International Journal of Computer and Internet Security.
ISSN 0974-2247 Volume 4, Number 2 (2012), pp. 51-58

http://www.irphouse.com

# Cyber Attack Detection and Classification Using Parallel Support Vector Machine

**Mital Patel and Yogdhar Pandey**

*Department of Computer Science & Engineering*
*Sagar Institute of Research & Technology, RGPV, Bhopal, M.P., India*
*E-mail: Mitalpatel95@gmail.com, P_yogdhar@yahoo.co.in*

**Abstract**

Cyber attack is becoming a critical issue of organizational information systems. A number of cyber attack detection and classification methods have been introduced with different levels of success that is used as a countermeasure to preserve data integrity and system availability from attacks. The classification of attacks against computer network is becoming a harder problem to solve in the field of network security. This paper describes a Subset Selection Decision Fusion method to choose features (attributes) of KDDCUP 1999 intrusion detection dataset. Selection algorithm for distributed cyber attack detection and classification is proposed. Different types of attacks together with the normal condition of the network are modeled as different classes of the network data. We proposed Parallel Support Vector Machine (pSVM) algorithm for detection and classification of cyber attack dataset. Support Vector Machines (SVM) are the classifiers which were originally designed for binary c1assification. The c1assificatioin applications can solve multi-class problems. Result shows that pSVM gives more detection accuracy for classes and comparable to false alarm rate.

**Keywords:** Distributed Cyber Attack Detection and Classification, Subset Selection Decision Fusion, Parallel Support Vector Machine, KDDCUP'99 and Confusion Matrix.

## 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. Subset selection decision fusion method plays a key role in cyber attack detection. It has been shown that redundant and/or irrelevant features may severely affect the accuracy of learning algorithms. The SDF is very powerful and popular data mining algorithm for decision-making and classification problems. It has been using in many real life applications like medical diagnosis, radar signal classification, weather prediction, credit approval, and fraud detection etc.

In this paper we proposed Parallel Support Vector Machine (pSVM) algorithm for detection and classification of cyber attack dataset. As we know that the performance of support vector machine is greatly depend on the kernel function used by SVM. Therefore, we modified the Gaussian kernel function in data dependent way in order to improve the efficiency of the classifiers. 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 PSVM performs better than SDF. The classification accuracy of PSVM remarkably improve (accuracy for Normal class as well as DOS class is almost 100%) and comparable to false alarm rate and training, testing times. The remainder of the paper is organized as follows. In Section II, we present KDDCUP'99 dataset. The Preliminary work of distributed cyber attack detection and classification is formulated in Section III. In section IV PSVM is proposed. The proposed Parallel Support Vector Machine algorithm is evaluated using KDD1999 intrusion detection datasets. The performance is analyzed by comparing to the feature subset selection and parallel support vector algorithm. Conclusions are provided in Section V.

## Related Work

Support Vector Machine is a powerful tool to classify cyber attacks. But still it has some drawback. The first drawback is that SVM is very sensitive for attacks .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 [13] is one of the improvements made on the traditional SVM. Several machine learning paradigms including Artificial Neural Network [14], Linear Genetic Programming (LGP) [15], Data Mining [16], etc. have been investigated for the classification of cyber attack. 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. Decision-tree-based support vector machine which combines support vector machines and decision tree can be an effective way for solving multi-class problems. This method can decrease the training and testing time, increasing the efficiency of the system [2]. Improved Support Vector Machine (iSVM) algorithm for classification of cyber attack dataset which gives 100% detection accuracy for Normal and Denial of Service (DOS) classes and comparable to false alarm rate, training, and testing times [2]. A new feature selection algorithm for distributed cyber attack detection and classification is proposed. Different types of attacks together with the normal condition of the network are modeled as different classes of the network data. Binary classifiers are used at local

sensors to distinguish each class from the rest.

