Direct Kernel Method for Machine Learning with Support Vector Machine

Sangpal Sopan Sarkate[1] Asst. Prof. Hirendra Hajare[2]
[1]PG Student [2]Head of Department
[1,2]Department of Computer Science Engineering
[1,2]BIT, Ballarpur (MH), India

Abstract— The Internet can be used for connecting the world from end to end user. The user end to end connection need to be secure transmission of the data. Various types of attacks occur during the transmission of data and this activity called as an intrusion. To identify the intrusion need to develop the Intrusion Detection System, this is able to distinguish between from normal and attacker users. These users can be distinguished by their behavior. The behaviors of attacker can suspicious which is identified by learning their activity which they are doing on the internet. To learn the behavior of the intruder and the normal user need to develop an efficient system. This system can be learned various aspects or features of the intrusive behavior and extract that feature for recognizing them from regular traffic data. The daily network traffic data consist of large amounts of data as well as having various features in more quantity.

Key words: Kernel Method, Machine Learning

I. INTRODUCTION

Internet is called as worldwide public networks by which it connects the entire world. The potential of the internet and evolution of the internet makes changes in the business model of various organizations. From 2nd decade of the start of 21st century use of internet increases day by day. Having facility of devices like smart-phones, tablets, laptops, and computer's networks which increase craze of the internet. These devices are easy to manage and link to the internet world. Internet facility used by the people to run businesses, to connect offices, for entertainment and for educational purpose also. Considering part of trade on the internet then it will be having two sides. From one of them which is having tremendous potential of the internet in terms of reaching the end users. At the same time another side of the business brings in lots of risk. On the internet, users can be divided into two types, i.e. some of them are harmless users and some of harmful users. The information system is available to the malicious user's i.e. harmful user as well as harmless users also. There is a possibility of malicious or harmful users can get access to internal system of an organization's in various reasons. The cost of the internet accessibility and processing of internet decreases, then the lots of organization are becoming defenseless to a varied variety of cyber threats, such as network intrusion. By using the internet, awareness of the network security increases day by day. The traditional solutions which are used for protecting internet applications and computer networks those are not capable to protect against the threats from occurs through cyber-attack techniques such as DoS attack and as well as computer malware developed by the attacker or intruder.

II. RELATED WORK

It can be considered that in the cyber world, there are various types of cyber-attacks and to restrict to those various protection mechanisms are implemented. IDS is one of the mechanism which are used for the reorganization the attacks and work to restrict it. Hence, now-days IDS is a major research problem in the network protection mechanism. For the development of IDS various machine learning approaches are used. According to detection strategy the IDS can be divided into two types one is misuse detection and another is anomaly detection which is popular in recent days. Misuse detection is a signature based detection method in which it uses the well-defined pattern of the intrusion to finding out the attacks or intruder. The misuse detection [12] technique gives better accurate results of the known attacks. Anomaly detection technique [13] is which it defines the behavior pattern of intrusion by which it can recognize the attacks or intrusion. This mechanism gives better accuracy results against unknown attacks.

The various machine learning approaches are present,

A. Pattern Classifier

It is an action can be performed on raw data and activity in the category of the data. The various pattern classification techniques are performed by using supervised and unsupervised techniques. In supervised learning it uses training datasets to create a function in which this training dataset contains pair of the input and output vector of the training data.

B. Single Classifier

In the IDS there is single machine learning algorithm is used as a classifier. These classifier algorithms are k- nearest neighbour algorithm, support vector machine, Artificial Neural Network, Decision Tree, Self-Organizing Maps, Naïve Bayes Networks, Genetic Algorithms, and Fuzzy Logic.

C. Hybrid Classifier

To get the best possible accuracy from the IDS it's a main goal of it. To achieve that goal the combination of the classifier are used. By combining several machines learning algorithm the performance of the IDS significantly increases. In hybrid classifier approach, there are two components present in which one for using the raw data to generate instant result and second for using that instant result to get the final result.

D. Ensemble Classifier

This classifier approach is used to improve the performance of the single classifier by combining various weak algorithms or weak learner. Weak learner used to train

different training data samples and result used for improving the performance.

