IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 12, December 2017

# Direct Kernel Method for Machine Learning With Support Vector Machine

**Sangpal Sopan Sarkate[1], Prof. Riya Qureshi[2]**

M.Tech Student, CS Department, BIT, Ballarsha, India [1]

HOD,CS Department, BIT, Ballarsha, India [2]

**Abstract** : Support vector machine (SVM) based intrusion detection system (IDS) presently working as the machine learning approach for classification. It helps to detect new attacks from the datasets which are used in the machine learning. At IDS, the task of the machine learning method is to construct a projectile model which can be distinguished between normal and illegitimate activity. Any IDS can be developed to get high accuracy, high detection rate and low false positive rate, which show the efficiency of that intrusion detection system. In this paper, we use a direct kernel method with SVM classifier to get the high accuracy and detection rate, also low false positive rate. For the performance evaluation of the projected system we use KDDCup99 dataset, NSL-KDD dataset and Kyoto 2006+ datasets.

**Keywords**: Machine learning, SVM, datasets, kernel methods, IDS.

## I. INTRODUCTION

Now a day's internet and online procedures are so much important in day to day life. They are used like the important component of online procedures and the business part. In the case of unwanted access or unauthorized user creates threats to the security of the online procedure hence it is essential to develop a security system for the user networks. Barriers such as the firewall and antivirus software provide security at local system. But for entire network its capacity reduced that's why needs to apply intrusion detection system which was introduced by the Anderson [1] in 1980. The goal of the IDSs is to find out the malicious activity or attacks to the system. Network based IDS is precious tool used in the defense in depth of the computer network. In general, there are two types of IDSs according to the detection policy they do misuse detection and anomaly detection. Misuse detection is like a signature based system in which pattern reorganization techniques, finding out the pattern of attacks with matching know patterns. The misuse detection technique is accurate and efficient to the known attack, but it creates high false alarm to the unknown patterns, because it powerless to detect it. In the case of anomaly detection, the datasets in which pattern of the normal system activity stored from that it creates an outline profile of the patterns which use in to the detection system. Anomaly detection technique is useful in the finding out identified and unspecified attacks by using the baseline profile learning. In anomaly IDS commonly used machine learning and statistical learning technique because its performance is more accurate and useful in finding intruders. Over the decades the various estimation and learning methods utilize kernels in the machine learning. Kernel methods have a stronger slope towards the mathematical concept. Hence, as compare to the former machine learning method, it has interest of statistics and mathematics society. Kernel methods are used in various intrusion detection techniques and for machine learning. Support vector machine is one of the tools which used for classification in the intrusion detection for machine learning with kernels. The Problem solving approach of support vector machine is like a typical quadratic optimization problem which was affected by dimensions of features which are present in the datasets and size of the dataset. Feature selection or attribute reduction helps to save the time and the space of the SVM classification. For large scale and high dimensional dataset general SVM faces the problem of memory storage and time consumption. A traditional SVM solver is able to resolve dual quadratic optimization problem easily i.e. the binary function is used to organize the datasets by a support vector machine classifier. The support vector machine classifier is determined by the set of support vector.

$$K(Xi,Xj) = \phi(x) \cdot \phi(y) \quad \ldots. \quad (1)$$

The kernel function (1) by which non-linear samples of the datasets from the input space can be mapped in high dimensional space and becomes linear separable in space. The common kernels are linear kernel, polynomial kernel, and radial basis function (RBF) kernels.

In this paper section 2 with related work, section 3 for the proposed system, section 4 gives an experimental analysis and result followed by section 5 with the conclusion.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 12, December 2017

