Classification of Attacks Using Support Vector Machine (SVM) on KDDCUP'99 IDS Database

[1]*Manjiri V. Kotpalliwar*
Department of Computer Science and Engineering
Y.C.C.E.
Nagpur, India
manjirikotpalliwar@yahoo.com

[2]*Rakhi Wajgi*
Department of Computer Science and Engineering
Y.C.C.E.
Nagpur, India
wajgi.rakhi@gmail.com

Abstract*-Intrusion Detection System (IDS) is used to preserve the data integrity and confidentiality from attacks. In order to identify the type of attack in IDS, different methodologies like various data mining techniques exist. But some are very time consuming and laborious. Therefore we have proposed the usage of SVM (Support Vector Machine) for classification of attack from large amount of raw intrusion detection datasets on standard personal computers. SVM is a method which is used in data mining to extract predicted data. We have use KDDCUP'99 IDS database for classification.*

Keywords- SVM; Data Mining; KDDCUP'99 IDS database .

I. INTRODUCTION

Intrusion is a set of actions which compromise the integrity of data, confidentiality, or availability of any resource on a computing platform [18]. An intrusion detection system (IDS) is a combination of hardware and software for detecting the intrusions in the network. IDS help to monitoring all the activities in the network and hence can detect the signs of intrusions. The main objective of IDS is to alarm the system administrator that any suspicious activity happened in the network [4]. There are two types of Intrusion detection techniques:

• Anomaly Detection: Detecting malicious activities based on deviations from the normal behavior are considered as attacks. Although it can detect unknown intrusions, so that rate of missing report is low.

• Misuse Detection: Detecting intrusions based on a pattern for the malicious activity [4]. It can be very helpful for known attack patterns. It significantly identifies missing data which is considerably high.

Many data mining techniques exist like classification, clustering, association rule. The accuracy of classification technique depends on quality and quantity of data collected. It demands for huge amount of data. This disadvantage will be overcome in clustering. In this, though labeling of data is time consuming but it identifies complex type of IDS data.

Association rule technique is based on attribute/ value pair. It identifies multi-feature correlation between attributes of IDS and based on this correlation classification of attack is done. Support vector machine (SVMs) was first proposed by Vapnik [19] which is used for binary classification. SVM prove that SVM is fit for the highly variable and high dimensional data. Data acquired in IDS domain is more complex. So we introduce SVM to the research for intrusion detection system (IDS). SVM is a classifier which can be store the result in two ways: partial or true. Generally, SVM is a simplest linear form which is the hyperplane for separating the positive and negative examples.

Support Vector Machine (SVM) Algorithm is specially used for solving the classification problem. The fundamental concept of SVM is to find decision surface that separates the best data vectors into two classes for a given problem which is defined over the vector space [9]. SVM is a simplest linear form; generally it is a hyperplane that separates the positive examples from the negative example. Figure. 1 shows the hyperplane that separates the training data by a maximal margin between two classes [20]. All vectors lies on one side of the hyperplane labeled as +1 and other vectors lies on another side of the hyperplane labeled as -1. The training objects are lying closest to the hyperplane are called the support vectors [15].

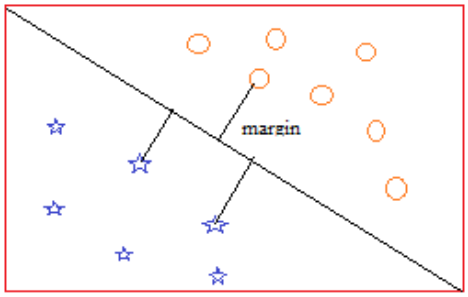


Figure 1. Linear Support Vector Machine

An advantage of SVM method is that the modeling of the hyperplane only deals with these support vectors than the whole training dataset, and therefore size of the training dataset is not usually an issue [9],[20]. A disadvantage is that the algorithm is time consuming for multiclass database and also it is sensitive to choice of the parameters, making it difficult to use.

This paper is divided into different sections as follows: Section II gives brief introduction about SVM. Section III deals with pseudo code of proposed algorithm and application of it on PC. This is followed by section IV exhibiting experimental results and the paper is concluded with section V containing conclusion and future scope.

