International Arab Conference on Information Technology (ACIT'2016)

Anomaly Traffic Detection Based on GPLVM and SVM

Houda Moudni[1], Mohamed Er-rouidi[1], Hassan Faouzi[1], Hicham Mouncif[2], and Benachir El Hadadi[2]
[1]Faculty of Sciences and Technology, Sultan Moulay Slimane University Beni Mellal, Morocco
[2]Faculty Polydisciplinary, Sultan Moulay Slimane University Beni Mellal, Morocco

Abstract:*An Intrusion Detection System (IDS) is a type of security management technique designed to automatically detect possible intrusions in a network or in a host. However, the performance of existing IDSs that have been proposed are unsatisfactory since novel kinds of attacks appear constantly. Hence, it is necessary to develop an intrusion detection system with improved accuracy, high detection rate and reduced training time. We thus propose a new adaptive intrusion detection method based on Gaussion Process Latent Variable Model (GPLVM) and Support Vector Machines (SVMs). In this paper, the GPLVM is applied for feature dimension reduction of network connection records. After the step of dimensionality reduction, samples are classified using SVMs algorithm in order to distinguish between the normal and the intrusions in the traffic data. The results of our experiment based on KDD Cup'99 dataset illustrate the effectiveness of our proposed method. A comparative analysis of intrusion detection system based on GPLVM and Artificial Neural Network (ANN) with our building IDS is also performed.*

Keywords:*IDS, network security, GPLVM, SVM, KDD Cup 99 Dataset.*

1. Introduction

In recent years, security problems in computer network systems have become more and more worldwide preoccupation, since various intrusions and other types of attacks [1] may cause significant losses to information assets. To defend these attacks, several protective measures such as firewalls, intrusion detection systems (IDSs), etc., have been developed. Among these security mechanisms, intrusion detection system plays a crucial role because it is dynamic defense technique.

Generally, intrusion detection systems can be classified into two major categories which are namely misuse or signature detection and anomaly detection. The misuse or signature based intrusion detection system can detect known attacks by comparing the data with predefined signature of known intrusion. While anomaly detection system defines a profile of normal activity as a model, then tries to check any pattern that deviates from the model and it is considered as an anomaly.

However, a major problem in intrusion detection is that new attacks cannot be detected by the misuse detection technique. In contrast, the anomaly detection approach is able to detect not only known intrusion but also unknown intrusion. Furthermore, this approach has drawbacks such as a significant number of false alarms the difficulty of handling gradual misbehavior and the detection rate is quite satisfactory.

To perform anomaly detection, many methods have been proposed based on machine learning or data mining. For example, the authors in [2] propose a new hybrid method by using Principal Component Analysis (PCA) and Simplified Fuzzy Adaptive resonance theory Map (SFAM) to improve anomaly detection performances. The results of their experiment based on KDD Cup'99 dataset show that their method improves classification performance of R2L attack significantly comparing to other techniques mentioned in the paper. In [3], the authors have proposed an efficient Back Propagation Neural Network (BPNN) architecture for the development of an anomaly based IDS with high accuracy and detection rate. Zhang et al. [4] have proposed a Fuzzy C-Means Clustering based on Particle swarm Optimization for intrusion detection system. The main advantage of their algorithm is that it uses clustering to identify anomaly intrusion and classification to find intruder.

This paper proposes a novel intrusion detection method based on Gaussion Process Latent Variable Model (GPLVM) [5] and Support Vector Machines (SVMs) [6]. In the proposed method, GPLVM is used to reduce the feature dimension of network connection records and SVMs are used for classification. We

tested the effectiveness of our system with the help of conducting several experiments on KDD-Cup99 dataset.

The remainder of this paper is organized as follows. Section 2 is dedicated to presents briefly the KDD Cup'99 dataset. Section 3 discusses our proposed architecture. In section 4, the performance of our work along with the comparison of it with an IDS based on GPLVM+ANN is discussed. Finally, Section 5 concludes the paper and outlines the future works.

