International Journal of Security and Its Applications
Vol.9, No.9 (2015), pp.371-386
http://dx.doi.org/10.14257/ijsia.2015.9.9.32

Cyber Attack Detection System based on Improved Support Vector Machine

Shailendra Singh and Sanjay Silakari

Member, IEEE
Department of Information Technology
Rajiv Gandhi Technological University
Bhopal, India
Department of Computer Science and Engineering
Rajiv Gandhi Technological University
Bhopal, India
shailendrasingh@rgtu.net, ssilakari@rgtu.net

Abstract

This paper presents a novel cyber attack classification approach using improved Support Vector Machine (iSVM) by modifying Gaussian kernel. The Support Vector Machine (SVM) is based on machine learning technique known to perform well at various pattern recognition tasks; such as image classification, text categorization and handwritten character recognition. The cyber attack detection is basically a pattern classification problem, in which classification of normal pattern is done from the abnormal pattern (attack). Although, traditional SVM is better classifier in terms of fast training, scalable and generalization capability. Performance of traditional SVM is enhanced in this work by modifying Gaussian kernel to enlarge the spatial resolution around the margin by a conformal mapping, so that the separability between attack classes is increased. It is based on the Riemannian geometrical structure induced by the kernel function. In the proposed method, class specific Cyber Attack Detection System which combines feature reduction technique and improved support vector machine classifier. This technique has two phases, in the first phase we reduced the redundant features of the original KDDCUP2009 dataset by Generalized Discriminant Analysis (GDA). In the second phase we used improved Support Vector Machine (iSVM) classifier to classify the reduced dataset obtained from first phase. Result shows that iSVM gives 100% detection accuracy for Normal and Denial of Service (DOS) classes and comparable to false alarm rate, training, and testing times.

Keywords: *Improved Support Vectors Machine, Gaussian kernel, Pattern Recognition, Generalized Discriminant Analysis, Machine learning, Riemannian geometrical structure.*

1. Introduction

The rapid increase in connectivity and accessibility of computer system has resulted frequent chances for cyber-attacks. Attack on the computer infrastructures are becoming an increasingly serious problem. Basically the cyber-attack detection is a classification problem, in which we classify the normal pattern from the abnormal pattern (attack) of the system. Support Vector Machine (SVM) [1] is a well-known machine learning algorithm used to solve the classification problem. SVM is based on recent advances in statistical learning theory and has been successfully applied in real world problems such as text categorization [2] image classification [3], handwritten character recognition [4]. The choice of the kernel greatly affects the SVM's ability to classify data points accurately. The Riemannian geometry induced by kernel function [5] proposes a method of modifying a Gaussian kernel to improve the performance of the SVM. The idea is to enlarge the

ISSN: 1738-9976 IJSIA

spatial resolution around the margin by conformal mapping; so that the separability between classes is increased (large margin), which means less misclassification of data hence improved classification accuracy.

In this paper we proposed a class specific cyber attack detection technique which is based on Generalized Discriminant Analysis (GDA) [6] feature reduction technique and improved Support Vector Machine (iSVM) classifier. Therefore, the features of original KDDCUP2009 dataset are reduced by GDA technique, which is implemented in this work. These reduced features dataset are applied to the iSVM classifier and traditional SVM. The relative results of the both the classifiers are also obtained to ascertain the theoretical aspects. The analysis is also taken up to show that iSVM performs better than SVM. The classification accuracy of iSVM remarkably improve (accuracy for Normal class as well as DOS class is almost 100%) and comparable to false alarm rate and training, testing times.

2. Related Work

SVM is a powerful tool for classification problems. But still has some drawbacks. The first problem is that SVM is sensitive to outliers or noises [7]. The second, SVM designed for the two class problems, it has to be extended for multiclass problem by choosing suitable kernel function. The performance of the SVM depends upon the kernel function. Some methods to improve the performance of SVM were proposed. Fuzzy SVM [8-9] is one of the improvement made on the traditional SVM. Several machine learning paradigms including Artificial Neural Network [10], Linear Genetic Programming (LGP) [11], Data Mining [12-13], etc. have been investigated for the classification of cyber-attack. However, the machine learning techniques as mentioned above is not suitable for the huge data set and its training, testing time and classification accuracy get affected with size of the dataset. As the size of the dataset grows the training and testing time of the above mentioned classifiers increases and accuracy decreases. Also the machine learning techniques are sensitive to the noise in the training samples. The presence of mislabeled data if any can result in highly nonlinear decision surface and over fitting of the training set. This leads to poor generalization ability and classification accuracy.

At same time, there is also some progress made in the feature extraction field. Andrew H. Sung [14] ranks the importance of the 41 features for the five categories in KDDCUP99 datasets by deleting one feature at a time when adopting the SVM and neural network as the classification method Melanie J. Middlemiss [15] uses the genetic algorithm to do the feature extraction for intrusion detection. The researchers have been working the combination of feature extraction technique and classification algorithms for cyber attack detection system. Taeshik Shon [16] proposes a machine learning frame work for cyber attack detection, using SVM and GA. It uses GA to extract the attacks features and SVM for classification. Because the datasets to be processed in the cyber attack detection are always very large. The main problem is to raise the detection rate and real time detection ability by combining the feature extraction technique with the classification technique and thus the performance of the classifier gets improved.

