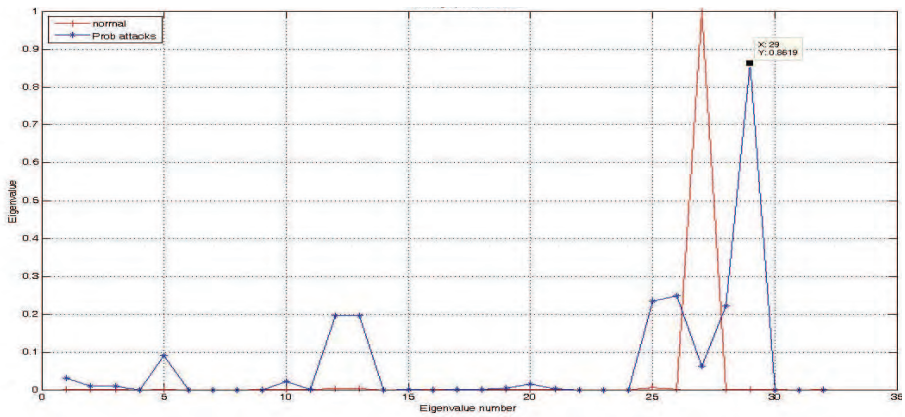


**Figure 2. A comparison between information value in the feature space for the Probing attack and the normal state of the network.**

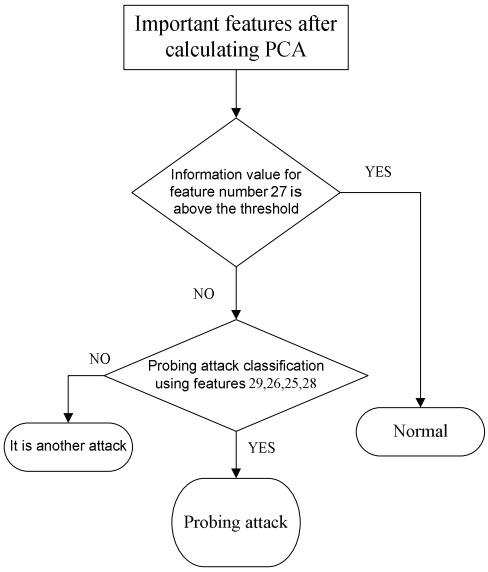


**Figure 3. Flowchart for detecting probing attack.**


[3] S. Chebrolu, A. Abraham, and J. Thomas, "Feature Deduction and Ensemble Design of Intrusion Detection Systems", *Computers and Security*, Elsevier Science, vol.24, Issue 4, pp. 295-307, 2005.

[4] A.H. Sung, S. Mukkamala, "Identifying important features for intrusion detection using support vector machines and neural networks", *Proceedings of International Symposium on Applications and the Interne(SAINT)*, pp. 209-216, 2003.

[5] A.H. Sung, S. Mukkamala. "The Feature Selection and Intrusion Detection Problems", *ASIAN 2004, LNCS, Springer Hieldelberg*, vol. 3321, pp. 468-482, 2004.

[6] B. Chakraborty, "*Feature Subset Selection by Neurorough Hybridization"* LNCS, Springer Hiedelberg, pp.519-526, 2005.

[7] A. Hassan, M.S. Nabi Baksh, A.M. Shaharoun, and H. Jamaluddin. "Improved SPC Chart Pattern Recognition Using Statistical Feature", *International Journal of Production Research*, vol 41, Issue.7, pp. 1587-1603, 2003.

[8] V. Vapnik, "*The Nature of Statistical Learning Theory*" Springer, Berlin Heidelberg, New York, 1995.

[9] F. Sabahi, A. Movaghar, "Intrusion Detection: A Survey", *3rd International conference on system and network communication*, ICSNC08, pp.23-26, 2008.

[10] W. Wang, R. Battiti, "Identifying Intrusions in Computer *Networks based on Principal Component Analysis",* http://eprints.bibli.unitn.it/archive/00000917 as visited on 30 May 2009.

[11] G.H. Golub, C.F. Van Loan, "*Matrix Computation*", Johns Hopkins Univ. Press, Baltimore, 1996.

[12] I.T. Jolliffe, "*Principal Component Analysis* 2nd Ed"., Springer-Verlag, NY, 2002.

[13] http://www.wireshark.org as visited on 29 Jan 2009.

[14] Knowledge discovery in databases DARPA archive, Task Description. http://www.kddics.uci.edu/databases/kddcup99/task.html

[15] http://www.Tcpdump.org as visited on 28 Jan 2009.

[16] MIT Lincoln Laboratory http://www.ll.mit.edu/IST/ideval/ as visited on 27 Jan 2009.

[17] http://www.wireshark.org/docs/man-ages/editcap.html as visited on 20 Jan 2009.

[18] Wenke Lee, "A Data Mining Framework for Constructing Feature and Model for Intrusion Detection System", PhD thesis University of Columbia, 1999.


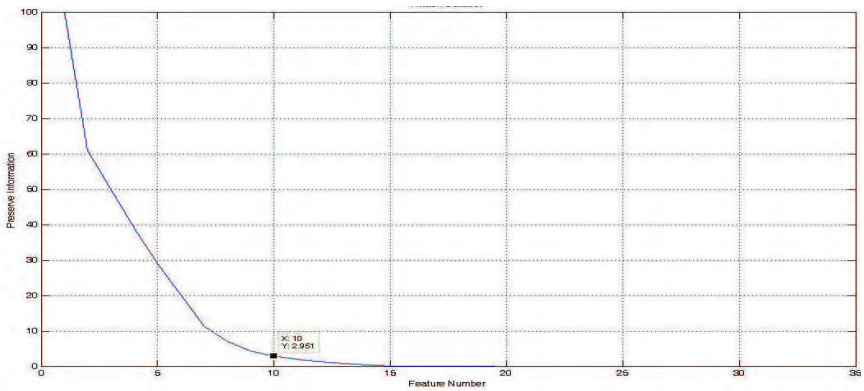


**Figure 4. Scree graph for the PCA coefficients calculated using the sampled dataset.**