978-1-4799-1797-6/15 $31.00 © 2015 IEEE
DOI 10.1109/CSNT.2015.185

II. SUPPORT VECTOR MACHINE AND ITS EXTENSION ON KDDCUP'99 IDS DATASETS FOR ATTACK CLASSIFICATION

Support vector machines (SVMs) conceptually implement the following idea: input vectors are mapped to a high dimensional feature space through some non-linear mapping chosen a priori. In this feature space a decision surface i.e. hyperplane is constructed. This decision surface is maximizes the "margin" between two classes in case of linearly separable data [15]. Special properties of this decision surface (hyperplane), ensures high generalization ability of the learning machine. This hyperplane can be written as: $\vec{w}.\ \vec{x} - b = 0$, where $\vec{x}$ denotes data point and the vector $\vec{w}$ and constant b are learned from the training data.

Let $y_i \in \{+1, -1\}$ is the classification label for input vector $\vec{x}$. Here, +1 is for positive class and -1 is used for negative class [15]. The linear SVM is used because they are fast to learn and classify new instances as compared to the non-linear SVMs. C-SVM is used for supervised learning which only deals with labeled corpus. One-Class SVM algorithm is deal with unlabeled data. It is an unsupervised learning method which is applied in intrusion detection estimation. The dataset in IDS is often high dimensional and heterogeneous. Traditional SVM can only deal with plain dataset and cannot tackle heterogeneous dataset directly, so we do some refinement of kernel function to extend SVM on the heterogeneous dataset. Here RBF (Radial Bias Function) kernel is used for classification of heterogeneous datasets.

III. PROPOSED METHODOLOGY

The experiments on KDDCUP99 IDS dataset for SVM model is a network connection record set which is restored from the raw data collected by Lincoln Labs at MIT for an intrusion detection system evaluation sponsored by DARPA in 1998. The data set contains total of 24 attack types which are classified into four categories [7]: Denial of service (Dos), Probe , User to Root (U2R) , Remote to User (R2L) . Each record is labeled either as normal or as attack or malicious, with exactly one specific attack type. We select two representative datasets "mixed" dataset and "10 percent KDDCUP'99" dataset. We employ the SVM type i.e. C_SVM algorithm to do classification on the dataset of DOS, Probe, U2R and R2L. We have classified the various types of attack to their belonging class. Classification is done using SVM classifier. It gives the result either in "Yes" or "No" results. During classification of attacks, any type of attack is classified properly to their belonging class then "Yes" result is displayed and accuracy of classification is increases as per the "Yes" result. If any type of attack is not classified to their belonging class means it appears in wrong class then the result shown "No". For e.g. Smurf is type of attack belonging to the class DoS, during classification, Smurf is classified in Dos class then the result is "Yes", and if it is classified in R2L class then the result shown "No". The broad overview of proposed work is shown by diagram:

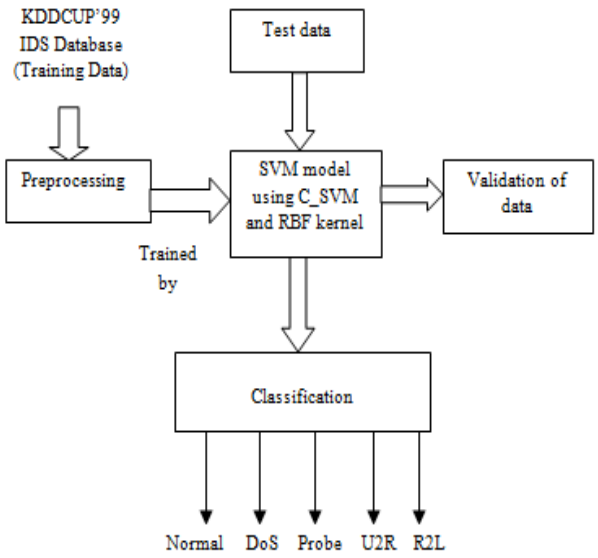


Figure 2. Broad Overview of Implemented work

The pseudo code of the proposed work is as follows:

Input:

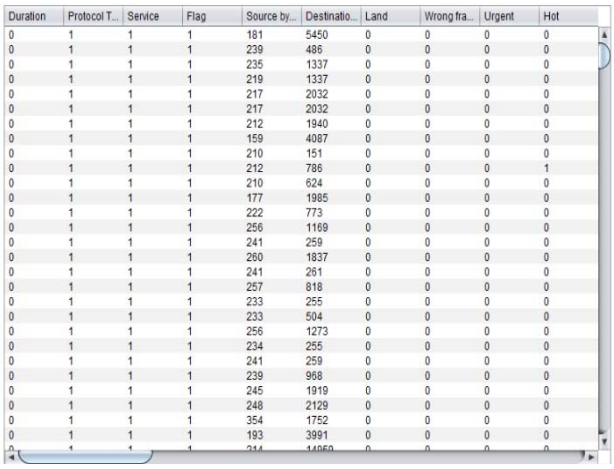
File containing records where each record belongs to specific attack. REC_i is a record of attack type i ($i \in 1 \ldots 24$).

1. Preprocessing Phase: Number each record REC_i with distinct index from 1 to I so that similar records are identified by unique index.
2. Training Phase: Train preprocessed records by creating SVM model using C_SVM and RBF kernel.
3. Validation Phase: Find number of records recognized correctly by SVM model.
4. Testing and Classification Phase: Test another 10% of unsupervised preprocessed data for classification.

Fig. 3 shows the sample attack database used for classification. Name of columns indicate features like duration, protocol type, service etc. [7]. After applying preprocessing we got the output which is shown in Fig. 4.

Durat...	Proto...	Service	Flag	Sour...	Desti...	Land	Wron...	Urgent	Hot	Faile...	Logg...	Num ...	Root ...	Su att...	I
0	tcp	http	SF	181	5450	0	0	0	0	0	1	0	0	0	0
0	tcp	http	SF	239	486	0	0	0	0	0	1	0	0	0	0
0	tcp	http	SF	235	1337	0	0	0	0	0	1	0	0	0	0
0	tcp	http	SF	219	1337	0	0	0	0	0	1	0	0	0	0
0	tcp	http	SF	217	2032	0	0	0	0	0	1	0	0	0	0
0	tcp	http	SF	217	2032	0	0	0	0	0	1	0	0	0	0
0	tcp	http	SF	212	1940	0	0	0	0	0	1	0	0	0	0
0	tcp	http	SF	159	4087	0	0	0	0	0	1	0	0	0	0
0	tcp	http	SF	210	151	0	0	0	0	0	1	0	0	0	0
0	tcp	http	SF	212	786	0	0	0	1	0	1	0	0	0	0
0	tcp	http	SF	210	624	0	0	0	0	0	1	0	0	0	0
0	tcp	http	SF	177	1985	0	0	0	0	0	1	0	0	0	0
0	tcp	http	SF	222	773	0	0	0	0	0	1	0	0	0	0
0	tcp	http	SF	256	1169	0	0	0	0	0	1	0	0	0	0
0	tcp	http	SF	241	259	0	0	0	0	0	1	0	0	0	0
0	tcp	http	SF	260	1837	0	0	0	0	0	1	0	0	0	0
0	tcp	http	SF	241	261	0	0	0	0	0	1	0	0	0	0
0	tcp	http	SF	257	818	0	0	0	0	0	1	0	0	0	0
0	tcp	http	SF	233	255	0	0	0	0	0	1	0	0	0	0
0	tcp	http	SF	233	504	0	0	0	0	0	1	0	0	0	0
0	tcp	http	SF	256	1273	0	0	0	0	0	1	0	0	0	0
0	tcp	http	SF	234	255	0	0	0	0	0	1	0	0	0	0
0	tcp	http	SF	241	259	0	0	0	0	0	1	0	0	0	0
0	tcp	http	SF	239	968	0	0	0	0	0	1	0	0	0	0
0	tcp	http	SF	245	1919	0	0	0	0	0	1	0	0	0	0
0	tcp	http	SF	248	2129	0	0	0	0	0	1	0	0	0	0
0	tcp	http	SF	354	1752	0	0	0	0	0	1	0	0	0	0
0	tcp	http	SF	193	3991	0	0	0	0	0	1	0	0	0	0