G. Baudat and F. Anouar, [6] proposed Generalized Discriminant Analysis (GDA) to deal with nonlinear discriminant analysis for feature reduction of dataset. In the transformed space, liner properties make it easy to extend and generalize the classical Linear Discriminant Analysis (LDA) [17] to nonlinear discriminant analysis. Recently, some work based on this method has been reported by researchers [18-19] in medical science for feature reduction of medical dataset. Which has been proven promising technique for feature reduction.

3. KDDCUP2009 Data Set

In the 1998 DARPA intrusion detection evaluation 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. This dataset contain 494021 connection records which are huge data and it contains the redundant connection also. To train such a huge dataset it take long time and even takes a day for low configuration machine. We were fortunate, enough to get refined version of kddcup1999 dataset, released on 2009 i.e. KDDCUP2009 dataset [20]. They have removed all the redundant records from the KDDCUP99 dataset. The kddcup2009 dataset contain only 1, 25, 973 connection records. 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 22 attack types that could be classified into four main categories:

- DOS: Denial of Service attacks.
- R2L: Remote to Local attacks.
- U2R: User to Root attacks.
- Probe: Surveillance.

A. Denial Of service Attack (DOS)

A denial of service attack is a class of attacks 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 e.g. neptune, teardrop, smurf, pod, back, land.

B. Remote to Local (R2L) Attacks

A remote to local attack is 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 e.g. guss_passwd, ftp_write, multihop, imap, phf, spy, warezmaster, warezclient.

C. User to Root (U2R) Attacks

User to root (U2R) attacks are 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 e.g. loadmodule, perl, buffer_overflow, rootkit.

D. Probe

Probe is a 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 e.g. ipsweep, portsweep, nmap, satan.

4. Support Vector Machine

Support Vector Machines (SVMs) [21] were first introduced in the mid of 1990s, and have since been established as one of standard tools for machine learning and data mining. SVM were originally designed for binary classification. However, cyber attack detection is a problem of multi-class classification. How to effectively extend SVM for multi-class classification is still an ongoing research issue. Currently there are two types of approaches for multi-class SVM. One is by combining several binary classifiers while the other is by directly considering all training samples into one optimization formulation.

The SVM identifies the best separating hyper plane (the plan with maximum margins) between the two classes of the training samples within the feature space by focusing on training cases placed at the edge of the class descriptors. In this way, not only an optimal hyper plane is fitted, but also less training samples are effectively used; thus high classification accuracy is achieved with small training sets [22]. We construct k SVM model where k is the number of classes. The i^{th}SVM is trained with all of the examples in the i^{th}class with positive labels, and all other examples with negative labels. Thus, given

$l$training data $(x_1, y_1), (x_2, y_2), \ldots\ldots, (x_l, y_l)$, $i = 1,2, \ldots\ldots, l$, where $x_i \epsilon R^l$ and $y_i \epsilon \{1,2,..k\}$ is the class of x_j the j^{th} SVM solves the following optimization problem [23]:

$$\min_{w^j b^j \xi_i^j} \left\{ \frac{1}{2} (w^j)^T w^j + C\left(\sum_{i=1}^{l} \xi_i^j\right) \right\}, \; j = 1,2, \ldots k \tag{1}$$

$$\begin{aligned} s.t. \quad (w^j)^T \emptyset(x_i) + b^j &\geq 1 - \xi_i^j \qquad if\ y_i = j \\ (w^j)^T \emptyset(x_i) + b^j &\leq -1 + \xi_i^j \qquad if\ y_i \neq j, \\ \xi_i^j &\geq 0, \quad i = 1, \ldots l, \end{aligned}$$

Where $\emptyset(x_i)$ is nonlinear function that maps x into a higher dimensional space [24] $w, b, and\ \xi$ are the weight vector, bias and slack variable, respectively. C is constant and determined a priori. Searching for the optimal hyperplane in equation (1) is quadratic programming problem. Minimizing $\frac{1}{2}(w^j)^T w^j$ means that we would like to maximize $2/\|w^j\|$, the margin between two classes of attack data. Where data are not linearly separable, there is penalty terms $C\left(\sum_{i=1}^{l} \xi_i^j\right)$ which can reduce the number of training errors the basic concepts behind SVM is to search for a balance between the regularization term $\frac{1}{2}(w^j)^T w^j$ and training errors. After solving the equation (1) to get k decision functions

$$\sum_{i=1}^{l} \alpha_i^j \; K(x, x_i) + b^1$$

$$\sum_{i=1}^{l} \alpha_i^j \; K(x, x_i) + b^k$$

We say x_i is in the class which has the largest value of the decision function:

$$class\ of\ x \equiv \underset{j=1,\ldots k}{\operatorname{argmax}} \sum_{i=1}^{l} \alpha_i^j \; \mathbf{K}(\mathbf{x}, \boldsymbol{x_i}) + b^j) \tag{2}$$

Here, kernel $\mathbf{K}(\mathbf{x}, \boldsymbol{x_i})$, is a Gaussian kernel function and α_j^i Lagrange multiplier. To improve the classification accuracy of the SVM classifier we will modify Gaussian kernel function $K(x, x_i)$ in data dependent way.