2. KDD Cup 99 Dataset

The dataset used in our experiments is KDD Cup 99 dataset [7], a most wildly used dataset for the evaluation of anomaly detection methods. It was created based on the data captured in DARPA'98 IDS evaluation program in a MIT Lincoln Laboratory. The dataset contains 42 features including class label which specifies the status of each connection record as either normal or specific attack type. The attack types are grouped into four main categories:

- **Denial of Service Attack (DoS):** in this attack, the attacker makes the traffic too busy over the network in order to consume network bandwidth or to consume the resources of a particular node.
- **Probe Attack:** is an attack in which the attacker attempts to gain information about the target host. The purpose of this attack is to get knowledge about available services in order to make some attacks in future.
- **Remote to Local Attack (R2L):** occurs when an attacker sends packets to a machine over a network since who does not have a local account on that machine. Then, exploits the machine's vulnerability to gain local access as a user.
- **User to Root Attack (U2R):** the attacker has local access to a normal user account on the system by compromising it through sniffing password and tries to gain super user privileges.

3. The Proposed Method

Our proposed method is based on three major steps: the first step is data pre-processing that converts and normalize data into appropriate format. The second step is the dimensionality reduction with Gaussion Process Latent Variable Model. The last step is to classify different group of normal and intrusion data by Support Vector Machine algorithm. Fig. 1 shows the overall description of our proposed intrusion detection system. The steps in our proposed system are discussed in detail below.

3.1. Data Pre-Processing

Three pre-processing operations are applied to the original dataset:

Encoding: the original input dataset used contains eight nonnumeric attributes, such as the feature number 2 "protocol_type", that has three different symbols "TCP", "UDP", "ICMP", feature number 3 "service" with 66 values as "HTTP", "SMTP", "TELNET", etc, and other features. The first step is to assign numeric values to these features in order to make the processing task much simpler, as numerical values can be easily operated upon.

For converting symbols into numerical format, an integer value is assigned to each symbol. For instance, in the case of "protocol_type" feature, 1 is assigned to "TCP", 2 to "UDP", and 3 to the "ICMP" symbol. The same procedure for other features.

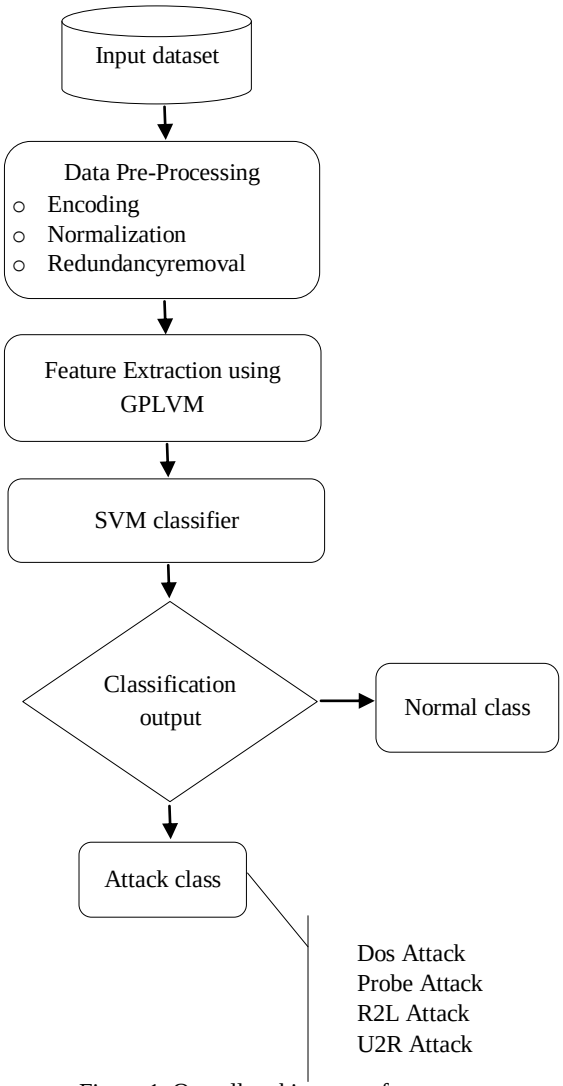

Figure 1. Overall architecture of system

For attack names, they are assigned to one of the five classes, 1 for Normal, 2 for DoS attack, 3 for

Probe attack, 4 for R2L attack, and 5 for U2R attack, according to the Table 1 below.