Figure 3. Sample Database of KDDCUP'99 IDS



Duration	Protocol T...	Service	Flag	Source by...	Destinatio...	Land	Wrong fra...	Urgent	Hot
0	1	1	1	181	5450	0	0	0	0
0	1	1	1	239	486	0	0	0	0
0	1	1	1	235	1337	0	0	0	0
0	1	1	1	219	1337	0	0	0	0
0	1	1	1	217	2032	0	0	0	0
0	1	1	1	217	2032	0	0	0	0
0	1	1	1	212	1940	0	0	0	0
0	1	1	1	159	4087	0	0	0	0
0	1	1	1	210	151	0	0	0	0
0	1	1	1	212	786	0	0	0	1
0	1	1	1	210	624	0	0	0	0
0	1	1	1	177	1985	0	0	0	0
0	1	1	1	222	773	0	0	0	0
0	1	1	1	256	1169	0	0	0	0
0	1	1	1	241	259	0	0	0	0
0	1	1	1	260	1837	0	0	0	0
0	1	1	1	241	261	0	0	0	0
0	1	1	1	257	818	0	0	0	0
0	1	1	1	233	255	0	0	0	0
0	1	1	1	233	504	0	0	0	0
0	1	1	1	256	1273	0	0	0	0
0	1	1	1	234	255	0	0	0	0
0	1	1	1	241	259	0	0	0	0
0	1	1	1	239	968	0	0	0	0
0	1	1	1	245	1919	0	0	0	0
0	1	1	1	248	2129	0	0	0	0
0	1	1	1	354	1752	0	0	0	0
0	1	1	1	193	3991	0	0	0	0

Figure 4. Preprocessed data

IV. EXPERIMENTS AND RESULT

We computed values of parameters related to performance evaluation of intrusion detector such as: Validation Accuracy = the number of samples recognize correctly and Classification Accuracy = the number of total samples classified accurately. Table 1 shows the different attack types with its total samples in 10 % KDD datasets and each attack type belongs to a particular class.

Table 2 shows the evaluation of the "mixed" dataset for validation accuracy and "10% KDD" dataset for classification accuracy.

TABLE I. CLASSES AND ITS ATTACK TYPES WITH ITS SAMPLES IN KDDCUP'99 IDS DATASETS

Classes	Attack types with its Samples
Normal	Normal. Samples: 97277
DOS	Smurf. Samples: 280790 Neptune. Samples: 107210 Back. Samples:2203 Teardrop. Samples: 979 Pod. Samples: 264 Land. Samples: 21
U2R	Buffer_Overflow. Samples: 30 Rootkit. Samples: 10 Loadmodule. Samples: 09 Perl. Samples: 03
R2L	Warezclient. Samples:1020 Guess_passwd. Samples:53 Warezmaster. Samples: 20 Imap. Samples: 12 ftp_write. Samples: 08 Multihop. Samples: 07 Phf. Samples: 04 Spy. Samples: 02
Probe	Satan. Samples: 1589 Ipsweep. Samples: 1247 Portsweep. Samples: 1040 Nmap. Samples: 231

TABLE II. VALIDATION AND CLASSIFICATION ACCURACY

Accuracy	For "mixed" and "10% KDD" datasets of attacks in %
Validation Accuracy	89.85
Classification Accuracy	99.9

V. CONCLUSION AND FUTURE WORK

In this paper we discussed the Support Vector Machines (SVMs), used for classification of attack from large amount of raw intrusion detection datasets on standard personal computers (PC's). SVM is a common method which is used in data mining to extract predicted data. IDS used to preserves the integrity of data, confidentiality and system availability from attacks. Intrusion Detection System (IDS) is the `burglar alarms' system of the computer security field [21]. KDDCUP'99 IDS database was firstly collected by Lincoln Labs at MIT for an intrusion detection system evaluation sponsored by DARPA in 1998.

We have implemented SVM Intel i3 2.4 GHz processor with 4 GB RAM. It took more than 1 minute to train as well as classify large size of data. So in future we are planning to implement SVM in Grid environment by adapting SVM which will use unique SVM model in new domain. This will helps to reduce the execution time of SVM.

