Principle Components Analysis and Support Vector Machine based Intrusion Detection System

Heba F. Eid[1], Ashraf Darwish[2], Aboul Ella Hassanien[3], and Ajith Abraham[4]

[1] *Faculty of Science, Al-Azhar University. Cairo, Egypt*
Email: heba.fathy@yahoo.com
[2] *Faculty of Science, Helwan University, Cairo, Egypt*
Email: amodarwish@yahoo.com
[3]*Faculty of Computers and Information, Cairo University,Cairo, Egypt*
Email: aboitcairo@gmail.com
[4] *Machine Intelligence Research Labs, MIR Labs, USA*
Email: abraham.ajith@gmail.com

***Abstract*—Intrusion Detection System (IDS) is an important and necessary component in ensuring network security and protecting network resources and infrastructures. In this paper, we effectively introduced intrusion detection system by using Principal Component Analysis (PCA) with Support Vector Machines (SVMs) as an approach to select the optimum feature subset. We verify the effectiveness and the feasibility of the proposed IDS system by several experiments on NSL-KDD dataset. A reduction process has been used to reduce the number of features in order to decrease the complexity of the system. The experimental results show that the proposed system is able to speed up the process of intrusion detection and to minimize the memory space and CPU time cost.**

***Keywords*-Network security; Intrusion detection system; Feature selection; Support Vector Machines (SVMs); Principal component analysis(PCA).**

I. INTRODUCTION

The concept of intrusion detection (ID) was proposed by Anderson in 1980 [1]. Intrusion detection system (IDS) dynamically monitors the events taking place in a system, and decides whether these events are symptomatic of an attack or constitute a legitimate use of the system [2]. In general, IDS techniques are classified into two categories: anomaly detection and misuse detection. Misuse intrusion detection uses well-defined patterns (signatures) of the malicious activity to identify intrusions [3,4]. Anomaly intrusion detection identifies malicious traffic based on deviations from the normal usage behavior patterns, where the normal patterns are constructed from the statistical measures of the system features [5]. IDS is a valuable tool for the defense-in-depth of computer networks. Network-based IDS looks for known or potential malicious activities in network traffic and raise an alarm whenever a suspicious activity is detected. Several machine-learning techniques including neural networks [6], support vector machines (SVM) [5], fuzzy logic [7] have been studied for the design of IDS.

In general, IDS deals with huge amount of data even for a small network, which contains irrelevant and redundant features. Extraneous features can make it harder to detect suspicious behavior patterns, causing slow training and testing process, higher resource consumption as well as poor detection rate [8].

Feature selection is one of the key topics in IDS, it improves classification performance by searching for the subset of features, which best classifies the training data [9]. In problem of high dimensional feature space, some of the features may be redundant or irrelevant. Removing these redundant or irrelevant features is very important; hence they may deteriorate the performance of classifiers. Feature selection involves finding a subset of features to improve prediction accuracy or decrease the size of the structure without significantly decreasing prediction accuracy of the classifier built using only the selected features [10]. This is very important if real-time detection is desired. Principal component analysis (PCA) is an essential technique in data compression and feature selection [11] which has been applied to the field of ID [12, 14]. PCA [15] is an efficient method to reduce dimensionality by providing a linear map of n-dimensional feature space to a reduced m-dimensional feature space. In this paper, PCA is applied for feature dimension reduction.

In this paper, we evaluate the NSL-KDD dataset and we propose an anomaly intrusion detection system based on SVM, where PCA is applied for feature selection. We examine the effectiveness of our system by conducting several experiments on NSL-KDD dataset. The rest of this paper is organized as follows. We outline mathematical overview of support vector machine methods and PCA in Section II. Section III describes the proposed IDS system.

978-1-4244-8136-1/10/$26.00 ©2010 IEEE

The experimental results and conclusions are presented in Section IV and V respectively.

II. Machine Intelligence: An overview

A. Support Vector Machines

Compared with conventional machine learning methods SVMs have some advantages [16, 17]:

1) There are only two free parameters to be chosen, namely the upper bound and the kernel parameter.
2) The solution of SVM is unique, optimal and global since the training of a SVM is done by solving a linearly constrained quadratic problem.
3) Good generalization performance and Good robustness. Because of the above advantages, SVM has been recently used in many applications.