Table 1. Different attacks falling into four major categories

DOS	Probe	R2L	U2R
back land neptune pod smurf teardrop	ipsweep nmap portsweep satan	ftp_write guess_passwd imap multihop phf spy warezclient warezmaster	buffer_overflow loadmodule perl rootkit

Normalization: After encoding the feature values, it is required to normalize the dataset in order to minimize the complexity involved in dealing with values spread over wide ranges. The attributes are scaled to the range [0,1] according to Equation 1.

$$X = (x - Min)/(Max - Min) \quad (1)$$

Where:

Min: minimum value of the attribute,
Max: maximum value of the attribute,
x: value of the attribute before normalization,
X: value of the attribute after normalization.

Redundancy removal: One of the major limitation with KDD Cup 99 dataset is the huge number of redundant records, which causes the evaluation results to be biased towards the frequent records. After redundancy removal, it has been found that most of the redundant records are present in the anomaly class than in the normal class.

3.2. Feature Extraction

Feature extraction is one of the most important tasks that improve the performance of classification by removing the irrelevant features. This is a dimensionality reduction process. In this paper we use the Gaussion Process Latent Variable Model (GPLVM) algorithm to select suitable subset of features from the dataset which are useful in detecting the intrusion.

The main idea behind the GPLVMs is to find a non-linear function that smoothly maps low-dimensional latent space vectors to a high dimensional observation space [8]. GPLVM is a probabilistic non-linear Principal Component Analysis (PCA) [9] which instead of applying linear transformation, using non-linear methods. It combines Gaussian Process (GP) and a linear transformation method such as PCA, Kernel PCA [10], Probabilistic PCA [11] and Dual Probabilistic PCA [12] to reduce the data dimensionality.

Let Y = [y_1, ..., y_N]^T be a matrix representing the training data, where $y_i \in R^D$, and X = [x_1, ..., x_N]^T denotes the matrix whose rows represent the corresponding positions of Y in latent space, $x_i \in R^Q$ with $Q < D$. The GPLVM relates a high-dimensional data set Y, and a low dimensional latent space X, using a Gaussian Process mapping from the latent space to high-dimensional observation data space.

Using the covariance function for the Gaussian process $K(x_i, x_j)$, the likelihood of the observed data given the low-dimensional latent coordinates, is

$$p(Y|X,\theta) = \frac{1}{\sqrt{(2\pi)^{ND}|K|^D}} \exp\left(-\frac{1}{2}tr(K^{-1}YY^T)\right) \quad (2)$$

Where K is the kernel matrix with the elements which defined by a kernel mapping function $K_{i,j} = k(x_i, x_j)$, and $tr()$ denotes the trace of a matrix. The covariance function is chosen as the sum of the Radial Basis Function (RBF) kernel, and the bias or constant term and noise term.

$$k(x_i, x_j) = \theta_1 \exp\left(-\frac{\theta_2}{2}\|x_i - x_j\|^2\right) + \theta_3 + \frac{\delta_{i,j}}{\theta_4} \quad (3)$$

Where $\theta = \{\theta_1, \theta_2, \theta_3, \theta_4\}$ are the kernel parameters, and $\delta_{i,j}$ is the Kronecker delta function. The latent space X is obtained by using the posterior distribution function.

$$p(X, \theta|Y) \propto p(Y|X,\theta)p(X) \quad (4)$$

The learning of the latent space consists of minimizing the negative log likelihood of the posterior in (4), with respect to the latent space configuration, X, and the hyper parameters, θ

$$L(X,\theta) = -\frac{DN}{2}\log 2\pi - \frac{D}{2}\log|K| - \frac{1}{2}tr(K^{-1}YY^T) \quad (5)$$

3.3. Classification

The purpose of this phase is to classify patterns of the system behavior in two categories (normal and intrusion) using patterns of known attacks that belongs to the abnormal class, and patterns of the normal behavior. Our proposed classifier in this work is Support Vector Machine (SVM) which gives good accurate results.