III. PROPOSED SYSTEM

A. Intrusion and Intrusion Detection in Machine Learning

It can be defined as several sets of actions that can be used to threaten the integrity, availability or privacy of a network system to access the important information. The subsequent scenarios are examples of the intrusions:

1) A member of staff browses by way of his/her boss, employee reviews.
2) A user takes benefit of a defect in file server program to obtain entry to and then to corrupt another user's files.
3) A client exploits a defect in system program to get super-user status.
4) An attacker uses a script to "crack" the password of another user on a computer.
5) An attacker installs a "snooping program" on a computer to examine network traffic, which frequently contains user's passwords and additional responsive data.
6) An attacker modifies router tables in the network to avoid the sending of a message to a specific computer.

B. Intrusion Detection

ID is the process of network and monitoring them for symbols of intrusion, definite as attempts to detour the protection method of a computer or network which cooperation the secrecy, honesty and accessibility of information resources. The various supervised and unsupervised machine learning approaches are implemented to detect the intrusion in the network. For the detection of intrusion in the network need to develop the systems which are used machine learning concept to learn the behavior of the intruder. This supervised and unsupervised machine learning approaches are Nuero-Fuzzy algorithm, Artificial Neural network, etc.

C. Intrusion Detection System for Machine Learning

It is a mixture of software and hardware that attempts to perform intrusion detection raise the alarm when the possibility of intrusion happens. By comparison an ID system does for a network what an antivirus programs package does for files that come into a system. It examines the content of network traffic to seem for and prevent probable attacks, presently as an antivirus program container inspects the stuffing of arriving records, e-mail attachments, lively web substance, and so onwards to seem to be for virus spot or for probable malicious events.

For the development of the IDS needs to implement the machine learning approach where they used different classifier. With the help of classifier they distinguish the intruder and the normal user behavior. Now days IDS working based on machine learning method hence to need to study the detail analysis of the IDS in this section. This study gives briefly introduction of functionality of the IDS, component of the IDS and working of IDS with machine learning approaches.

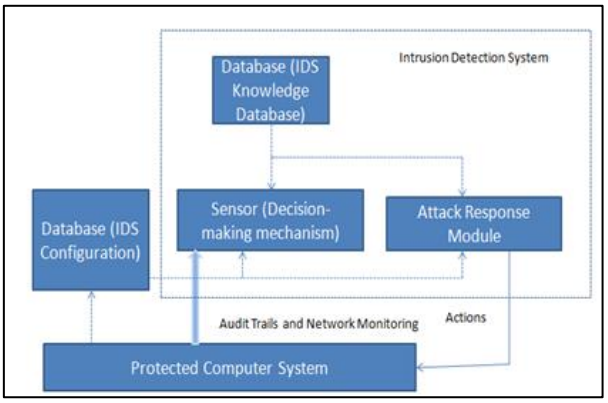

Figure 3.1: Basic IDS System

D. Structure and Architecture of Intrusion Detection System

An ID system forever has its primary element a sensor (an investigation engine) that is dependable for detecting intrusion. This sensor contains a decision-making mechanism concerning intrusion. Sensor receives crude data from three main information sources. These are IDS information base, SYSLOG and audit trails. The SUSLOG might involve, for example, design of the file system, user authentication, record information of the packets, etc. this information makes the root for a further decision making procedure.

In figure 3.1 shows that, the component of basic IDS system. The flow of information through the IDS knowledge database to decision making mechanism as well as to attack response module is present.

ID system can be set as whichever centralized (for example, bodily included within a firewall) or circulated. Distributed IDS consist of many ID Systems above a big network, each and every one of which communicates with each one extra. Further complicated system follows an agent construction principle where small independent modules are prepared on a per-host root crosswise the secured network [6].