## II. RELATED WORK

In the IDSs there are variety of machine learning approaches, which are used for misuse detection and anomaly detection. An anomaly-based network IDs technique like genetic algorithm, self-organized feature map (SOFM) is used with SVM [9]. Multilayer support vector machine used to compare with SVM on NSL- KDD datasets [2]. For selecting the features by using wrapper method and the filter method Least Square Support Vector Machine (LSSVM) [11] is applied. For handling the linear and non-linear input samples apply mutual information based algorithm in the Least Square Support Vector Machine based IDS (LSSVM-IDS) [12] in which to finding out best possible feature for categorization. The SVM classifier is used with combining K-means, Fuzzy neural network for classification [13]. To capture the requirement of any concurrent IDSs the incremental learning methods can be useful and also get better computational correctness of real-time applications [5]. In [14] data mining techniques like decision trees and Naïve Bayes are combined with support vector machine to reduce false positive rate. A hierarchical Clustering algorithm with SVM classifier [15] proposed for intrusion detection. In case of selecting the features and determining parameter of the SVM a method like particle swarm optimization (PSO) is used which called as a PSO+SVM method [16]. In opposition to Knowledge Discovery in Databases with principle component analysis (PCA) are evaluated performances of different kernels of Support Vector Machine (SVM) [17].

## III. PRAPOSED SYSTEM

*A) Data Collection*

In the network security evaluation there are few datasets which can be available to use in the machine learning. The KDD cup99 [4], NSL- KDD dataset [2] and Kyoto 2006+ dataset [3] are publicly available dataset which are used in various research papers. In this dataset, they contain the different size of files and that dataset having various features in quantity and quality. These features can be helpful to get more information and can be used for the pervasive test in the validation of feature selection methods. That's why we are selecting those datasets for fair and equal comparison of the other methods.

*B) Preprocessing*

*a. Remove Duplicate*

Data redundancy and duplicity is problematic to present the various datasets, hence to overcome that we use 'remove duplicate method'. By removing the duplicate data the size of datasets decreases, which help for the machine learning and classification.

*b. Normalization*

It is another method which helps to categorize the records in a proper manner. By using normalization all the records are transformed to its score value or in weights. By using normalization process the records feature is re-cantered and rescaled. By re-cantering and rescaling those features retain to zero mean and unit variance.

*c. Machine learning*

For machine learning using a direct kernel method which is a mathematical expression formulated in SVM. Kernel direct method is the arithmetical and the statistical formulation by which we can calculate the score or weights of the features. This score value is the similarity measures between two records. This score values helps to find important and useful feature for selecting classification and evaluation. The training pattern or samples in the input space can be linear or non-linear separable in input space. Those samples or patterns can be mapped in the high dimensional space for the linear separable in the space. After using the direct kernel some specific feature are chosen to use in classification for the support vector machine

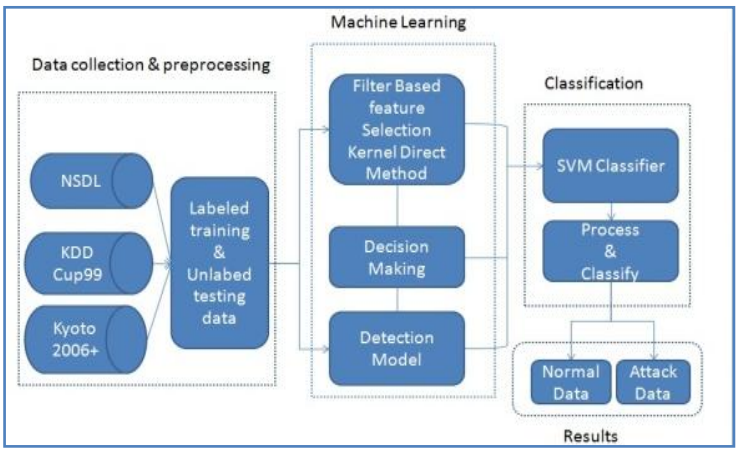


Figure 1: Architecture of Direct Kernel Method


DOI 10.17148/IJARCCE.2017.61202

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 12, December 2017

*d. Classification*

In this process the records in the datasets can be classified in normal or attack type class (i.e. Probe, DoS, R2L, U2R). For classification, we are using the SVM classification. In this classifier class are divided by using hyper plane, which categorized the record in the multidimensional space. SVM having binary classification and all records feature are in the numerical forms which are plotted in the space. These plotted points can be further divided in its class by hyper plane generating. We are getting all results in the matrix form which helps to find out the performance measures of the developed system.

## 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 17$^{th}$ 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.

Table 1: KDDCup99 Dataset

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

Table 2: NSL-KDD Dataset

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

Table 3: Kyoto 2006+ Dataset

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

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.

## V. PERFORMANCE MEASURE

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.