5. Proposed Cyber Attack Detection System

We present a novel concept to build a class specific Cyber Attack Detection System by integrating two different techniques: Feature extraction technique and Classification technique. The Figure 1: shows the architecture for the proposed cyber attack detection system.

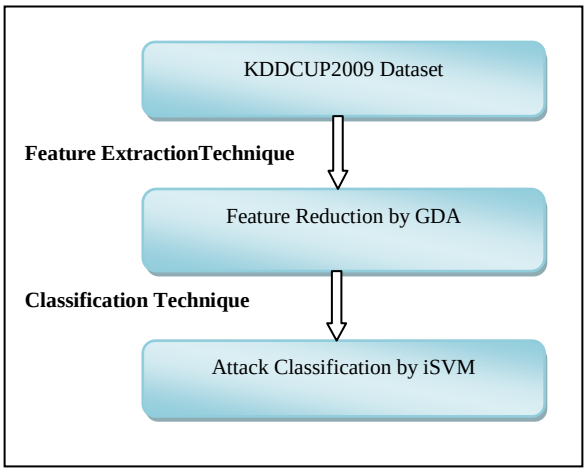

Figure 1. Architecture of Proposed Cyber Attack Detection System

E. Feature Extraction Technique

Feature extraction [25] techniques are commonly used as preprocessing to machine learning and statistical 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 [26]. 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 [27]. Thereby improving the overall performance of classifier based cyber attack detection. The Generalized Discriminant Analysis GDA [6, 17-18] is a method designed for nonlinear classification based on a kernel function $\emptyset$ which transform the original space X to a new high-dimensional feature space $Z: \emptyset: X \rightarrow Z$.

The between-class scatter matrix and within-class scatter matrix of the nonlinearly mapped data is

$$B^{\emptyset} = \sum_{c=1}^{C} M_c \, m_c^{\emptyset} (m_c^{\emptyset})^T \tag{3}$$

$$W^{\emptyset} = \sum_{c=1}^{C} \sum_{x=X_c} \emptyset(x) \, \emptyset(x)^T \tag{4}$$

Where:

$m_c^{\emptyset}$ 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 projection matrix $U^{\emptyset}$ that maximizes the ratio

$$U_{opt}^{\emptyset} = argmax \frac{|(U^{\emptyset})^T B^{\emptyset} U^{\emptyset}|}{|(U^{\emptyset})^T W^{\emptyset} U^{\emptyset}|} = [u_1^{\emptyset}, \dots \dots, u_N^{\emptyset}] \tag{5}$$

We have to find eigenvalues λ_i and eigenvectors $u^{\emptyset}$ solution of the equation:

$$B^{\emptyset} u_i^{\emptyset} = \lambda_i W^{\emptyset} u_i^{\emptyset} \tag{6}$$

The largest eigenvalue of (6) gives maximum of the following quotient of the inertia:

$$\lambda_i = \frac{B^{\emptyset} u_i^{\emptyset}}{W^{\emptyset} u_i^{\emptyset}} \tag{7}$$

As the eignvectors are linear combinations of Z elements, there exist coefficients α_{ci} $(c = 1 \dots, C; i = 1 \dots .. M_c$;) such that

$$u^{\emptyset} = \sum_{c=1}^{C} \sum_{i=1}^{M} \alpha_{ci} \emptyset(x_{ci}) \tag{8}$$

Where x_{ci} is the i^{th} sample of the class c. The solution is obtained by solving

$$\lambda_i = \frac{\propto^T KDK \propto}{\propto^T KK \propto} \tag{9}$$

Where: coefficient vector $\propto = (\propto_c)$, $c = 1 \dots .. C$ is a vector of weights with $\propto_c = (\propto_{ci}), \; i = 1 \dots . M_c$.

Where $\propto_c$ is coefficient of the vector $u^{\emptyset}$ in the class c. and α^T is the transpose of coefficient vector.

The kernel matrix $K(M \times M)$ is composed of the dot products of nonlinearly mapped data, i.e.

$$K = (K_{kl})_{k=1 \dots C, \; l=1 \dots C} \tag{10}$$

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

$$D = (D_c)_{c=1..C} \quad (11)$$

Where

The c^{th} on the diagonal has all elements equal to $1/M_c$. Solving the eigenvalue problem yields the coefficient vector $\propto$ which define the projection vectors $u^{\emptyset} \in Z$. A projection of a testing vector x_{test} is computed as

$$(u^{\emptyset})^T \emptyset(x_{test}) = \sum_{c=1}^{C} \sum_{i=1}^{M} \propto_{ci} k(x_{ci}, x_{test}) \quad (12)$$

The algorithm of proposed Generalized Discriminant Analysis (GDA) technique is given below:

Step1: Compute the matrices K and D by solving the equation (10) and (11),
Step2: Decompose K using eigenvectors decomposition,
Step3: Compute eigenvectors $u^{\emptyset}$ and eigenvalues λ_i of the equation (6),
Step4: Compute eigenvectors $u^{\emptyset}$ using α_{ci} from equation (8) and normalize them,
Step5: Compute projections of test points onto the eigenvectors $u^{\emptyset}$ from equation (12).

