2009 International Conference on Machine Learning and Computing
IPCSIT vol.3 (2011) © (2011) IACSIT Press, Singapore

An efficient feature reduction technique for intrusion detection system

Shailendra Singh[1 +], Sanjay Silakari and Ravindra Patel [1]

[1] Rajiv Gandhi Technological University, Bhopal, INDIA

Abstract. The information security is an issue of serious global concern. The network traffic data provided for the design of intrusion detection system always are large with ineffective information, thus we need to remove the worthless information from the original high dimensional database. To improve the generalization ability, we usually generate a small set of features from the original input variables by feature extraction. The conventional Principal Component Analysis (PCA) feature reduction technique has its limitations. It is not suitable for non-linear dataset. Thus we propose an efficient algorithm based on the Generalized Discriminant Analysis (GDA) feature reduction technique which is novel approach used in the area of intrusion detection. This not only reduces the number of the input features but also increases the classification accuracy and reduces the training time of the classifiers by selecting most discriminating features. We use Self-Organizing Map (SOM) and C4.5 classifiers to compare the performance of the proposed technique. The result indicates the superiority of GDA.

Keywords: Principal Component Analysis Generlalized Descriminant Analysis, Self-Organizing Map (SOM), C4.5.

1. Introduction

In recent years, many intrusion detection systems are studied and proposed to meet the challenges of vulnerable internet environment [1] [4]. According to the statistics of American Computer Emergency Response Team /Coordination Center (CERT) [2], network cases annually showed index growth in recent years and according to the report of information security [3], internet attacks have became new weapon of world war. Further the report said that Chinese Military Hacker had drew up plan, with the view of attacking American Aircraft Carrier Battle Group to making in it weak fighting capacity thorough internet. Such information reveals that there is an urgent need to effectively identify and hold up internet attacks. It is not an exaggerated statement that an intrusion detection system is must for modern computer systems. Anomaly detection and misuse detection [4] are two general approaches to computer intrusion detection system. Unlike misuse detection, which generates an alarm when a known attack signature is matched, anomaly detection identifies activities that deviate from the normal behaviour of the monitored system and thus has the potential to detect novel attacks [5]. The data we use here originated from MIT's Lincoln Lab. It was developed for KDD (Knowledge Discovery and Data mining) competition by DARPA and is considered a standard benchmark for intrusion detection evaluation program [6]. Empirical studies indicate that feature reduction technique is capable of reducing the size of dataset. The time and space complexities of most classifiers used are exponential function of their input vector size [7]. Moreover, the demand for the number of samples for the training the classifier grows exponentially with the dimension of the feature space. This limitation is called the 'curse of dimensionality.'

The feature space having reduced features that truly contributes to classification that cuts pre-processing costs and minimizes the effects of the 'peaking phenomenon' in classification [8]. Thereby improving the over all performance of classifier based intrusion detection systems. The most famous technique for

+ Corresponding author. Tel.: +91755-2678863; fax: +91755-2742006.
E-mail address: shailendrasingh@rgtu.net

dimensionality reduction is Principal Component Analysis [9] [10]. This technique searches for directions in the data that have largest variance and subsequently project the data into it. By this we obtain a lower dimensional representation of the data that removes some of the “noisy” directions. But this suffers from many difficult issues with how many directions one needs to choose. It fails to compute principal component in high dimensional feature spaces, which are related to input space by some nonlinear map.

In this paper we present Generalized Discriminant Analysis (GDA) [11] technique to overcome the limitations of PCA technique. This is unique approach to reduced size of attack data The Each network connection is transformed into an input data vector. GDA is employed to reduce the high dimensional data vectors and identification is handled in a low dimensional space with high efficiency and low use of system resources. The normal behaviour is profiled based on normal data for anomaly detection and the behaviour of each type of attack are built based on attack data for intrusion identification. Each reduced feature dataset is applied to the Self-Organizing Map (SOM) and C4.5 decision tree classifiers and their performance are compared.

2. The Data

In the 1998 DARPA intrusion detection evaluation [6] program, an environment was setup to acquire raw TCP/IP dump data for a network by simulating a typical U.S. Air Force LAN. The LAN was operated like a true environment, but being blasted with multiple attacks. For each TCP/IP connection, 41 various quantitative (continuous data type) and qualitative (discrete data type) features were extracted among the 41 features, 34 features are numeric and 7 features are symbolic. The data contains 24 attack types that could be classified into four main categories:

- DOS: Denial Of Service attack.
- R2L: Remote to Local (User) attack.
- U2R: User to Root attack.
- Probing: Surveillance and other probing.