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


DOI 10.17148/IJARCCE.2017.61202

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 12, December 2017

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

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

Table 4: Terminology of Records Classification

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

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 2. The accuracy of the Kyoto 2006+ dataset is achieving low as compare to the other two datasets.

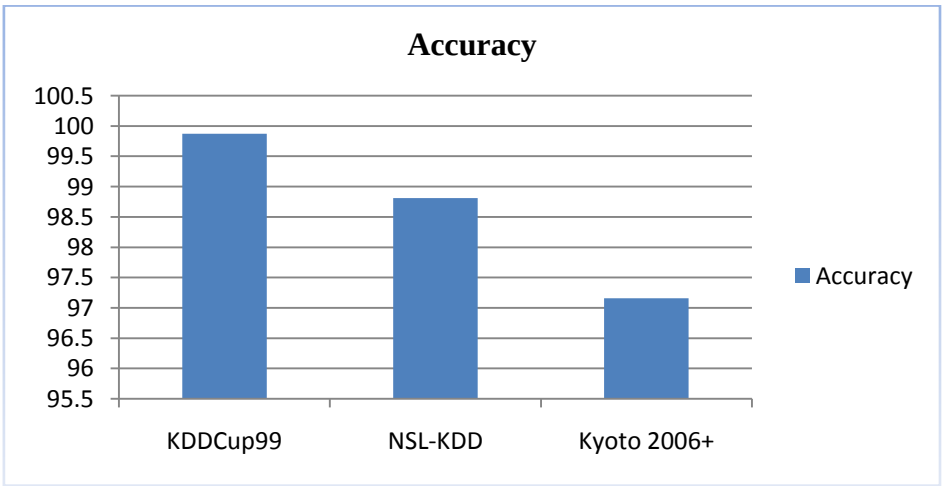


Figure 2: Comparison of Accuracy

The detection rate of three datasets is shown in the figure 3 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.

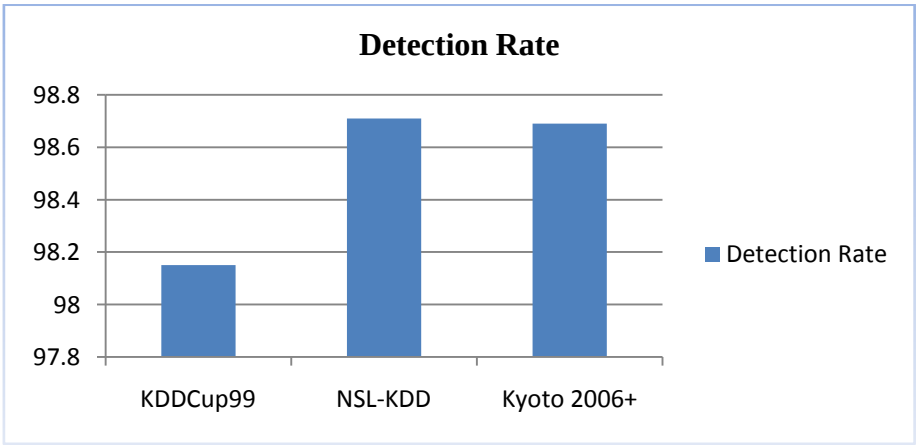


Figure 3: comparison of detection rate

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 12, December 2017

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.