The Linear Discriminant Analysis (LDA) [17] 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. The number of classes of KDDCUP2009 dataset is five. Therefore, the optimal number of eigenvectors for the data transformation is equal to four. After feature reduction of KDDCUP2009 dataset the reduced features are fed to the both SVM and iSVM classifiers and the performance is measured.

F. Improved SVM (iSVM) algorithm by modifying kernels

Kernels provide support vector machines with the capability of implicitly mapping non-linearly separable data points into a different dimension, where they are linearly separable. Mapping the data points to a higher dimensions, involve cost. More dimensional means larger vectors which means larger memory requirements and longer calculation times. Fortunately, SVMs do not need to store these high dimensional vectors explicitly. They map the input data into the higher dimension and then are only to store inner products [23]. Different kernel functions provide different mappings. Unfortunately, there is no silver bullet choice of kernel. Each kernel has its advantages and disadvantages. The choice of the kernel greatly affects the SVM's ability to classify data points accurately. We modify existing Gaussian kernel according to our need. This modified kernel gives better performance compare with the original Gaussian kernel.

A nonlinear SVM maps each samples of input space R into a feature space F through a nonlinear mapping $\emptyset$. The mapping $\emptyset$ defines an embedding of S into F as a curve submanifold.

Denote $\emptyset(x)$ the mapped samples of S in the featured space; small vector dx is mapped to:

$$\emptyset(dx) = \nabla\emptyset . dx = \sum_i \frac{\partial}{\partial x^{(i)}} \emptyset(x) dx^{(i)} \quad (13)$$

Where $\nabla\emptyset = \frac{\partial}{\partial x^{(i)}} \emptyset(x)$.

The squared length of $\emptyset(dx)$ is written as:

$$ds^2 = |\emptyset(dx)|^2 = \sum_{i,j} g_{ij}(x) dx^{(i)} dx^{(j)} \quad (14)$$

Where:

$$g_{ij}(x) = \left(\frac{\partial}{\partial x^{(i)}}\emptyset(x)\right).\left(\frac{\partial}{\partial x^{(j)}}\emptyset(x)\right), \quad (15)$$

The dot denoting the summation over index α of $\emptyset$. The $n \times n$
Positive-definite matrix $G(x) = (g_{ij}(x))$ is the Riemannian metric tensor induced in S.

$$g_{ij}(x) = \frac{\partial}{\partial x^{(i)}}\frac{\partial}{\partial x^{(j)}}K(x, x_i) \quad (16)$$

We can increase the margin or the distances (ds) between classes to improve the performance of the SVM. Taking eq.(14) in to account, this leads us to increase the Riemannian metric tensor around the boundary and to reduce it around other samples. In view of eq.(16), we can modify the kernel K such that $g_{ij}(x)$ is in large around the boundary.

Modifying kernel based on the structure of the Riemannian geometry: Assume the kernel can be modified as:

$$\tilde{K}(x, x_i) = p(x)p(x_i)K(x, x_i) \quad (17)$$

is called a conformal transformation of a kernel by factor p(x). We take the kernel function used in SVM is Gaussian Kernel, i.e.:

$$K(x, x_i) = \exp(-\|x - x_i\|^2/\sigma^2) \quad (18)$$

Here, the parameter σ is kernel width. It is proved that the corresponding Riemannian metric tensor is changed into:

$$g_{ij}(x) = \frac{1}{\sigma^2}\delta_{ij} \quad (19)$$

After modifying the kernel Riemannian metric tensor is changed into:

$$\tilde{g}_{ij}(x) = p_i(x)p_j(x) + p^2(x)g_{ij}(x) \quad (20)$$

To ensure that p(x) has large value around the support vector (SV), by the conformal transformation of the Gaussian kernel,

$$p_i(x) = \partial p(x)/\partial x_i \quad (21)$$

For maximum p(x) the value of $p_i(x) = 0$.

In order to ensure that p(x) has large values at the support vector positions, it can be constructed in a data dependent way as:

$$p(x) = \sum_{i \in SV}\alpha_i \exp(-\|x - x_i\|^2/2\tau^2 \quad (22)$$

Where τ is a free parameter and summation runs over all the support vectors. As we see $p_i(x)$ and $p(x)$ are large when x is close to support vectors and those are small when x is for away from SVs then, when x is close to support vectors the $g_{ij}(x)$ around support vectors is increased. So the spatial resolution around the boundary is enlarged and classification ability of SVM becomes stronger.

We summarized the procedure of the proposed Algorithm as follows:

Step1: Train SVM with primary Gaussian kernel $K(x, x_i)$ to extract the information of SVs, then modify Gaussian kernel K according to the formula (17) and (22).
Step2: Train the SVM with the modified Gaussian kernel $\tilde{K}$.
Step3: Iteratively apply the above two steps until the best performance is achieved.

6. Experiments and Results

All the experiments were performed on an Intel Xeon with a 2.4GHz CPU and 2 GB of RAM. We used $SVM^{Multiclass}$ version V2.12 software [28]. To evaluate the performance of our proposed cyber attack detection system, we used the KDDCUP2009 dataset.