Support vector machine (SVM) approach is a classification technique based on Statistical Learning Theory (SLT). It is based on the idea of a hyper plane classifier, or linearly separability. The goal of SVM is to find a linear optimal hyper plane so that the margin of separation between the two classes is maximized [18,19]. Suppose we have N training data points $\{(x_1, y_1), (x_2, y_2), (x_3, y_3), ..., (x_N, y_N)\}$, where $x_i \in R^d$ and $y_i \in \{+1, -1\}$. Consider a hyper plane defined by (w, b), where w is a weight vector and b is a bias. A new object x can be classify with the following function:

$$f(x) = sign(w.x + b) = sign(\sum_{i=1}^{N} \alpha_i y_i (x_i, x) + b) \quad (1)$$

In practice, the data is often not linearly separable. However, one can still implement a linear model by transform the data points via a non-linear mapping to another higher dimensional space (feature space) such that the data points will be linear separable. This mapping is done by a kernel function K.

The nonlinear decision function of SVM is given by the following function:

$$f(x) = sign(\sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b) \quad (2)$$

where $K(x_i, x)$ is the kernel function.

B. Principle Components Analysis

It is well known that principal component analysis (PCA) is an essential technique in data compression and feature extraction [11], and it has been also applied to the field of ID [12-14]. It is well known that PCA has been widely used in data compression and feature selection. Feature selection refers to a process whereby a data space is transformed into a feature space, which has a reduced dimension. Some basic knowledge of PCA is briefly described in the next.

Assume that $\{x_t\}$ where $t = 1, 2 \ldots, N$ are stochastic n-dimensional input data records with mean (μ). It is defined by the following Equation:

$$\mu = \frac{1}{N} \sum_{t=1}^{N} x_t \quad (3)$$

The covariance matrix of x_t is defined by

$$C = \frac{1}{N} \sum_{t=1}^{N} (x_t - \mu).(x_t - \mu)^T \quad (4)$$

PCA solves the following eigenvalue problem of covariance matrix C:

$$Cv_i = \lambda_i v_i \quad (5)$$

where λ_i $(i = 1, 2, ..., n)$ are the eigenvalues and $v_i (i = 1, 2, ..., n)$ are the corresponding eigenvectors.

To represent data records with low dimensional vectors, we only need to compute the m eigenvectors (called principal directions) corresponding to those m largest eigenvalues $(m < n)$. It is well known that the variance of the projections of the input data onto the principal direction is greater than that of any other directions.

Let

$$\phi = [v_1, v_2, ..., v_m], \Lambda = diag[\lambda_1, \lambda_2, ..., \lambda_m] \quad (6)$$

Then

$$C\Phi = \Phi\Lambda \quad (7)$$

The parameter v denote to the approximation precision of the m largest eigenvectors so that the following relation holds.

$$\frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{n} \lambda_i} \geq v \quad (8)$$

Based on (7) and (8) the number of eigenvectors can be selected and given a precision parameter v, the low-dimensional feature vector of a new input data x is determined by

$$x_f = \Phi^T x \quad (9)$$

III. The proposed Intrusion Detection System

The proposed intrusion detection system is composed of the following three phases.

- Preprocessing: contains mapping symbolic valued attributes to numeric, scaling data and attack names
- Feature selection: select the optimum feature subset
- Intrusion detection: classify the intrusion type into its classes

Figure (1) shows the overall description of the proposed intrusion detection system.

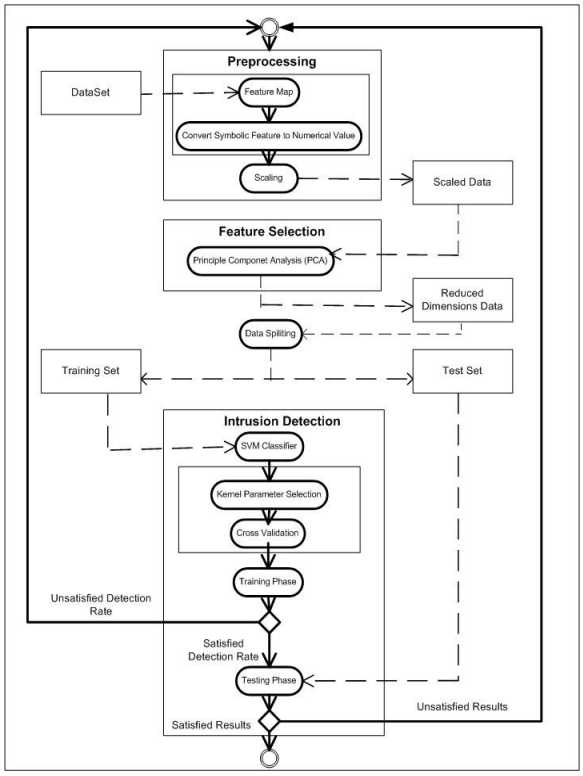

Figure (1) The proposed Intrusion Detection System

A. Preprocessing phase

SVM classification system are not able to process NSL-KDD dataset in its current format. Hence preprocessing was required before SVM classification system could be built. Preprocessing contains the following processes:

- Mapping symbolic features to numeric value.
- Implementing scaling since the data have significantly varying resolution and ranges. The attribute data are scaled to fall within the range [-1, 1].
- Attack names were mapped to one of the five classes, 0 for $Normal$, 1 for DoS (Denial of Service), 2 for $U2R$ (user-to-root: unauthorized access to root privileges) , 3 for $R2L$ (remote-to-local: unauthorized access to local from a remote machine), and 4 for $Probe$ (probing:information gathering attacks) .

B. Feature selection phase

PCA reduces the amount of dimensions required to classify new data and produces a set of principal components, which are orthonormal eigenvalue/eigenvector pairs [20]. It reduces the dimensionality of data by restricting attention to those directions in the feature space in which the variance is greatest. The proportion of the total variance accounted for a feature is proportional to its eigenvalue [15]. We apply PCA on NSL-KDD dataset to reduce the dimensions (D) of the data where D = 41. We use the Critical Eigenvalue test and screeplot test [15] to determine the k features that are required for classification. The Critical Eigenvalue test recommends selecting all principal components whose eigenvalues exceed the threshold $\frac{D^{0.6}}{15}$, where D is the dimension of the original dataset. The remaining d-k features are assumed to contain noise or redundancy. For NSL-KDD dataset we select 23 features which reduce the input data dimension by 56%.

C. Intrusion Dtection:Parameter selection

The radial basis function kernel (RBF) is used within this study, mainly for some reasons [21]:

- RBF makes it possible to map the non-linear boundaries of the input space into a higher dimensional feature space unlike the linear kernel.
- The RBF kernel has less numerical difficulties because the kernel values lie between zero and one, while the polynomial kernel values may go to infinity or zero while the degree is large.
- When looking at the number of hyper parameters, the polynomial kernel has more hyper parameters than the RBF kernel. On the basis of these arguments, the RBF kernel is used as the default kernel function.

In this study, a K-fold cross-validation is used on the dataset to search best values of RBF kernel width parameter γ and constant C. The search is realized by evaluating exponential sequences of γ and C.

IV. Experimental results and analysis

A. Dataset characteristics

Under the sponsorship of Defense Advanced Research Projects Agency (DARPA) and Air Force Research Laboratory (AFRL), MIT Lincoln Laboratory has collected and distributed the datasets for the evaluation of researches in computer network intrusion detection systems [22]. The KDD'99 dataset is a subset of the DARPA benchmark dataset prepared by Sal Stofo and Wenke Lee [23].

KDD'99 train dataset is about four gigabytes of compressed binary TCP dump data from seven weeks of network traffic, processed into about five million connection Record each with about 100 bytes. The two weeks of test data have around two million connection records. Each KDD'99 training connection record contains 41 features (e.g., protocol type, service, and flag) and is labeled as either normal or an attack, with exactly one specific attack type. The training set contains a total of 22 training attack types, with an additional 17 types in the testing set only. The attacks fall into four categories:DoS e.g Neptune, Smurf, Pod and Teardrop, R2L e.g Guess-password, Ftp-write, Imap and Phf, U2R e.g Buffer-overflow, Load-module, Perl and Spy, and Probing eg. Port-sweep, IP-sweep, Nmap and Satan.