Denial of service Attack (DOS)

Denial of service (DOS) is class of attack where an attacker makes a computing or memory resource too busy or too full to handle legitimate requests, thus denying legitimate user access to a machine.

Remote to Local (User) Attacks

A remote to local (R2L) attack is a class of attacks where an attacker sends packets to a machine over network, then exploits the machine’s vulnerability to illegally gain local access to a machine.

User to Root Attacks

User to root (U2R) attacks is a class of attacks where an attacker starts with access to a normal user account on the system and is able to exploit vulnerability to gain root access to the system.

Probing

Probing is class of attacks where an attacker scans a network to gather information or find known vulnerabilities. An attacker with map of machine and services that are available on a network can use the information to notice for exploit.

3. Feature extraction techniques

Feature extraction [12] includes feature construction, space dimensionality reduction, sparse representations, and feature selection. All these techniques are commonly used as pre processing to machine learning and statistics tasks of prediction, including pattern recognition and regression. Although such problems have been tackled by researchers for many years, there has been recently a renewed interest in feature extraction. A number of new applications with very large input spaces critically need space dimensionality reduction for efficiency of the classifiers.

In this section we discuss two techniques PCA and proposed GDA for reducing dimensionality of KDDCup99 intrusion detection dataset. Each feature vectors is labeled as an attack or normal. The distance

between a vector and its reconstruction onto those reduced subspaces representing different types of attacks and normal activities is used for identification.

3.1. Principal Component Analysis (PCA)

Contributions to Principal Component Analysis is technique used for feature extraction, data used in intrusion detection problem are high dimensional in nature. It is desirable to reduce the dimensionality of the data for easy exploration and further analysis. The PCA is often used for this purpose. PCA is concerned with explaining the variance-covariance structure of a set of variables through a few new variables. If there are M features in each datum and there are N data which is represented by x11, x12, x13…..x1M. x21,. x22, x23….x2M. Similarly the final datum can be represented by xN1, xN2, xN3….xNM.

The matrix $A = [\phi_1, \phi_2 ..... \phi_M])$ (N×M matrix)

The sample covariance matrix C of the data set is defined as

$$C = \frac{1}{M} \sum_{i=1}^{M} \phi_i \phi_i^T \tag{1}$$

We compute eigenvalues $(\lambda_1, \lambda_2, ... \lambda_N)$ and eignvectors $(u_1, u_2, ... u_N)$ of covariance matrix C. The K eigenvectors having the largest eigenvalues are selected. The dimensionality of the subspace K can be determined by using the following criterion.

$$\frac{\sum_{i=1}^{K} \lambda_i}{\sum_{i=1}^{N} \lambda_i} > threshold \quad (\alpha) \tag{2}$$

The linear transformation RN>RK that performs the dimensionality reduction is

$$Z_n = U^T (x - \overline{x}) = U^T \phi_n \tag{3}$$

PCA technique is applied to the KDDCup99 dataset and 19 features selected out of 41 features as shown in Table I. Resulting confusion matrices are obtained from SOM and C4.5 classifiers are shown in Table II and Table III respectively.

TABLE I. FEATURES SLECTED BY PCA TECHNIQUE

S. No	Feature	Type
1	Duration	Continuous
2	protocol_type	Discrete
3	Service	Discrete
4	Flag	Discrete
5	Src_bytes	Continuous
6	Dst_bytes	Continuous
7	Hot	Continuous
8	Num_compromised	Continuous
9	Num_root	Continuous
10	is_host_login	Discrete
11	is_guest_login	Discrete
12	Count	Continuous
13	Srv_count	Continuous
14	rerror_rate	Continuous
15	Diff_srv_rate	Continuous
16	Srv_diff_host_rate	Continuous
17	Dst_host_count	Continuous
18	Dst_host_same_srv_rate	Continuous
19	Dst_host_srv_diff_host_rate	Continuous

TABLE II. CONFUSION MATRIX FOR SOM CLASSIFIER BY PCA TECHNIQUE.

Predicted / Actual	Normal	Probe	DOS	R2L	U2R	%Correct
Normal	56520	3748	315	2	8	93.3
Probe	1302	2506	350	2	6	60.2
DOS	8593	1243	220010	2	5	95.7
R2L	11400	3027	7	1755	0	10.8
U2R	102	69	9	1	47	20.6
%Correct	72.6	23.7	99.7	99.6	70.8	

TABLE III. CONFUSION MATRIX FOR C4.5 CLASSIFIER BY PCA TECHNIQUE.