G. Preprocessing of data

The KDDCUP2009 dataset is not a normalized dataset. Therefore, it needs preprocessing of dataset before given to feature reduction algorithm. Logarithmic scaling (with base 10) was applied to the very large features. The KDDCUP2009 dataset consist of 1, 25, 973 records for training and 25,192 records for testing. We have written JAVA code to read data from the KDDCUP2009 dataset line by line where each line contains records. As we have seen from the dataset it contains both symbol and numeric form we have to convert symbolic form to numeric form and stored these data into two dimensional tables. We created five files for five classes to store specific class data into these files like Normal.data, DOS.data, Probe.data, U2R.data and R2L.data for processing of the data.

H. Experimental Setting

Our experiment is split into three main steps. In the first steps, we prepare different dataset for training and testing. Second, we apply feature reduction algorithm (GDA) to the dataset. The original KDDCUP2009 dataset to select most discriminant features for cyber attack detection. Third, we classify the cyber attacks by using traditional SVM and improved SVM (iSVM) as two different classifiers.

In the first step of the experiment, we prepare the data set for the training and testing. We used a training set of 125973 records and testing set of 25,192 records. We choose 68253, 45927, 85, 52 and 11656 samples for Normal, DOS, R2L, U2R and Probe respectively for training. The testing set consists of 25,192 kinds of data and then used test data 13637, 9234, 21, 11, and 2289 for Normal, DOS, R2L, U2R and Probe respectively.

The second step, we apply feature reduction technique on the original KDDCUP2009 dataset with 41 features. We use Generalized Discriminant Analysis (GDA) algorithm for selecting most discriminant features. Each record is located in the n-dimensional space, with each dimension corresponding to a feature of the record.

Finally, the evaluation is done using SVM and iSVM classifiers. During the SVM training process the default regularization parameter is set to C=1000 with optimization done for 88 iterations. In the experiment, the Gaussian kernel is modified based on equation (15). The features are represented by d dimensional feature vectors. Selecting an appropriate d is very critical for cyber attack detection. the value of d is set for 28 in the experiment based on 3-fold cross validation. We set the kernel width σ to be equal to the optimal one $\sigma = 1.0$ it has been experimentally proved that the value of τ is around $\sigma/\sqrt{d}$. Therefore we set the value of τ is equal to 0.18. The regularization parameter is set to C=1000 for iSVM.

7. Exprimental Results

We apply Generalized Discriminant Analysis (GDA) algorithm for feature reduction of the original KDDCUP2009 dataset with 41 features. With this algorithm the optimal number of eigenvectors for the data transformation is equal to four.

Table 1. Comparison of Training Time (in Seconds)

Predicted Actual	Normal	Probe	DOS	R2L	U2R
SVM	10.10	6.20	25.02	6.50	8.01
iSVM	5.01	3.11	14.56	3.02	4.17

Therefore, the new feature set not only reduces the number of the input features but also increases the classification accuracy by selecting most discriminating features for the better discrimination of the different cyber attack classes.

Table 2. Comparison of Testing Time (in Seconds)

Predicted Actual	Normal	Probe	DOS	R2L	U2R
SVM	1.12s	1.01	4.02	1.01	2.01
iSVM	0.26	0.21	2.13	0.21	1.07

Table 3. Comparison of False Alarm Rate

Classifiers	Normal	Probe	DOS	R2L	U2R
SVM	0.18	0.75	0.52	40.91	33.33
iSVM	0.06	0.14	0.06	06.66	20

The comparison of SVM and iSVM classifiers is done with respect to different performance indicators: Detection Rate, training time, testing time and false alarm rate. The results of the training time, testing time and false alarm rate for both the classifiers are presented in Table I, II and Table III.

The detection of attack and normal pattern 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.

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 behavior

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

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 IV shows the confusion matrix.

Table 4. Confusion Matrix

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

Table 5. Confusion Matrix for svm Classifier before Modifying the Kernel

Predicted Actual	Normal	Probe	DOS	R2L	U2R	%Correct
Normal	13589	12	30	4	2	99.64
Probe	16	2256	13	3	1	98.55
DOS	4	2	9227	1	0	98.92
R2L	2	2	4	13	0	62.0
U2R	2	1	1	1	6	55.01
%Correct	99.82	99.25	99.48	59.09	66.67	

The performance of the traditional SVM is shown in Table V in the form of confusion matrix.

After modifying the Gaussian Kernel the performance of the iSVM is remarkably improved, which is shown in Table VI. We compare the performance of other classifiers with iSVM as shown in Table VII. iSVM is superior to all the mentioned classifiers.

Table 6. Confusion Matrix for Improved svm Classifier after Modifiying the Kernel

Predicted Actual	Normal	Probe	DOS	R2L	U2R	%Correct
Normal	13637	0	0	0	0	100
Probe	5	2280	2	1	1	99.90
DOS	0	0	9234	0	0	100
R2L	2	2	2	14	1	66.66
U2R	1	1	1	0	8	72.72
%Correct	99.94	99.86	99.94	93.34	80.0	

Table 7. Performance Comparison of Improved svm with Other Classifiers

Classifiers Attack classes	BN[13] DR %	CART[13] DR %	SVM DR %	iSVM DR %
Normal	99.57	95.50	99.64	100
Probe	96.71	96.85	98.55	99.98
DOS	99.02	94.31	98.92	100
R2L	97.87	97.69	63.20	85.54
U2R	56.00	84.00	59.60	78.61

I. Discussion

By using Generalized Discriminant Analysis (GDA) as feature reduction technique and iSVM as cyber attack classification approach, the system performance (detection rate, training time, testing time and false alarm rate) and the scalability is improved. We applied reduced dataset for training and testing to both the classifiers SVM and iSVM.