SVM approach is a supervised learning technique mainly used for classification. Originally, it was developed for the classification of two linear classes (SVM binary model) with a margin, the margin means that the minimum distance from the separating hyperplane to the closest data points. The SVM learning machine seeks for an optimal separating

hyperplane, where the margin of separation between the two classes is maximal. Therefore, in the case of the data points are not linearly separable, SVM uses suitable non-linear method (kernel function) to map them into higher dimensional space. These kernels allow to produce the maximum margin in a transformed feature space, such that the data points will be linear separable.

Here are some commonly used kernel functions by SVM:

Normal kernel:

$$k(x, y) = x^T y + c \quad (6)$$

Sigmoid kernel (also known as hyperbolic tangent kernel):

$$k(x, y) = \tanh(ax^T y + c) \quad (7)$$

Radical Basis Function (RBF) kernel (also known as Gaussian kernel):

$$k(x, y) = exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \quad (8)$$

Polynomial kernel:

$$k(x, y) = (ax^T y + c)^d \quad (9)$$

SVM method is destined for binary classification. However, in the case of a multi-class classification there are two approaches: the "one against the rest" approach and the "one against one" [13].

- One against the rest: This is a simpler approach, it considers the problem as a binary problem and teaches N decision function f that allows to discriminate against a class of the N classes, to classify an individual it presents all the classifiers and the membership class obtained depends on the highest value returned by the classifier.
- One against one: In this approach, it is about to discriminate against a class of another, then N-1/2 functions of decision will be learned and each performs a vote allocation a new individual x, the membership class of the individual becomes the majority class.

4. Experimental results

The proposed system is simulated on Microsoft Windows 7 Operating System with the configuration: Intel(R) Processor 2.60 GHz with 4G RAM, by using MATLAB R2012a software [14].

In our experiments, the SVM parameters such as the kernel functions, the kernel parameter (γ) and the punishment factor (C) are chosen as RBF kernel functions, $C = 4$ and $\gamma = 4$, which were optimized manually to obtain higher levels of accuracy.

After applying GPLVM algorithm on training and testing dataset, we extracted 3-dimentional dataset from each one.

The performance of our proposed IDS is measured by the Detection Rate (DR) and False Alarm Rate (FAR). They can be calculated using the confusion matrix in Table 2 and defined as follows:

$$DR = \frac{TP}{TP+FN} \times 100 \quad (10)$$

$$FAR = \frac{FP}{TN+FP} \times 100 \quad (11)$$

With:

TP= attack connection record classified as attack (TP).

FP= attack connection record classified as normal (FP).

TN= normal connection record classified as normal (TN).

FN= normal connection record classified as attack (FN).

Table 2. Confusion Matrix for Evaluation of Intrusions (Attacks) in KDD 99 Dataset

Confusion Matrix (Standard Metrics)		Predicted Connection Label	
		Normal	Intrusions (Attacks)
Actual Connection label	Normal	True Negative(TN)	False Alarm(FP)
	Intrusions(Attacks)	False Negative (FN)	Correctly detected Attacks(TP)

In order to reduce the size of the dataset, we select some records as [4], including training dataset and testing dataset. Table 3 shows detailed information about the number of all records.

Table 3. Number and distribution of training and testing dataset

Different types	Training dataset	Testing dataset
Normal	100	300
DoS	99	294
Probe	78	218
R2L	53	157
U2R	42	50

We tested the proposed algorithm with an IDS based on GPLVM and Artificial Neural Network (ANN) [15]. Table 4 presents the results of the detection rate for GPLVM+ANN and our system and Table 5 gives the false alarm rate for the both algorithms. Simulation results show that the proposed method outperforms the other method. It provides averagely high performance of detection rate and also minimizes the false alarm rate.