## KDD CUP ‘’99 Data Set Description

To check performance of the proposed algorithm for distributed cyber attack detection and classification, we can evaluate it practically using KDD’99 intrusion detection datasets [6]. In KDD99 dataset these four attack classes (DoS, U2R,R2L, and probe) are divided into 22 different attack classes that tabulated in Table I. The 1999 KDD datasets are divided into two parts: the training dataset and the testing dataset. The testing dataset contains not only known attacks from the training data but also unknown attacks. Since 1999, KDD’99 has been the most wildly used data set for the evaluation of anomaly detection methods.This data set is prepared by Stolfo et al. [11] and is built based on the data captured in DARPA’98 IDS evaluation program [12]. DARPA’98 is about 4 gigabytes of compressed raw (binary) tcpdump data of 7 weeks of network traffic, which can be processed into about 5 million connection records, each with about 100 bytes. 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 (numeric) and 7 features (symbolic). To analysis the different results, there are standard metrics that have been developed for evaluating network intrusion detections. Detection Rate (DR) and false alarm rate are the two most famous metrics that have already been used. DR is computed as the ratio between the number of correctly detected attacks and the total number of attacks, while false alarm (false positive) rate is computed as the ratio between the number of normal connections that is incorrectly misclassified as attacks and the total number of normal connections.

**Table 1:** Different Types of Attacks in 10% KDD99 Dataset

| 4 Main Attack Classes | 22 Attack Classes | Samples |
|---|---|---|
| Normal | | 97277 |
| Denial of Service (DoS) | back, land, neptune, pod, smurt, teardrop | 391458 |
| Remote to User (R2L) | ftp_write, guess_passwd, imap, multihop, phf,spy, warezclient, warezmaster | 1126 |
| User to Root (U2R) | buffer_overflow, perl, loadmodule, rootkit | 52 |
| Probing(Information Gathering) | ipsweep, nmap, portsweep, satan | 4107 |

In the KDD Cup 99, the criteria used for evaluation of the participant entries is the Cost Per Test (CPT) computed using the confusion matrix and a given cost matrix .A Confusion Matrix (CM) is a square matrix in which each column corresponds to the predicted class, while rows correspond to the actual classes. An entry at row i and column j, CM (i, j), represents the number of misclassified instances that originally belong to class i, although incorrectly identified as a member of class j. The entries of the primary diagonal, CM (i, i), stand for the number of properly detected instances. Cost matrix is similarly defined, as well, and entry C (i, j) represents the cost penalty

for misclassifying an instance belonging to class i into class j. Cost matrix values employed for the KDD Cup 99 classifier learning contest are shown in Table 2. A Cost Per Test (CPT) is calculated by using the following formula:

$$\mathrm{PT} = 1/N \sum_{i=1}^{m} \quad \sum_{j=1}^{m} \mathrm{CM(i,j)} * \mathrm{C(i,j)} \tag{1}$$

Where CM and C is confusion matrix and cost matrix, respectively, and N represents the total number of test instances, m is the number of the classes in classification. The accuracy is based on the Percentage of Successful Prediction (PSP) on the test data set.

$$PSP = \frac{number\ of\ successful\ instance\ classification}{number\ of\ instance\ in\ the\ test\ set} \tag{2}$$

## Proposed Work

We proposed a new method for cyber attack classification based on parallel support vector machine based on distant feature set of attack attribute. All of the features are ranked based on their KullbackLeibler (K-L) distances, which is an alternative way to measure the importance of a feature in discriminating two classes. The features discriminating based on the equiliden distance formula for finding a similarity of features based on attack category. After calculation of discriminate we apply parallel support vector machine. SVM which was developed by Vapnikis one of the methods that is receiving increasing attention with remarkable results. SVM implements the principle of Structural Risk Minimization by constructing an optimal separating hyper plane in the hidden feature space, using quadratic programming to find a unique solution. Originally SVM was developed for pattern recognition problems. Recently, a regression version of SVM has emerged as an alternative and powerful technique to solve regression problems by introducing an alternative loss function. Although SVM has been successfully applied in many fields, there is a conspicuous problem appeared in the practical application of SVM. In parallel SVM machine first we reduced non-classified features data by distance matrix of binary pattern. From this concept, the cascade structure is developed by initializing the problem with a number of independent smaller optimizations and the partial results are combined in later stages in a hierarchical way, as shown in figure 1, supposing the training data subsets and are independent among each other.