Support Vector Machines (SVM) is learning machines that plot the training vectors in high-dimensional feature space. Label each vector by its class. SVMs view the classification problem as a quadratic optimization problem. The SVM classify data by determining a set of support vectors, which are the members of the set of training inputs that outline the hyper plane in feature space [21]. The SVM are based on the idea of structural risk minimization, which minimizes the generalization error on unseen data. The number of free parameters used in the SVMs depends on the margin that separate the data points. The SVM provide a generic mechanism to fit the surface of the hyper plane to the data through the use of a kernel function. We used Gaussian Kernel function to the SVM during training process, which selects the support vectors along the surface of this function. This capability allows classifying a broader range of problems.

The SVM and Improved Support Vector Machine (iSVM) separate the data into two classes, classification into additional classes by applying one against all (OAA) method. In the OAA method, a set of binary classifiers (k parallel SVMs, where k denotes the number of classes) is trained to be able to separate each class from all others. Then each data object is classified to the class for which the largest decision value has been determined. Then voting strategy aggregates the decisions and predicts that each data object is in the class with the largest vote [29].

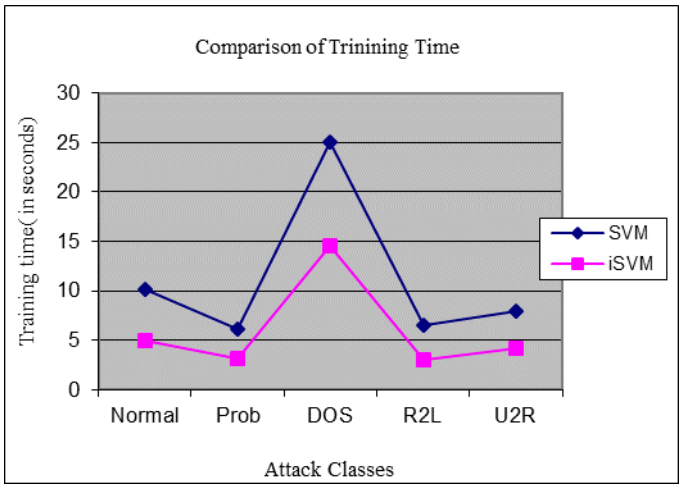

Figure 2. Comparision of Training Time of SVM and iSVM Classifiers

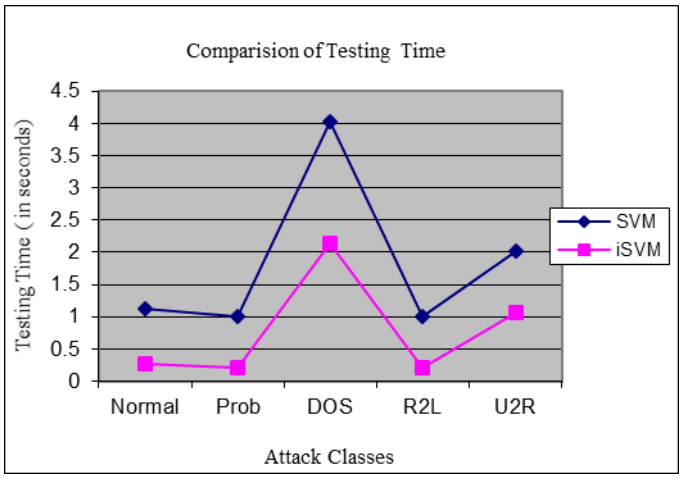

Figure 3. Comparision of Training Time of SVM and iSVM Classifiers

We train the SVM with the Gaussian Kernel and observe the training time of the classifier. Then we test with the test dataset to the SVM classifier, follow the same procedure until the best performance is achieved. Then record the training, testing and detection rate for this classifier.

Then we train the same classifier with modify Gaussian kernel K according to the formula (15) and (19).We train with the modified Gaussian Kernel $\tilde{K}$ and apply test dataset to it and the performance of the classifier is observe repeat the above process until the best performance of the classifier is obtained. As shown in Figure 2 the training time of iSVM classifier reduced, in case of DOS class of attack it reduces form 25 seconds to only 15 seconds. In case of iSVM the overall time reduces for all the classes of attacks. Testing of iSVM is also comparable with SVM as shown in Figure 3.

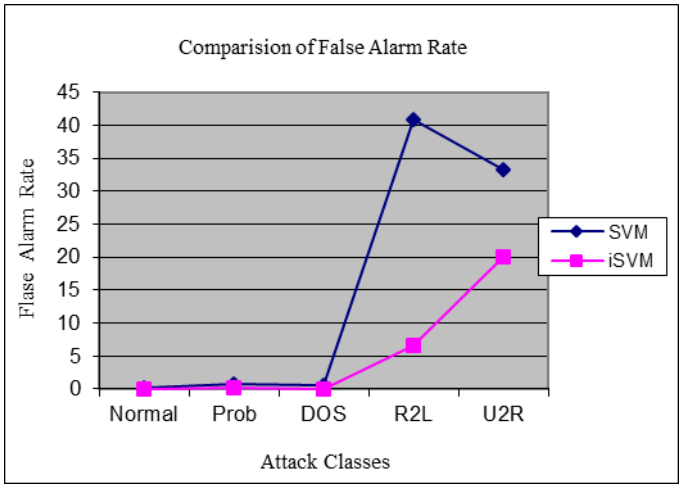

Figure 4. Comparison of False Alarm Rate for SVM and iSVM

In the Figure 4 it is clearly shown that the false alarm rate for the Normal, Probe, DOS classes of attack is low for both SVM and iSVM, but the false alarm rate for the R2L and U2R classes of attack are high SVM and comparatively low for the iSVM.