Predicted / Actual	Normal	Probe	DOS	R2L	U2R	%Correct
Normal	60187	334	68	2	3	99.32
Probe	143	4010	66	1	1	96.2
DOS	2795	605	226449	0	2	98.5
R2L	7852	470	1815	6159	1	38.0
U2R	99	52	4	2	73	32
%Correct	84.6	73.3	99.14	99.9	91.3	

3.2. Generalized Discriminant Analysis (GDA)

Generalized Discriminant Analysis is used for multi-class classification problems. Due to the large variations in the attack patterns of various attack classes, there is usually a considerable overlap between some of these classes in the feature space. In this situation, a feature transformation mechanism that can minimize the between-class scatter is used.

The Generalized Discriminant Analysis GDA [13] is a method designed for nonlinear classification based on a kernel function ϕ which transform the original space X to a new high-dimensional feature space $Z:\phi: X \to Z$. The within-class scatter and between-class scatter matrix of the nonlinearly mapped data is

$$B^{\phi} = \sum_{c=1}^{C} M_c m_c^{\phi} (m_c^{\phi})^T \tag{4}$$

$$W^{\phi} = \sum_{c=1}^{C} \sum_{x \in X_c} \phi(x)\phi(x)^T \tag{5}$$

Where m_c^{ϕ} is the mean of class X_c in Z and M_c is the number of samples belonging to X_c . The aim of the GDA is to find such projection matrix U^{ϕ} that maximizes the ratio

$$U_{opt}^{\phi} = \arg\max \frac{|(U^{\phi})^T B^{\phi} U^{\phi}|}{|(U^{\phi})^T W^{\phi} U^{\phi}|} = [u_1^{\phi}, ..., u_N^{\phi}] \tag{6}$$

The vectors, u^{ϕ}, can be found as the solution of the generalized eigenvalue problem i.e. $B^{\phi} u_i^{\phi} = \lambda_i W^{\phi} u_i^{\phi}$. The training vectors are supposed to be centered (zero mean, unit variance) in the feature space Z .from the theory of reproducing kernels any solution $u^{\phi} \in Z$ must lie in the span of all training samples in Z, i.e.

$$u^{\phi} = \sum_{c=1}^{C} \sum_{i=1}^{M_c} \alpha_{ci} \phi(x_{ci}) \tag{7}$$

Where α_{ci} are some real weights and x_{ci} is the ith sample of the class c. The solution is obtained by solving

$$\lambda = \frac{\alpha^T KDK \ \alpha}{\alpha^T KK \ \alpha} \tag{8}$$

Where $\alpha = (\alpha_c)$ c=1…C is a vector of weights with $\alpha = (\alpha_{ci}), i = 1...M_c$ .The kernel matrix $K(M \times M)$ is composed of the dot products of nonlinearly mapped data, i.e.

$$K = (K_{kl})_{k=1...C, l=1....C} \tag{9}$$

Where $K_{kl} = (k(x_{ki} x_{lj}))_{i=1...M_k, j=1....M_l}$ The matrix $D(M \times M)$ is a block diagonal matrix such that

$$D = (D_c)_{c=1...C} \tag{10}$$

Where the cth on the diagonal has all elements equal to 1/Mc.. Solving the eigenvalue problem yields the coefficient vector α that define the projection vectors $u^{\phi} \in Z$. A projection of a testing vector x_{test} is computed

as

$$(u)^T \phi(x_{test}) = \sum_{c=1}^{C} \sum_{i=1}^{M_c} \alpha_{ci} k(x_{ci}, x_{test}) \tag{11}$$

The procedure of the proposed algorithm for performing GDA could be summarized as follows:

- Compute the matrices K and D by solving the equation(9) and(10),
- Decompose K using eigenvectors decomposition,
- Compute eigenvectors α and eigenvalues of the equation(6),
- Compute eigenvectors u^{ϕ} using α_{ci} from equation (7) and normalize them,
- Compute projections of test points onto the eigenvectors u^{ϕ} from equation (11).

The input training data is mapped by a kernel function to a high dimensional feature space, where different classes is supposed to be linearly separable. The Linear Discriminant Analysis (LDA) [14] scheme is then applied to the mapped data, where it searches for those vectors that best discriminate among the classes rather than those vectors that best describe the data [15]. Furthermore, gives a number of independent features which describe the data, LDA creates a linear combination of the features that yields the largest mean differences to the desired classes [16] The number of original 41 features is reduced to 12 features by GDA as shown in the Table IV.