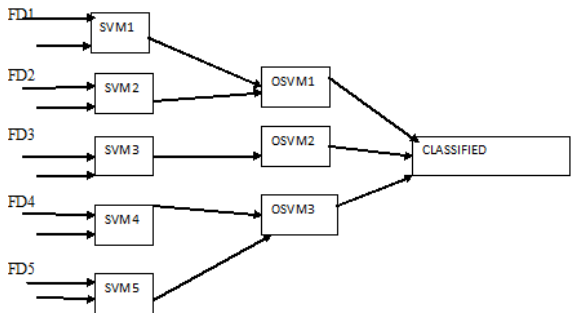


**Figure 1:** Cascaded SVM

This figure shows that cascaded support vector machine, in this machine we passed five stage of features discernment and all these passes to optimized support vector machine for the processing of classification.

**Step for Data Preprocessing**

- Transform data to the format of an SVM
- Conduct scaling on the data
- Consider the RBF kernel K(x; y)
- Use cross-validation to 2nd the best parameter C and
- Use the best parameter C and to train the whole training set
- Generate formatted data.

**Step of Cyber Data Classification**

- Read preprocessing data
- For all the classes are represented

BEGIN
Find class with no attribute
Find class at Max cross product rate
Find the class at half cross product
REPEAT
Pointer= False
Find the intervals of hyper plane
If the end condition is met
Pointer = True
If the first interval has better results we should Use this, otherwise the other
Find the class evaluation after cross product class
Instances middle times
UNTIL pointer= False

**END**

- Multiply all the classes with the best factor obtained;
- Data classified.

## Experiments and Results

All the experiments were performed on an Intel ® Core ™ i3 with a 2.27GHz processor with 2 GB of RAM. We used MATLAB version 2009 software. Figure 2 and Figure 3 shows the results obtained by using various classification techniques. The results of the comparison of proposed algorithm with SDF are shown in Table 3.

**Table 2:** Comparison of Detection rate and False Alarm Rate

| Classifiers | Train Data | Detection Rate | False Alarm Rate |
|---|---|---|---|
| SDF | .5 | 89.192 | 4.57657 |
| pSVM | .5 | 92.559 | 6.03219 |

Our experiment is split into three main steps. In the first steps, we prepare different dataset for training and testing. Second, we apply subset selection decision fusion algorithm (SDF) to the dataset.

The original KDDCUP1999 dataset, to select most discriminate features for cyber attack detection. Third, we classify the cyber attacks by using parallel SVM (pSVM) as classifier. Table II shows the comparison.

**The Detection of Attack and Normal Pattern Can be Generalized as Follows**

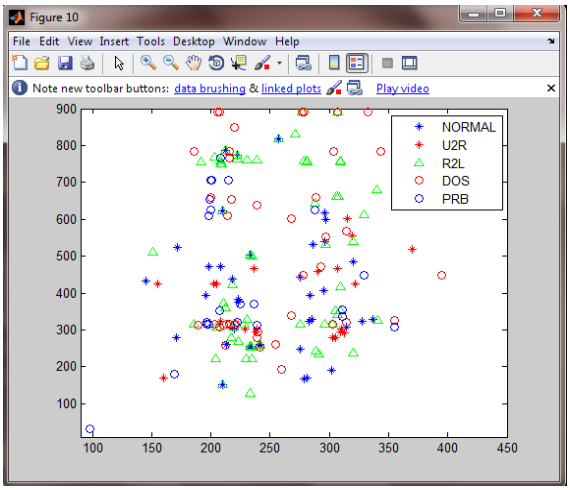


**Figure 2:** Result of Various Classification Techniques

***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.

$$\mathrm{DR} = \frac{Total\ no\ of\ detected\ attacks}{Total\ no\ of\ attack\ detection} \times 100\%$$

***Comparison of False Alarm Rate****:* False Alarm Rate (FAR) refers to the proportion that normal data is falsely detected as attack behavior.

$$\mathrm{FAR} = \frac{Total\ no\ of\ normal\ processes}{Total\ no\ of\ misclassified\ processes} \times 100\%$$

Confusion matrix contains information actual and predicted classifications done by a classifier. The performance of cyber attack detection system is commonly evaluated using the data in a matrix. Table III shows the confusion matrix.