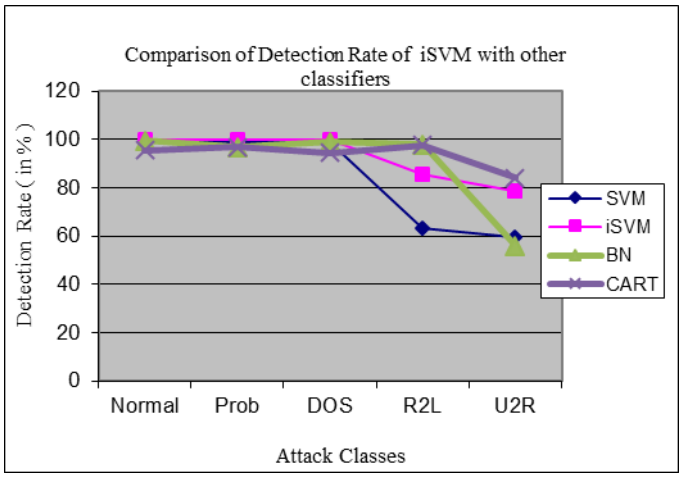

Figure 5. Comparision of Detection Rate of iSVM with Other Classifiers

The cyber attack detection rate of iSVM is higher than the traditional SVM as shown in Figure 5 We also compare the detection rate of iSVM with Bayesian Network (BN) shows high detection accuracy for R2L class of attack i.e. 97.87% which is higher as compare to all other classifiers and Classification and Regression Tree (CART) exhibits highest detection accuracy for U2R class of attack i.e. 84%. Although, the detection accuracy of iSVM has improved as compare to detection accuracy of SVM but still there is need to improve the detection accuracy for the classes like U2R and R2L. iSVM shows execellent detection rate for the classes like Normal and DOS i.e. 100% and 99.98% for the Probe.

8. Conclusion and Future Works

In this paper we presented class specific cyber attack detection system. Two approaches have been used in this work. The first approach, feature reduction technique; we used Generalized Discriminant Analysis (GDA) in which the GDA is able to significantly decrease training and testing times while retaining high detection rates. The second approach is using improved support vector machine classifier. We have modified the Gaussian kernel in data dependable way to improve the classification accuracy of the traditional SVM. As result shows the accuracy of iSVM is remarkably improved for the Normal and DOS classes of attacks. The proposed model has yield the classification accuracy of 100% for DOS, and Normal classes. A number of experiments were conducted to evaluate the proposed class specific cyber attack detection system. The experimental results demonstrate that the proposed class specific cyber attack detection system can reduced training time and, testing time where false alarm rate with high cyber attack detection accuracy is improved. Therefore, combining the two approaches, feature reduction and classification approach give better performance.

Future work will involve building cyber attack detection system that integrates the different class specific cyber attack detection system, which will be able to give 100% detection rate for all the classes, and investigate the possibility and feasibility of implementing this approach in real time cyber attack detection system.

References

[1] V. Vapnik, “Statistical Learning Theory”, (**1995**).
[2] T. Jachims, “Text categorization with Support Vector Machines: learning with many relevent features”, Proc. European Conference on Machine learning, (1398), (**1998**), pp. 137-142.
[3] O. Chapelle, P. Haffner and V. Vapnik, “Support Vector Machines for histogram-based image classification. Neural Networks”, IEEE Transactions on, vol. 10, no. 5, (**1999**), pp. 1055-1064.