TABLE IV. FEATURES SELECTED BY GENERALIZED DISCRIMINANT ANALYSIS

S.No	Feature	Type
1	Service	Discrete
2	src_bytes	Continuous
3	dst_bytes	Continuous
4	logged_in	Discrete
5	Count	Continuous
6	srv_count	Continuous
7	serror_rate	Continuous
8	rv_rerror_rate	Continuous
9	srv_diff_host_rate	Continuous
10	dst_host_count	Continuous
11	dst_host_srv_count	Continuous
12	dst_host_diff_srv_rate	Continuous

TABLE V. CONFUSION MATRIX FOR SOM CLASSIFIER BY GDA TECHNIQUE.

Predicted / Actual	Normal	Probe	DOS	R2L	U2R	%Correct
Normal	57101	3769	275	1	9	94.23
Probe	1371	2670	360	3	4	64.1
DOS	11643	1105	224860	2	3	97.82
R2L	11562	3027	8	1956	1	12.08
U2R	99	67	8	1	55	24.12
%Correct	69.83	25.10	99.7	99.64	76.38	

TABLE VI. CONFUSION MATRIX FOR C4.5 CLASSIFIER BY GDA TECHNIQUE..

Predicted / Actual	Normal	Probe	DOS	R2L	U2R	%Correct
Normal	60400	151	38	1	3	99.68
Probe	40	4120	4	1	1	98.89
DOS	2028	160	228369	2	3	99.35
R2L	7468	984	1010	6726	1	41.5
U2R	96	53	4	1	74	32.45
%Correct	86.24	75.34	99.53	99.92	90.24	

The resulting confusion matrices of SOM and C4.5 classifiers are obtained as shown in the Table V and VI respectively. We obtain two reduced datasets by PCA and GDA techniques in addition to the original dataset as shown in Table VII.

TABLE VII. SUMMARY OF DATASET OBTAINED AFTER FEATURE EXTRACTION

Dataset Name	Features	Method
ORIGDATA	41	None
PCADATA	19	PCA
GDADATA	12	GDA

3.3. Experimental results

We will conduct two experiments one with Self Organizing Map (SOM) [17] and another with C4.5[18] for training and testing. There are approximately 4,94,020 kinds of data in training dataset and 3,11,029 kinds of data in test dataset of five classes (Normal, DOS,R2L,U2R and Probe). We choose 97277, 391458, 1126, 52 an d 4107 samples for Normal, DOS, R2L, U2R and Prob respectively to train the PCA and

proposed GDA and then used test data 60593, 229853, 16189, 228, and 4166 for Normal, DOS, R2L, U2R and Prob respectively to compare the training and testing time and recognition rate. Each sample vector is of dimensionality 41. We use Gaussian kernel $k(x, y) = \exp(-\|x - y\| / 0.1)$ to calculate the kernel matrix. All these experiments are run on the platform of Windows XP with 2.0GHz CPU and 1GB RAM by Weka3.5.8 software to implement the proposed technique. The detection and identification of attack and non-attack behaviours can be generalized as follows:

True Positive (TP): the amount of attack detected when it is actually attack.

True Negative (TN): the amount of normal detected when it is actually normal.

False Positive (FP): the amount of attack detected when it is actually normal (False alarm).

False Negative (FN): the amount of normal detected when it is actually attack.

Confusion matrix contains information actual and predicted classifications done by a classifier. In the performance of such a system is commonly evaluated using the data in a matrix. Table VIII shows the confusion matrix.

TABLE VIII. CONFUSION MATRIX

Predicted / Actual	Normal	Attack
Normal	True Negative (TN)	False Pasitive (FP)
Attack	False Negative (FN)	True Positive (TP)

In the confusion matrix above, rows correspond to predicted categories, while columns correspond to actual categories.

Comparison of detection rate : Detection Rate (DR) is given by.

$$DR = \frac{TP}{TP + FN} \times 100\ \%$$

Comparison of false alarm rate : False Alarm Rate (FAR) refers to the proportion that normal data is falsely detected as attack behaviour.

$$FAR = \frac{FP}{FP + TN} \times 100\ \%$$

The reported results in term of detection rate, false alarm rate, training time and testing time of SOM and C4.5 decision tree classifiers are summarized in Tables IX, X.