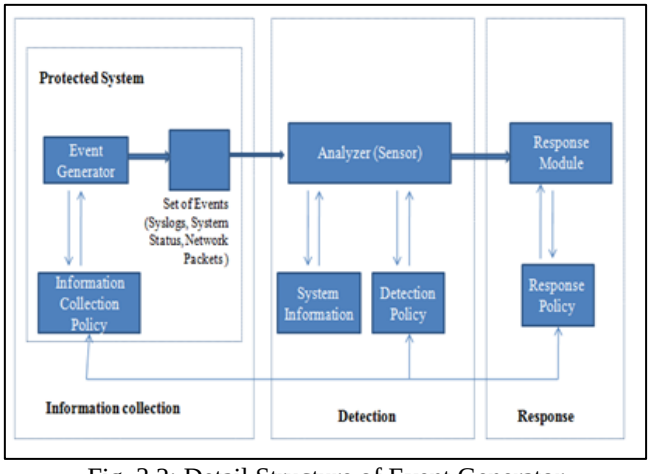

Fig. 3.2: Detail Structure of Event Generator

One multi-agent structural design solution, which originated in 1994, is AAFID (autonomous Agents for ID). It uses agents that examine a definite part of the activities of the system they are located in on at the time

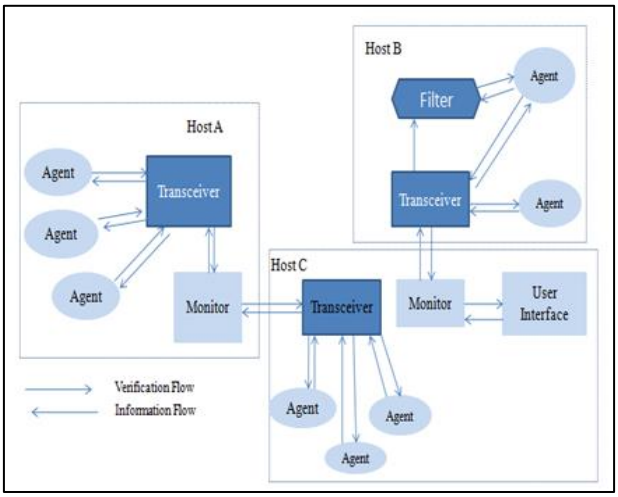

Fig. 3.3: Autonomous Agents for Intrusion Detection

IV. EXPERIMENTAL RESULT

For the experiments the used datasets are KDDCup99 dataset which contains training and testing dataset. In the kddcup99 datasets there are 5 million records present in training and 2 million records in testing. For both it contains 41 features and one class label which shows its category i.e. normal type or attack types (probe, dos, r2l, u2r).

The NSL-KDD is the revised version of the KDDCup99 datasets which is having the same pattern of records. It contains training and testing data which contain 41 features and class label normal and attack type. One number feature added which helps to know in which category their present, but for our conditions we remove that.

Kyoto 2006+ dataset is the newly dataset in research concept which are contains the real network traffic data. In this dataset the data which contains from the various honey pots and regular servers present in the Kyoto University. This dataset collects in the period of November 2006 and August 2009.

For the result extraction of KDDCup99 dataset we select the "kddcup.data_10_percent"dataset as training dataset. For testing, we choosing 15500 records from the "kddcup.data.corrected" and for unlabeled testing dataset select 15455 records from "kddcup.newtestdata_10%_unlabeled". For the NSL-KDD training dataset "KDDTrain+_20Percent"is used. Labeled testing dataset "KDDTest-21" and for unlabeled testing datasets 22544 records are selected from "KDDTest+". For the Kyoto 2006+ dataset select the last day collection of records which present in "20090831.txt". Here we select only those features which are similar to the KDDCup99 and NSL-KDD datasets .The total number of features selected for the process is 14, which are similar to the both previous datasets and 17th number feature which contain the status of the records.

Tables (1) contains the outcome of the record KDDCup99, table (2) contain outcome of the record NSL-KDD and table (3) showing the outcome of the record Kyoto 2006+ datasets which are used in the process.