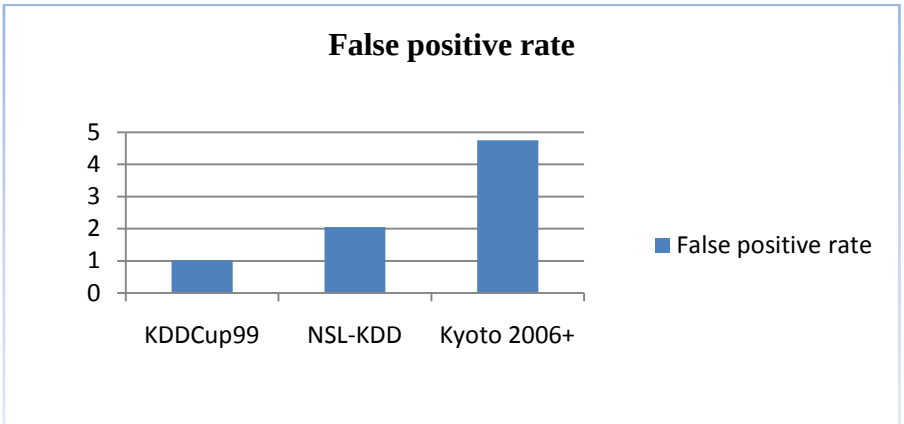


Figure 4: Comparison of False Positive Rate

## VI. CONCLUSION

The kernel direct method is the arithmetical and statistical formulation which can be used easily but for implimenting it increases complexity. For machine learning, it can be used by various researchers with SVM with only KDDCup99 datasets, not for the NSL-KDD and Kyoto 2006+. By applying kernel direct method to those datasets extract the more information and important feature selection. The kernel direct method with SVM classification gives the more accuracy and low false positive rate which is most important in machine learning. For the optimizing the datasets remove duplicate method is a good concept which reduced the dataset in small size which is essential to the machine learning. For the future research in the machine learning with various kernel methods such as linear, polynomial, RBF, cosine with the different classifiers are possible.

## REFRENCES

[1]. J.P. Anderson, "Computer security threat monitoring and surveillance", Technical Report, James P. Anderson Co., Fort Washington, PA, 1980.
[2]. M. Tavallaee, E. Bagheri, W. Lu, and A.-A. Ghorbani, "A detailed analysis of the kdd cup 99 data set," in Proc. 2nd IEEE SCISDA, 2009.
[3]. 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.
[4]. KDD Cup 1999 Data, University of California, Irvine, [online] 1999, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html (Accessed on 28th December 2016).
[5]. Roshan Chitrakar, Chuanhe Huang "Selection of Candidate Support Vectors in incremental SVM for network intrusion detection" in 45th Int. J computers& security, Elsevier, 2014.
[6]. Gisung Kim, Seungmin Lee, Sehun Kim, "A novel hybrid intrusion detection method integrating anomaly detection with misuse detection" in 41th Int. journal Expert Systems with Applications, Elsevier ,2014.
[7]. Depren, O., Topallar, M., Anarim, E., & Ciliz, M. K."An intelligent intrusion detection system for anomaly and misuse detection in computer networks". In 29th Int. journal Expert Systems with Applications, 2005.
[8]. Vapnik, V. "Statistical learning theory" John Wiley, 1998.
[9]. Anil S,Remya R, "A hybrid method based on Genetic Algorithm, Self-Organised Feature Map, and Support Vector Machine for better Network Anomaly Detection" in 4th ICCCNT, Tiruchengode, India, 2013.
[10]. Milad Aghamohammadi, Morteza Analoui "A Comparison of Support Vector Machine and Multi-Level Support Vector Machine on Intrusion Detection" in WCSIT, Vol. 2, No. 7, 2012.
[11]. Mohammed A. Ambusaidi, Xiangjian He, Zhiyuan Tan, Priyadarsi Nanda, Liang Fu Lu and Upasana T.Nagar, "A novel feature selection approach for intrusion detection data classification" in IEEE 13th ICTSPCC,2014.
[12]. Mohammed A. Ambusaidi, Xiangjian He, Priyadarsi Nanda, Zhiyuan Tan, "Building an Intrusion Detection System Using a Filter-Based Feature Selection Algorithm" in IEEE TRANSACTIONS ON COMPUTERS, VOL. 65, NO. 10, OCTOBER 2016.
[13]. A. M. Chandrashekhar, K. Raghuveer, "Intrusion detection technique by using k-mean, fuzzy neural network and SVM classifiers" in ICCCI-2013.
[14]. Kathleen Goeschel, "Redusing False Positive in Intrusion Detection Systems Using Data-Mining Techniques Utilizing Support Vector Machines, Decision Trees, And Naive Bayes For Off-Line Analysis" in IEEE 2016.
[15]. Shi-Jinn Horng, Ming-Yang Su , Yuan-Hsin Chen , Tzong-Wann Kao, Rong-Jian Chen, Jui-Lin Lai, Citra Dwi Perkasa ,"A novel intrusion detection system based on hierarchical clustering and support vector machines" in 38th Int. journal Expert Systems with Applications,2011.
[16]. 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", in Int. J.35th Expert Systems with Applications,2008.