TABLE IX. DETECTION RATE, FALSE ALARM RATE, TRAINING TIME AND TESTING TIME OF SOM AND C4.5 CLASSIFIER WITH PCA TECHNIQUE

	SOM				C4.5			
	DR	**FAR**	**TR. TIME**	**TE. TIME**	**DR**	**FAR**	**TR. TIME**	**TE. TIME**
Normal	93.3	27.4	45s	31s	99.3	15.4	41s	30s
Probe	60.2	76.3	16s	15s	96.2	26.7	17s	16s
DOS	95.7	0.3	56s	27s	98.5	0.9	52s	27s
R2L	10.8	0.4	16s	14s	38.0	0.1	16s	13s
U2R	20.6	29.2	11s	10s	32	8.7	10s	9s

DR-detection rate, FAR-false alarm rate, TR- training, TE-testing

TABLE X. DETECTION RATE, FALSE ALARM RATE, TRAINING TIME AND TESTING TIME OF SOM AND C4.5 CLASSIFIER WITH GDA.TECHNIQUE

	SOM				C4.5			
	DR	**FAR**	**TR. TIME**	**TE. TIME**	**DR**	**FAR**	**TR. TIME**	**TE. TIME**
Normal	94.23	30.2	37s	26s	99.68	13.7	32s	23s
Probe	64.1	74.9	14s	13s	98.89	24.7	13s	11s
DOS	97.82	0.3	48s	24s	99.35	0.5	45s	22s
R2L	12.08	0.36	13s	11s	41.5	.08	12s	9s
U2R	24.12	23.6	8s	7s	32.45	9.7	7s	6s

4. Conclusion

We have presented an efficient technique for performing GDA especially for the case of large scale dataset where the number of training samples is large. GDA gives better detection rate, less false positives, reduced training and reduced testing times than PCA for the both classifiers. Moreover, when we compared two classifiers, the C4.5 classifier shows better performance for all the classes (Normal, DOS, R2L, U2R, Prob,) and comparables training and testing times as shown in Table IX and X.

Dataset KDDCup99 applied in the research paper is popularly used in current cyber attack detection system; however, it is data of 1999 and network technology and attack methods changes greatly, it can not reflect real network situation prevailing nowadays. Therefore, if newer information is available and tested and compared, they can more accurately reflect current network situation.

We propose ensemble approach for intrusion detection system in which Generalized Discriminant Analysis (GDA) is used as feature reduction technique and C4.5 as an intrusion detection classifier for future research.

5. References

[1] Dasarathy, B.V.: Intrusion Detection, *Information Fusion*. (4) 243-245, 2003.

[2] *American Computer Emergency Response Team /Coordination Center* (CERT),http://www.cert.org/.

[3] *Information Security* Report, http://www.isecu-tech.com.tw/.

[4] Bace, R.G.: *Intrusion Detection*. Macmillan Technical Publishing. 2000.

[5] H. Debar etal. "Towards a taxonomy of intrusion detection systems" *Computer Network*,pp.805-822, April1999.

[6] KDDCup99dataset,August2003 http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

[7] R.O.Duda, P.E.Hart, and D.G.Stork, *Pattern Classification*, vol. 1. New York: Wiley, 2002.

[8] A.K.Jain, R.P.W.Duin, and J.Mao, "Statistical Pattern Recognition: A Survey," *IEEE Transactions on Pattern Analysis and Mission Intelligence*, vol. 22, pp.4-37, January 2000. Computer Science

[9] Jolliffe, I.T. *Principal Component Analysis*. Springer-Verlag, NY, 2002.

[10] W. Wang, X. Guan and X. Zhang. "A Novel Intrusion Detection Method Based on Principal Component Analysis in Computer Security", *Advances in Neural Networks*-ISNN2004,

[11] G.Baudt and F. Anouar "Generalized Discriminant Aanlyis Using a Kernal Approach" *Neural Computation*, 2000

[12] Gopi K. Kuchimanchi,Vir V. Phoha, Kiran S.Balagani, Shekhar R. Gaddam, Dimension Reduction Using Feature Extraction Methods for Real-time Misuse Detection Systems, *Proceedings of the IEEE on Information*, 2004

[13] Kemal Polat,et.al.. A cascade learning system for classification of diabetes disease: Generalized Discriminant Aanalysis and Least Square Support Vector Machine. *Expert Systems with Applications* 34 pp-482-487. 2008.

[14] K. Fukunaga. *Introduction to Statistical Pattern Classification*. Academic Press, San Diego,California, USA, 1990

[15] Kim HC et al. Face recognition using LDA mixture model. In: *Proceedings int conf. on pattern recognition*, 2002.

[16] Martinez AM, Kak AC. PCA versus LDA. *IEEE Trans Pattern Anal Mach Intel*; 23(2):228-33, 2001.

[17] Kohan T. *Self-Organizing and Associative Memory*.2nd Edition, Springer Vela, NY.

[18] J.R. Quinlan, *C4.5 Programs for machine learning* Morgan Kaufmann 1993.