[4] C. Cortes and V. Vapnik, "Support Vector Netwroks", Machine Learning, vol. 20, no. 3, (**1995**), pp. 273-297.
[5] S. Amari, "Improving support vector machine classifiers by modifying kernel functions", Neural Networks, vol. 12, (1999), pp. 783-792.
[6] G. Baudat and F. Anouar, "Generalized Discriminant Analysis Using a Kernel Approach", Neural Computation, vol. 12, (**2000**), pp. 2385-2404.
[7] Y.-h. Liu and Y.-t. Chen, "Face recognition using total margin based adaptive fuzzy support vector machines", IEEE Transactions on Neural Networks, vol. 18, vol. 1, (**2007**), pp. 178-192.
[8] S.-W. Xiong, H.-b. Liua nd X.-x. Niu, "Fuzzy support vector machines based on FCM clustering", Proceddings of the fourth international conference on Machine Learning and Cybernetics, Guangzhou, China, August 18-21, IEEE, (**2005**), pp. 2608-2613.
[9] A. K. Ghosh and A. Schwartzbard, "A study in Using Neural Networks for Anomaly and Misuse detection", Proceeding of the 8th USENIX Security Symposium, pp. 23-36. Washington, D.C. US. (**1999)**.
[10] S. Mukkamala, AH. Sung and A. Abraham, "Modeling Inrusion Detection Systems Using linear genetic programming approach", The 17th international conference on industrial & engineering applications of artificial intelligence and expert systems, innovation in applied artificial intelligence.
[11] S. Mukkamala, A. H. Sung, A. A. Ramos V., "Inrusion detection systems Using adaptive regression splines", In: Seruca I, Filipe J, Hammoudi S, Cordeiro J, editors, Proceedings of 6th international conference on enterprise information systems, ICEIS'04, vol. 3, Portugal, (**2004),** pp. 26-33.
[12] W. Lee, S. J. Stolfo and K. Mok, " Data mining in work flow environments: Experiences in intrusion detection, Proceedings of the Conference on Knowledge Discovery and Data Mining (KDD-99), (**1999**).
[13] C. S. Sung A. and A. Abraham, "Feature Deduction and ensemble Design of Intrusion Detection Systems, Computers and Security", vol. 24, no. 4, Elsevier, (**2005**) June, pp. 295-307.
[14] A H. Sung and S. Mukkamala, "Idenfify important features for intrusion detection using support vector machines and neural networks", Proceedings of 2003 Symposium on Applications and Internet, Piscataway, NJ, USA: IEEE Computer Scociety, (**2003**) January 27-31, pp. 209-217.
[15] M. J. Middlemiss and G. Dic, "Weighted feature extraction using a genetic algorithm for intrusion detection", Conqress on Evolusionary Computation: vol.3, , Carberra, Australia. Piscataway, NJ, USA: IEEE, (**2003**) December 8-12, pp. 1669-1675.
[16] T. Shon, Y. Kim, C. Lee and J. Moon, "A machine learning framework for network anomly detection using svm and ga", Proceedings of 6th Annual IEEE Workshop on infromation Assurance and Security, Piscataway, NJ, USAS: IEEE Computer Scociety, (**2005**) June 15-17, pp. 176-183.
[17] HC. Kim, "Face recognition using LDA mixture model", In: Proceedings int conf. on pattern recognition, (**2002**).
[18] P. Kemal, "A cascade learning system for classification of diabetes disease: Generalized Discriminant Aanalysis and Least Square Support Vector Machine", Expert Systems with Applications, vol. 34, (**2008**), pp. 482-487.
[19] F. Yaghouby, A. Ayatollahi and R. Soleimani, "Classification of Cardiac Abnormalities Using Reduced Features of Heart Rate Variability Signal", World Applied Science Journal, ISSN 1818-4952, vol. 6, no. 11, (**2009**), pp. 1547-1554.
[20] KDDCup99dataset,2009http://kdd.ics.uci.edu/databases/kddcup99/ kddcup2009.html.
[21] G. Mercier and M. Lennon, "Support vector machine for hyper-spectral image classification with spectral-based kernels", In proceedings of the international geosciences and remote sensing symposium, (**2003**), pp. 288-90.
[22] T. Joachims, "Making Large-Scale SVM Learning Practical", LS8-Report, University of Dortmund, LS VIII-Repprt.
[23] C W. Hsu and C J Lin, "A comparision of methods for multiclass support vector machines", IEEE Transactions on Neural Networks, vol. 13, no. 2, (**2002**), pp. 415-425.
[24] M.-H. Yang, "Kernel Eigenface vs. Kernel Fisherface: Face recognition using kernel methods", In Automatic Face and Gesture Recognition, Proceedings, Fifthe IEEE International Conference, (**2002**), pp. 215-220.
[25] G. K. Kuchimanchi, V. V. Phoha, K. S. Balagani and S. R. Gaddam, "Dimension Reduction Using Feature Extraction Methods for Real-time Misuse Detection Systems", Proceedings of the IEEE on Information, (**2004**).
[26] V. Venkatachalam and S. Selvan, "An approach for reducing the computational complexity of LAMSTAR intrusion detection system using principal component analysis", International Journal of Computer Science, vol. 2, no. 1, (**2007**), pp. 76-84.
[27] C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition", Data Mining and Knowledge Discovery, vol. 2, no. 2, (**1998),** pp. 121-167.
[28] http://download.joachims.org/svm_multiclass/current/svm_multiclass.tar.gz
[29] J. Weston and C. Watkins, "Multi-Class Support Vector Machines", CSD-TR-98-04, Royal Holloway. 1st Edn. Department of Computer Science, University of London, London, (**1998**).

Authors

Shailendra Singh, he is an Asistant Professor in, Department of Information Technology at Rajiv Gandhi Technological University, Bhopal, India. He has publised more than 15 papers in international journals and conference proceedings His research interest include datamining and network security.He is a life member of ISTE, Associte member of Institution of Engineers (India) member of International Association of Computer Science and Information Technology (IACSIT) Singapore and member IEEE.

Sanjay Silakari, he is an Professor and Head, Department of Computer Science and Engineering at Rajiv Gandhi Technological University, Bhopal, India. He has awarded Ph.D. degree in Computer Science & Engg. He posses more than 16 years of experience in teaching under-graduate and post-graduate classes. He has publised more than 60 papers in international, national journals and conference proceedings. He is member of International Association of Computer Science and Information Technology (IACSIT) Singapore.