Statistical analysis on KDD'99 dataset found important issues which highly affects the performance of evaluated

Table I
KDD'99 DATASET REDUCTION STATISTICS

	Train set		Test set	
	Repeated records	Reduction rate	Repeated records	Reduction rate
Intrusion	3663472	93.32%	221058	88.26%
Normal	159967	16.44%	12680	20.92%

Table II
KDD'99 AND NSL-KDD DATASET TESTING ACCURACY COMPARISON

Class name	Original KDD'99 test set Test Accuracy	NSL-KDD test set Test Accuracy
Normal	99.8%	99.5%
DoS	92.5%	97.5%
U2R	5.1%	86.6%
R2L	70.2%	81.3%
Probe	98.3%	92.8%

Table III
TESTING ACCURACY COMPARISON

Class name	SVM system with 41-dimension feature Test Accuracy	SVM system with 23-dimension feature Test Accuracy
Normal	99.8%	99.5%
DoS	97.5%	99.9%
U2R	86.6%	81.2%
R2L	81.3%	54.6%
Probe	92.8%	95.3%

systems and results in a very poor evaluation of anomaly detection approaches. Leung and Leckie [24] reported two problems in the KDD'99 dataset. First, KDD'99 dataset contains huge number of redundant records.10% portions of the full dataset contained only two types of DoS attacks (Smurf and Neptune).These two types constitute over 71% of the testing dataset which completely affects the evaluation. Secondly, since these attacks consume large volumes of traffic, they are easily detectable by other means and there is no need of using anomaly detection systems to find these attacks.

To solve these issues, NSL-KDD a new dataset is suggested [25]. NSL-KDD consists of selected records of the complete KDD'99 dataset,where the repeated records in the entire KDD'99 train and test set are removed.KDD'99 dataset contains 4898431 records in train set and 311027 records in test set.Table I gives statistics of the reduction of repeated records in the KDD train and test sets.The KDD'99 train set is reduced by 78.05% and the test set is reduced by 75.15%.

NSL-KDD dataset has the following advantages over the original KDD'99 dataset [25]:

1) The train set does not include redundant records; hence the classifiers will not be biased towards more frequent records.
2) The proposed test sets have no duplicate records; therefore, the performances of the learners are not biased by the methods which have better detection rates on the frequent records.
3) The number of records in the train and test sets is reasonable, which makes it affordable to run the experiments on the complete set without the need to randomly select a small portion. Consequently, evaluation results of different research works will be consistent and comparable.

B. Experiments and analysis

Case one: to compare the evolution performance of the two dataset KDD'99 dataset and NSL- KDD dataset. We applied the SVM intrusion detection system on two test sets; the original KDD'99 test set (KDDTest) and NSL-KDD test set (KDDTest+). All experiments were performed using Intel Core 2 Duo 2.26 GHz processor with 2 GB of RAM.

Tables II, the redundant records in the original KDD'99 cause the learning algorithms to be biased towards the frequent records (DOS and Prob attacks). Thus prevent the detection algorithms from learning unfrequented records (U2R and R2L attacks).These unbalanced distribution of the KDD'99 testing dataset completely affects the evaluation of the detection algorithms. NSL-KDD test sets have no redundant records; hence, the performances of the learners are not biased by frequent records.

Case two: the NSL- KDD dataset are taken to evaluate the proposed SVM intrusion detection system with PCA for dimension reduction. We have randomly taken 10776 training records and 7970 test records. The performance of the proposed classifiers system includes testing accuracy and the processing speed during testing. The experimental results are shown in Table III and IV.

Table III shows the accuracy achieved for SVMs using full dimension data (without PCA) and after the features reduction (with PCA). The testing accuracies indicate that PCA can be used to reduce data dimension without sacrificing much performance in accuracy.

Table IV, illustrate that SVMs system with PCA will improve training and testing speed, which is important for real time network applications. It is clear that the proposed SVMs system with PCA faster in training and testing than SVMs without PCA.

V. CONCLUSIONS AND FUTURE WORKS

In this paper, we test the new dataset NSL-KDD which solve important issues of KDD'99 dataset. The experiments show that, NSL-KDD dataset can be applied as an effective benchmark dataset to help researchers compare different intrusion detection models. We proposed a PCA feature

Table IV
TESTING ACCURACY COMPARISON

	Train time (ms)	Test time (ms)
SVM system with 41-dimension feature	4.5	3.1
SVM system with 23-dimension feature	1.7	1.3

selected SVM intrusion detection system. PCA algorithm was used in order to select a best subset of features for classifying. We build SVM system to evaluate the selected subset. We developed a series of experiments on NSL-KDD dataset to examine the effectiveness of our building IDS. The experiment results show that our system is able to speed up the training and testing process of intrusions detection which is important for high-speed network applications. In our future work, we will extend our SVM system to build an efficient intrusion detection system.

VI. Acknowledgements

The authors thank Dr. Afaf Abo El-Ftouh Saleh, Professor of mathematics, Faculty of Science, Al-AZhar University for supporting this research.

