Intrusion Detection System based on Support Vector Machine and BN-KDD Data Set

Razieh Baradaran, Department of information technology, university of Qom, Qom, Iran
R.baradaran@stu.qom.ac.ir

Mahdieh HajiMohammadHosseini, Department of information technology, university of Qom, Qom, Iran
M.hajihoseini@stu.qom.ac.ir

Name of the Presenter: Razeih Baradaran

Abstract

In today's world that information is a great wealth,protection of this wealth has a certain importance. On the other hand, with spread of global internet network and growing use of it, information security has become more vulnerable.

In this paper, we present a support vector machine based intrusion detection system that use BN-KDD data set for train and test. This data set is provided from science and technology university data mining lab that in fact it is an improvement form of KDD CUP 99. Mentioned system classifies each instancewhich is expressed with 41 features as normal or attack and has over 90 percent accuracy in experiments on training and testing data.

Key words: Intrusion detection system, data mining, support vector machine, BN-KDD

1. Introduction

Due to the increasing use of Internet and e-commerce transactions, the security of data Sent and received on the network has become more important. Attackers are always trying to find security holes and unauthorized access to data; But in the online environment, data access and disrupting network is easier. Unauthorized access includes incorrect, illegal, improper and unusual access to network information.

According toEndorf et al. (2004) Attacker interventions prevent servers to provide services for users. The intrusion detection system monitors incoming and outgoing data and prevents suspicious transactions. Intrusion detection system compares input traffic with known samples to detect attacks. In fact, the intrusion detection system try to detect computer attacks by examining various data records observed in network processes.

Network intrusion detection systems are of two types, the first type is based on the known samples of previous attacks and the incoming traffic are compared with the known attacks.The problem of these systems is the dynamic and changing attacks.It needs to update database of known samples constantly, otherwise it will not detect new attacks.The second type of intrusion detection systems is anomaly detection systems. They check abnormal traffic entering the network and prevent them from intruding.In this method, a set of normal traffic are collected and abnormal traffic can be discovered.

According to importance of maintaining the security of sent and received data on the network, there is a long time ongoing research activity for this purpose. A study conducted in the late 90's, has been worked in the Bonifácio (1998) paper, where attacks are discovered based on known attacks. Security agents place in a critical point in the network and Sent data are collected over several steps. Data preprocessing and the use of expert system for recognizing various behaviors and accepted samples are carried out by the agent. Then in semantic analysis, attack information, describing suspicious session information is stored in a database and the data is sent to the neural network to be assigned to a danger level.

In Sung and Mukkamala (2003) study, different experiment by Support Vector Machine and Neural Network using DARPA data was performed. Both showed good accuracy, and Support Vector Machine able to get more accurate results. In the same year, Bivens and colleagues (2002) used SOM algorithm for clustering and multi-layer neural network to detect suspicious connections. SOM clusters the input data based on the severity of traffic. Then neural network is used for identifying suspicious communication. In the Moradi and Zulkernine(2004) study, multi-layer neural network was used to detect intrusion. But it didn't work in the mode of two categories: Normal or abnormal. In the output of the neural network, three categories were created.

InBashah et al.paper (2005) artificial intelligence was used to detect intrusion. Its main techniques were neural network and fuzzy logic. Fuzzy rules can make the rules if-then that indicate attacks against the network security.

In Vokorokoset al. (2006) Article, SOM neural network have been used to detect intrusion. Requests and system status in processing them were monitored, and then converted to vector; Clustered by SOM algorithm and a value was assigned to each vector. Exceeding this value from a certain threshold, represent abnormal behavior.

Linda and colleagues (2009) in a paper again used the neural network to detect intrusion. In the process, the Sent packets to the network are analyzed, and the headers are extracted. Then a window move on packets; then from different packets, the required statistics on certain parameters are extracted, such as the number of IP address and minimum and maximum number of packets associated with each IP. Then neural network has been used to define the boundary for normal behavior.

In the Ahmadet al. paper (2009) multi-layer neural network is used to detect DOS attacks. In the paper Mhammadand Mehrotra(2010)the combination of fuzzy logic and neural network is used to detect intrusion.