**Table 3:** Confusion Matrix

| Predicted<br>Actual | Normal | Attack |
|---|---|---|
| Normal | True Negative (TN) | False Positive (FP) |
| Attack | False Negative (FN) | True Positive (TP) |

## Conclusion

This paper presents new cyber attack detection and classification system to classify cyber attacks. In this paper, we developed the performance of IDS using parallel support vector machine for distributed cyber attack detection and classification. The new PSVM is shown more efficient for detection and classification of different types of cyber attacks compared to SDF. The experimental results on KDD99 benchmark dataset manifest that proposed algorithm achieved high detection rate on different types of network attacks.

## Reference

[1] Hoa Dinh Nguyen and Qi Cheng, “An Efficient Feature Selection Method For Distributed Cyber Attack Detection and Classification”,978-1-4244-9848-2/11$26.00 © 2011 IEEE

[2] Shailendra Singh Member, IEEE, IAENG, Sanjay Agrawal, Murtaza, A. Rizvi and Ramjeevan Singh Thakur, “Improved Support Vector Machine for Cyber Attack Detection”, Proceeding of The World Congress on Engineering and Computer Science 2011 Vol I WCECS 2011, October 19-21, 2011, San Francisci, USA

[3] Snehal A.Mulay, P.R.Devale, G.V.Garje,”Intrusion Detection System Using Support Vector Machine and Decision Tree”, International Journal Of Computer Applications (0975-8887), Volume 3- No3,June 2010

[4] Dewan Md. Farid, Nouria Harbi, EmnaBahri, Mohammad Zahidur Rahman, Chowdhury Mofizur Rahman, "Attack Classification in Adaptive Intrusion Detection Using Decision Tree", World Academy of Science, Engineering and technology 63 2010.

[5] P. Srinivasulu, R. Satya Prasad and I. Ramesh Babu, "Intelligent Network Intrusion Detection Using DT and BN Classification Techniques", Int. J. Advance. Soft Compt. Appl., Vol.2, No. 1, March 2010 ISSN 2074-8523; Copyright © ICSRS Publication, 2010

[6] KDD CUP 1999. Availabe on: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html October 2007

[7] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set", Proceeding of the 2009 IEEE Symposium on Computational Intelligence In Security and Defence Applications (CISDA 2009)

[8] S. B. Kotsiantis, "Supervised Machine Learning:" A Review of Classification Techniques", Informatica 31 (2007) 249- 268

[9] Dr. Adnan Mohsin Abdulazeez Brifcani and Adel Sabry Issa, "Intrusion Detection and Attack Classifier Based on Three Techniques: A Comparative Study", Eng. & Tech. Journal, Vol.29, No.2, 2011

[10] T. Subbulakshmi, S. Mercy Shaliinie and A. Ramamoorthi, "Detection and Classification of DDoS Attacks Using Machine Learning Algorithms", European Journal of Scientific Research ISSN 1450-216X Vol.47 No. 3 (2010), pp. 334-346 © EuroJournals Publishing, Inc. 2010

[11] G. MeeraGandhi, Kumaravel Appavoo and S.K. Srivatsa, "Effective Network Intrusion Detection using Classifier Decision Trees and Decision Rules", Int. J. Advanced Networking and Applications Volume: 02, Issue: 03, Pages: 686-692 (2010)

[12] Keng-Pei Lin and Ming-Syan Chen, Fellow, IEEE, "On the Design and Analysis of the Privacy-Preserving SVM Classifier", IEEE Transactions on Knowledge and Data Engineering, Vol. 23, No. 11, November 2011

[13] Xiong, Sheng-Wu, Liu Hong-bing, Niu Xiao-xiao, "Fuzzy Support Vector Machine Based on FCM Clustering", Proceeding of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, China, Aug 18-21, IEEE, p. 2608-2613, 2005

[14] A. K. Ghosh and A. Schwartzbard, "A Study in Using Neural Networks for Anomaly and Misuse Detection", Proceeding of the 8th USENIZ Security Symposium, pp. 23-36. Washington, D.C. US. 1999

[15] Mukkamala S., Sung AH, Abraham A., "Modeling Intrusion Detection Systems using Linear Genetic Programming Approach", The 17th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, Innovation in Applied Artificial Intelligence.

[16] W. Lee, S. J. Stolfo and K. Mok, "Datamining in Workflow Environments: Experence in Intrusion Detection, Proceeding of the Conference on Knowledge Discovery and Datamining (KDD-99), 1999