Number of records	Before removing duplicates	After removing duplicates
Labeled Training	49406	38614
Labeled Testing	15500	15500
Unlabeled testing	15455	15400

Table 1: KDDCup99 Dataset

Number of records	Before removing duplicates	After removing duplicates
Labeled Training	25192	25192
Labeled Testing	11850	11832
Unlabeled testing	22544	22487

Table 2: NSL-KDD Dataset

Number of records	Before removing duplicates	After removing duplicates
Labeled Training	25000	18832
Labeled Testing	20000	16976
Unlabeled testing	15000	13458

Table 3: Kyoto 2006+ Dataset

All datasets are sent to the process with selecting last feature value is the class value. This class value helps SVM to classify the records in the high dimensional space and categorizing in the normal and attack types.

A. Performance measures

For the measuring implementation of the any machine learning system there are various parameters. For our system we choose three parameters these are accuracy, detection rate and false positive rate.

Accuracy = $\frac{TP+TN}{TP+TP+FP+FP}$,

Detection Rate = $\frac{TP}{(TP+FN)}$,

False Positive Rate = $\frac{FP}{(FP+TN)}$.

Classification of Records	Attack type	Normal
Attack type	TP	FN
Normal	FP	TN

Table 4: Terminology of Records Classification

In table 4 terms are represent to get category of the records according to their type which they come in and help to measure the functioning of the method.

The outcome of the machine learning with respect to accuracy, detection rate and false positive rate are shown in chart form.

Accuracy of the KDDCup99 dataset, NSL-KDD dataset and Kyoto 2006+ datasets are 99.87, 98.81, and 97.15 respectively which are shown in figure 4.1. The accuracy of the Kyoto 2006+ dataset is achieving low as compare to the other two datasets.

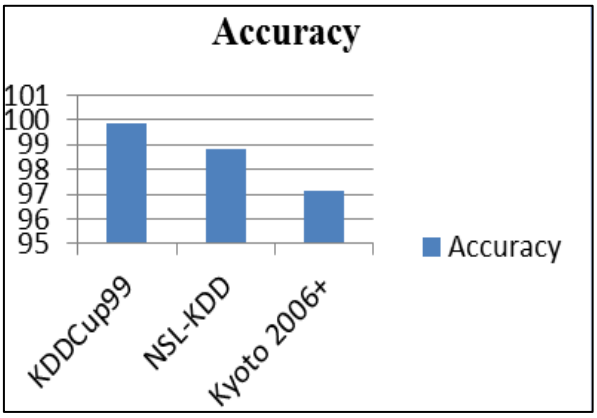

Fig. 4.1: Comparison of Accuracy

The detection rate of three datasets is shown in the figure 4.2 which shows that our method archives high detection rate, but for the KDDCup99 dataset detection rate is less. With the kernel direct method it achieves in high range, which is essential for the machine learning.

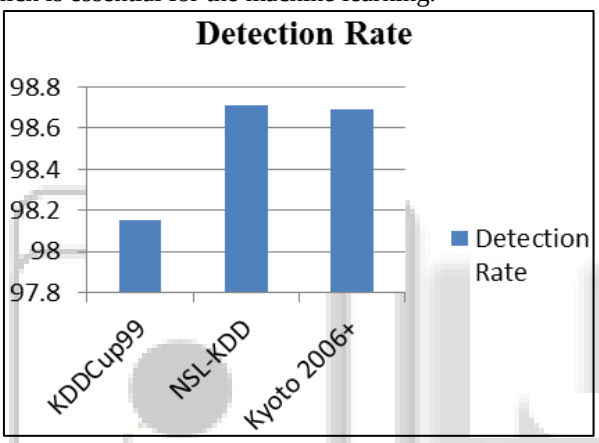

Fig. 4.2: Comparison of Detection Rate

The low false positive rate of the proposed method is the major advantage of the given system. For KDDCup99 and NSL-KDD dataset direct kernel method gives good performance, which is shown in the figure 4.3.