References

[1] J.P. Anderson, "Computer security threat monitoring and surveillance",Technical Report, James P. Anderson Co., Fort Washington, PA, April 1980.

[2] H. Debar, M. Dacier and A. Wespi, "Towards a taxonomy of intrusion-detection systems" Computer Networks, vol. 31, pp. 805-822, 1999.

[3] K. Ilgun, R.A. Kemmerer and P.A. Porras, "State transition analysis:A rule-based intrusion detection approach" IEEE Trans. Software Eng. vol. 21, pp. 181-199, 1995.

[4] D. Marchette, "A statistical method for profiling network traffic". In proceedings of the First USENIX Workshop on Intrusion Detection and Network Monitoring (Santa Clara), CA. pp. 119-128, 1999.

[5] S. Mukkamala, G. Janoski and A.Sung, "Intrusion detection: support vector machines and neural networks" In proceedings of the IEEE International Joint Conference on Neural Networks (ANNIE), St. Louis, MO, pp. 1702-1707, 2002.

[6] Z. Zhang, J. Li, C. N. Manikopoulos, J. Jorgenson, and J. Ucles, "HIDE: a hierarchical network intrusion detection system using statistical preprocessing and neural network classiffcation" In proceedings of the 2001 IEEE Workshop on Information Assurance,(New York), pp. 85-90, IEEE Press, June 2001.

[7] S. Wu and W. Banzhaf, "The use of computational intelligence in intrusion detection systems: A review", Applied Soft Computing, vol.10, pp. 1-35, 2010.

[8] W. Lee, S. Stolfo and K. Mok, "A data mining framework for building intrusion detection models" In proceedings of the IEEE symposium on security and privacy; pp.1999.

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

[10] D. Koller, and M. Sahami, "Toward optimal feature selection" In proceedings of international conference on machine learning, Bari, (Italy) pp. 284-92, 1996.

[11] E. Oja, "Principal components, minor components, and linear neural networks" Neural Networks, vol. 5, pp. 927-935, 1992.

[12] G.K. Kuchimanchi, V.V. Phoha, K.S. Balagami and S.R. Gaddam, "Dimension reduction using feature extraction methods for Real-time misuse detection systems" In proceedings of the IEEE Workshop on Information Assurance and Security, West Point, (New York), pp. 195-202, 2004.

[13] K. Labib and V.R. Vemuri, "Detecting and visualizing denial-of-service and network probe attacks using principal component analysis" In Third Conference on Security and Network Architectures, La Londe, (France), 2004.

[14] M. Shyu, S. Chen, K. Sarinnapakorn and L. Chang, "A Novel Anomaly Detection Scheme Based on Principal Component Classifier" In Proceedings of ICDM'03, pp. 172-179, 2003.

[15] E. E. Cureton and R. B. D'Agostino, "Factor Analysis: An Applied Approach", London: Lawrence Erlbaum Associates, vol. I, 1983.

[16] S. Kim, K. S. Shin and K. Park, "An application of support vector machines for customer churn analysis: Credit card case" Lecture Notes in Computer Science, vol. 3611, pp. 636-647, 2005.

[17] S. Kim, S. Yang and K. S. Seo, "Home photo categorization based on photographic region templates" Lecture Notes in Computer Science, vol. 3689, pp. 328-338, 2005.

[18] V. Vapnik, "Statistical learning theory" New York: Wiley, 1998.

[19] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition", Data Mining and Knowledge Discovery, vol. 2, pp.121-167,1998.

[20] I. T. Jolliffe, "Principal Component Analysis" Springer-Verlag, (New York), 2002.

[21] C.W. Hsu, C. C. Chang and C.J. Lin. "A practical guide to support vector classification". Technical Report, Department of Computer Science and Information Engineering, National Taiwan University, 2004.

[22] MIT Lincoln Laboratory, DARPA Intrusion Detection Evaluation, http://www.ll.mit.edu/CST.html,MA, USA. July, 2010.

[23] KDD'99 dataset, http://kdd.ics.uci.edu/databases, Irvine, CA, USA, July, 2010.

[24] K. Leung and C. Leckie, "Unsupervised anomaly detection in network intrusion detection using clusters", In Proceedings of the Twenty-eighth Australasian conference on Computer Science, vol. 38, pp. 333- 342, 2005.

[25] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set" In Proceeding of the 2009 IEEE symposium on computational Intelligence in security and defense application (CISDA), 2009.