Table 4. Detection rate of the proposed method and GPLVM+ANN

Detection Rate	Normal	Dos	Probe	R2L	U2R
GPLVM+ANN	99.76%	99.24%	100%	36.77%	67.84%
Our approach	100 %	98.54%	100%	62.76%	88.91%

Table 5. False positive rate of the proposed method and GPLVM+ANN

False Alarm Rate	Normal	Dos	Probe	R2L	U2R
GPLVM+ANN	0.24 %	0.76%	0 %	63.23%	32.16%
Our approach	0 %	1.46%	0%	37.24%	12.09%

Table 6 shows the comparison of training and testing speed of GPLVM+SVMs and GPLVM+ANN. It is clear that the proposed GPLVM+SVMs classifier is 4 times faster in training and 2 times faster in testing than the IDS based on GPLVM+ANN.

Table 6. Speed comparison of the proposed method and GPLVM+ANN

Classifiers	Training time (s)	Testing time (s)
GPLVM+ANN	118.6	27.4
Our approach	30.5	12.8

5. Conclusion

In this paper, we have proposed an intrusion detection system based on SVMs combined with GPLVM. The main contribution of the present work is to construct an intrusion detection model with high accuracy but also make the IDS model to be processed as fast as possible. The simulation experiments on KDD-Cup99 intrusion detection dataset show that the proposed method has the higher DR and lower FAR than the GPLVM+ANN, especially for Dos, Probe and U2R attacks, and it is much faster in processing speed than the other approach. For future work, we have planned to implement our proposed method in the real life environment.

References

[1] H. Moudni, M. Er-rouidi, H. Mouncif and B. El Hadadi, "Performance analysis of AODV routing protocol in MANET under the influence of routing attacks," *2016 International Conference on Electrical and Information Technologies (ICEIT)*, Tangiers, 2016, pp. 536-542.

[2] Somwang, P., & Lilakiatsakun, W. (2015). Anomaly traffic detection based on PCA and SFAM. *Int. Arab J. Inf. Technol.*, *12*(3), 253-260.

[3] Sen, N., Sen, R., & Chattopadhyay, M. (2014, November). An effective back propagation neural network architecture for the development of an efficient anomaly based intrusion detection system. In *Computational Intelligence and Communication Networks (CICN), 2014 International Conference on* (pp. 1052-1056). IEEE.

[4] Zhang, Z., & Gu, B. (2016). Intrusion Detection Network Based on Fuzzy C-Means and Particle Swarm Optimization. In *Proceedings of the 6th International Asia Conference on Industrial Engineering and Management Innovation* (pp. 111-119). Atlantis Press.

[5] Lawrence, N. D. (2004). Gaussian process latent variable models for visualisation of high dimensional data. *Advances in neural information processing systems*, *16*(3), 329-336.

[6] Hearst, M. A., Dumais, S. T., Osman, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their Applications*,*13*(4), 18-28.

[7] KDD Cup 1999 dataset. http://kdd.ics.uci.du/databases/kddcup99/kdd-cup99.html

[8] Quirion, S., Duchesne, C., Laurendeau, D., & Marchand, M. (2008). Comparing GPLVM approaches for dimensionality reduction in character animation.

[9] Jolliffe, I. (2002). *Principal component analysis*. John Wiley & Sons, Ltd.

[10] Schölkopf, B., Smola, A., & Müller, K. R. (1997, October). Kernel principal component analysis. In *International Conference on Artificial Neural Networks*(pp. 583-588). Springer Berlin Heidelberg.

[11] Tipping, M. E., & Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *61*(3), 611-622.

[12] Lawrence, N. (2005). Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *Journal of Machine Learning Research*, *6*(Nov), 1783-1816.

[13] Guo, J., Takahashi, N., & Hu, W. (2008, December). An efficient algorithm for multi-class support vector machines. In *2008 International Conference on Advanced Computer Theory and Engineering* (pp. 327-331). IEEE.

[14] Grant, M., Boyd, S., & Ye, Y. (2008). CVX: Matlab software for disciplined convex programming.

[15] Zhang, Y., Ding, X., Liu, Y., & Griffin, P. J. (1996). An artificial neural network approach to transformer fault diagnosis. *IEEE Transactions on Power Delivery*,*11*(4), 1836-1841.