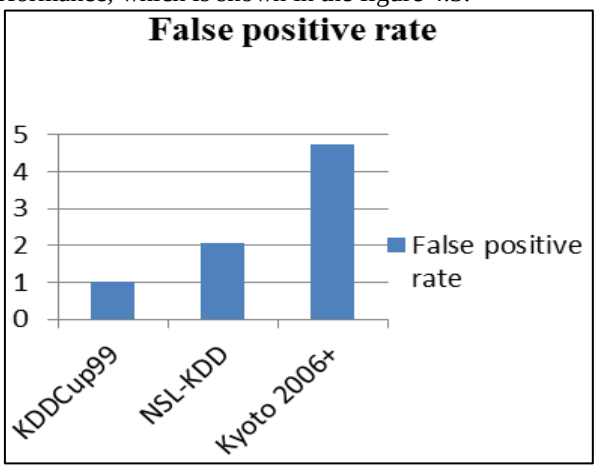

Fig. 4.3: Comparison of False Positive Rate

V. CONCLUSION

We present a direct method to build kernel method for machine learning. This method is same as the kernel functions which are used in the SVM. Various machine learning algorithms are present to work with SVM which are capable to extract the results of classification. But SVM is used for classification with the machine learning algorithm not to use in feature extraction. In implemented system we use the SVM as classifier as well as the machine learner to extract the features to plot in the high dimensional feature space which is used for classification. To enhance the performance of the SVM during machine learning and classification the direct kernel method performs well.

The kernel direct method is the arithmetical and statistical formulation which can be used easily but for implimenting it increases complexity. For machine learning, kernel methods are used by various researchers such as RBF, Polynomial kernel, linear kernel. SVM is used for the classificatin of KDDCup99 datasets lots of time but for the NSL-KDD and Kyoto 2006+ dataset is not used by researchers. The datasets KDDCup99, NSL-KDD and Kyoto 2006+ datset gives the network information which is important to know the internet network threats and valuable information.

REFERENCES

[1] J. Song, H. Takakura, Y. Okabe, M. Eto, D. Inoue, and K. Nakao, "Statistical analysis of honeypot data and building of Kyoto 2006+ dataset for nids evaluation," in Proc. 1st Workshop BADGERS, 2011.

[2] R. Kemmerer, G. Vigna, "Intrusion detection: a brief history and overview", Security and Privacy, Supplements to Computer, IEEE Vol. 35, Issue 4, pp 27-30,2002.

[3] Boser B. E., Guyon I. M.,Vapnik V. N.,"A training algorithm for optimal margin classifiers". Proceedings of the fifth annual workshop on Computational learning theory – COLT '92, p. 144, 1992.

[4] Cortes C., Vapnik V., "Support-vector networks", in Machine Learning, 20 (3), 273–297., 1995.

[5] Mingrui Wu, Bernhard Sch¨olkopf, G¨okhan Bakir, "A Direct Method for Building Sparse Kernel Learning Algorithms"Journal of Machine Learning Research, 7th Edition, 603–624, 2006.

[6] 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.

[7] 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.

[8] Chih-Fong Tsai, Yu-Feng Hsu, Chia-Ying Lin, Wei-Yang Lin, "Intrusion detection by machine learning: A review", Expert Systems with Applications 36, 11994–12000, 2009.

[9] Salama, M., Eid, H., Ramadan, R., Darwish, A., & Hassanien, A. "Hybrid intelligent intrusion detection scheme" Soft computing in industrial applications, 293-303, 2011.

[10] Mukkamala, Srinivas, and Andrew H. Sung. "Artificial intelligent techniques for intrusion detection." Systems, Man and Cybernetics, IEEE International Conference on. Vol. 2. IEEE, 2003.

[11] Srinivas Mukkamalaa, Andrew H. Sunga, Ajith Abraham, "Intrusion detection using an ensemble of intelligent paradigms" Journal of Network and Computer Applications 28, 167–182, 2005.
[12] Shih-Wei Lin , Kuo-Ching Ying , Shih-Chieh Chen , Zne-Jung Lee, " Particle swarm optimization for parameter determination and feature selection of support vector machines" , Expert Systems with Applications 35, 1817–1824, 2008.
[13] Chunhua Gu,Xueqin Zhang, "A Rough Set and SVM Based Intrusion Detection Classifier", Second International Workshop on Computer Science and Engineering, 2009.