Horng and colleagues (2011) did an activity in order to discover attacks in networks using hierarchical clustering and support vector machine. Using hierarchical clustering have been achieved better quality data from KDD99 data. Then using support vector machine classifies higher quality data. This system resulted better accuracy in identification of DOS and Probe attacks and the overall accuracy was 95/72 percent.

In Visumathiand et al. (2012) study, the combination of genetic algorithms and fuzzy c-mean clustering method was used to detect abnormalities. In this method, the fuzzy clustering assign a membership level to each of its member and genetic algorithm done clustering, then cluster members again are classified to determine normal users and attackers.

In this paper, a Support Vector Machine classifier has been trained by using BN-KDD data set[1]to detect attacks on the network.

The paper is structured as follows: Section 2 introduces the BN-KDD dataset.In Section3, the pre-processing is performed on the data and presented system is introduced.Next section, namely section 4, shows the results of the evaluation system developed on the test data set.In the final section describe conclusions and future directions.

2. Data and Material

In this paper, we use BN-KDD data setto train and test our systems. This data set is subset of KDD CUP 99 that fixesdisadvantages of KDD CUP 99and NSL-KDDdatasets. In BN-KDD data set although the number of train and test instance is decreased, it can more accurately evaluates intrusion detection systems and compares them.

BN-KDD data set is a set of connection records which are located in folders as hierarchical order. Main set is consists of train and test subsets, which are five subset as normal dataset and 4 attack data sets in each of them.

Attack records fall into four categories:

- Denial of Service (DoS): is an attempt to make a machine or network too busy to accept legitimate users access these resources.
- Probe (PRB): host and port scans to gather information or find known vulnerabilities.
- Remote to Local (R2L): unauthorized access from a remote machine in order to exploit machine's vulnerabilities.
- User to Root (U2R): unauthorized access to local super user (root) privileges using system's susceptibility.

Connection records are organized in files based on two main features as protocol-type and service in each of subfolders. For example all normal records from train dataset which have protocol-type feature as icmp and service feature as eco_i, organized in train->normal->icmp.eco_i file path.

[1] http://dml.iust.ac.ir/fa/index.php/2011-09-23-13-27-22/2----bn-kdd.

The number of records as normal and attacks subsets in train and test datasets is presented in table.1.

Connection type	Record number of train dataset	Record number of test dataset
Normal	315462	35482
DoS	186	989
PRB	1616	174
R2L	355	222
U2R	50	355

Table 1: records number in train and test BN-KDD dataset

Each data instance is represented as 41 features. Two main features as protocol-type and service are recognized from file name; and 39 other featureis described in the file.

These features have all forms of continuous, discrete, and symbolic variables and falling in four categories:

- The first category consists of the intrinsic properties of a connection, which include the basic features of individual TCP connections such as the duration of the connection, the type of the protocol (TCP, UDP, etc.), and network service (http, telnet, etc.).
- The content features within a connection such as the number of failed login attempts.
- The same host features that calculate the statistics related to the protocol behavior, service, etc.
- The same service features.

3. Research Methodology

3.1 Data Preprocessing

Data preprocessing is the most important and time consuming parts of each data mining projects. In this step, data must be converted to the acceptable format of desired data mining algorithm. In support vector machine (SVM) classification, all feature values must be presented as numerical values.

As describe in previous section, data are organized in files based on protocol-type and service features. For train SVM using this dataset, first, all data must be integrated in one train data set. So assign numeric values to each main feature (protocol-type and service) and add these feature to feature set. For instance, protocol-type with tcp value and service with ftp value are mapped to 1 and 16 values respectively.

Also, for other non-numerical feature such as flag, its values are mapped to appropriate numerical values.Finally, Status feature are added to feature set, which represent instance

status as normal or attack instance. This feature, in fact is prediction feature for intrusion detection system. In other word, our system must predict the status value (attack or normal) based on other features values.

3.1 Intrusion DetectionSystemsBased on Support Vector Machine

SVM is type of machine learning methods that shows good performance for classification rather than many other approaches.Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships.The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes. In other word, each data instance are mapped into one side of decision plane and predicted to belong to a category based on which side of the plane they fall on [15].

We develop an intrusion detection system using SVM classifier. This classifier, train with 40 features of normal and attack records from train dataset. In train dataset normal dataset size is higher than attack data set size, While in SVM classification positive and negative records must be balanced to create a reasonable model and correct prediction. Otherwise prediction is inclined to class with more instances. For solve this problem, we sample normal records to equivalent with attack total size.