REFERENCES

[1] Chih-Wei Hsu and Chih-Jen Lin, "A Comparison of Methods for Multiclass Support Vector Machines", IEEE Transaction on Neural Networks, 2002.

[2] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, "Boosting for Transfer Learning," in Proceedings of 24th International Conference of Machine Learning, pp. 193-200, 2007. Deepthy K Denatious and Anita John, "Survey on Data Mining Techniques to Enhance Intrusion Detection", International Conference on Computer Communication and Informatics, pp. 10 – 12, 2012.

[3] Deepthy K Denatious and Anita John, "Survey on Data Mining Techniques to Enhance Intrusion Detection", International Conference on Computer Communication and Informatics, pp. 10 – 12, 2012.

[4] Mohammadreza, Ektefa,Sara, Memar, Fatimah, Sidi, Lilly, Suriani Affendey, "Intrusion Detection Using Data Mining Techniques", IEEE, pp. 200 -203, 2010.

[5] L. Bottou, C. Cortes, J. Denker, H. Drucker, I. Guyon, L. Jackel, Y. LeCun, U. Muller, E. Sackinger, P. Simard, and V. Vapnik, "Comparison of classifier methods: A case study in handwriting digit recognition," in Proceedings of International Conference of Pattern Recognition, pp. 77–87, 1994.

[6] Thanh-Nghi Do and François Poulet, "Classifying one billion data with a new distributed SVM algorithm", in Proceedings of 4th International Conference on Computer Science, Research, Innovation and Vision for the Future, pp. 59-66, 2006.

[7] Shaik Akbar, Dr.K.Nageswara Rao, Dr.J.A.Chandulal, "Intrusion Detection System Methodologies Based on Data Analysis", International Journal of Computer Applications, Volume 5, No.2, 2010.

[8] Durgesh K. Srivastava, Lekha Bhambhu, "Data Classification Using Support Vector Machine", Journal of Theoretical and Applied Information Technology, 2005 – 2009.

[9] Ali Meligy and Manar Al-Khatib, "A Grid-Based Distributed SVM Data Mining Algorithm", European Journal of Scientific Research, 2009.

[10] Simon Tong and Daphne Kolle, "Support Vector Machine Active Learning with Applications to Text Classification", Journal of Machine Learning Research, 2001.

[11] Isabelle Moulinier, "Feature Selection: A Useful Preprocessing Step", in Proceedings of the 19th Annual BCS-IRSG Colloquium on IR Research, pp. 140-158, 1997.

[12] T. Joachims, "Text categorization with support vector machines: learning with many relevant features", in Proceedings of ECML-98, pp. 137-142, 1998.

[13] F. Poulet and T-N. Do, "Mining Very Large Datasets with Support Vector Machine Algorithms", in Enterprise Information Systems V, Camp O., Filipe J., Hammoudi S. et Piattini M. Eds., Kluwer Academic Publishers, pp. 177-184, 2004.

[14] J. C. Platt, N. Cristianini, and J. Shawe-Taylor, "Large margin DAG's for multiclass classification," in Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, vol. 12, pp. 547–553, 2000.

[15] Inderjit S. Dhillon, Subramanyam Mallela and Rahul Kumar, "Enhanced Word Clustering for Hierarchical Text Classification", University of Texas at Austin, 2002.

[16] J. Suykens and J. Vandewalle, "Least Squares Support Vector Machines Classifiers", Neural Processing Letters, 9(3): pp. 293-300, 1999.

[17] C. Cortes and V. Vapnik, "Support-vector network," Machine Learning, vol. 20, pp. 273–297, 1995.

[18] Khaled Labib, "Computer Security and Intrusion Detection" ,from Crossroads TheACM students magazine.

[19] V. Vapnik, "The Nature of Statistical Learning Theory", Springer-Verlag, New York, 1995.

[20] Pang-Ning Tan, M. Steinbach, and V. Kumar, "Data Mining Addison Wesley", London, 2006.

[21] http://citeseer.nj.nec.com/axelsson00intrusion.html, 2000.