After model creation, we evaluate this model with test dataset and receive records status values as normal or attack in output. We repeated train and test process for 5 times and shuffle normal data set before sampling in each time to reduce negative impact of sampling data on model creation. So at each run time, different normal records involved in train process. Also we scale features values in 1 and -1 values so all features have same contribution in model creation and prediction.

4. Results and Analysis

In this study we evaluate our experiment using performance measures metrics precision, recall and f-measure. We use Rapid Miner Software and LIBSVM classifier (which is kind of SVMclassifier) to implement these systems. Averagemeasures values from first system evaluation in test data set and for 5 time is presented in table2. Also the results of evaluating the system on each normal and attack data sets can be seen in Table 3 and Fig 3.

	Attack Class (%)	**Normal Class (%)**
Precision	86.25	99.32
Recall	86.23	99.32
F1	86.24	99.31
Accuracy	98.71	98.71

Table 2: Average result performance of intrusion detection system based on SVM in run 5 times.

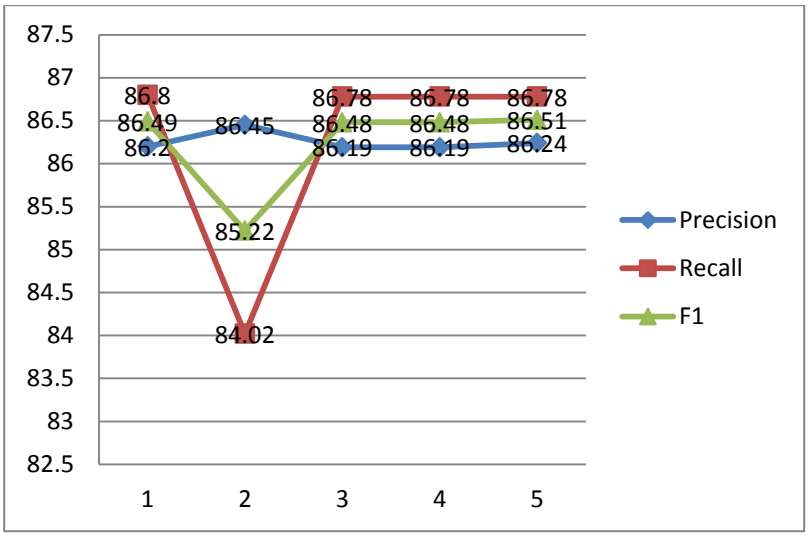

Fig 1. Precision and Recall of attack class in run 5 time

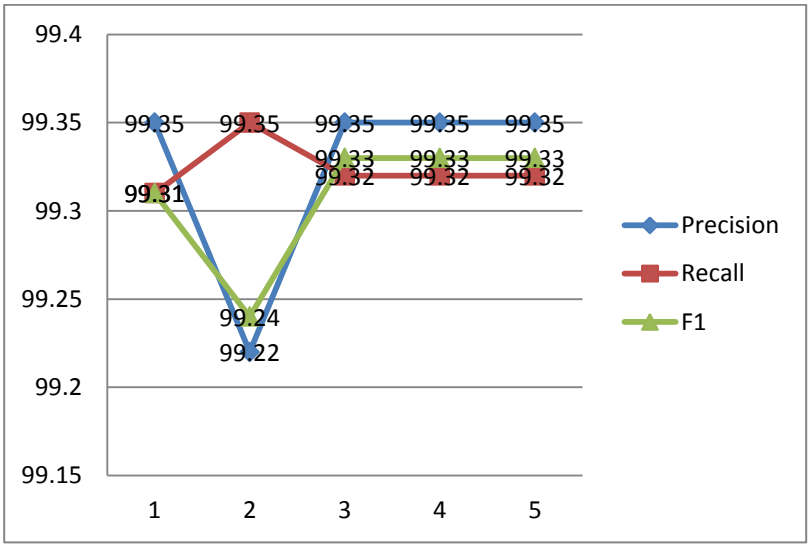

Fig 2. Precision and Recall of normal class in run 5 time

	Accuracy
Normal	99.32
DoS	87.26
PRB	85.06
R2L	65.77
U2R	99.72

Table 3: System accuracy attacks and normal datasets

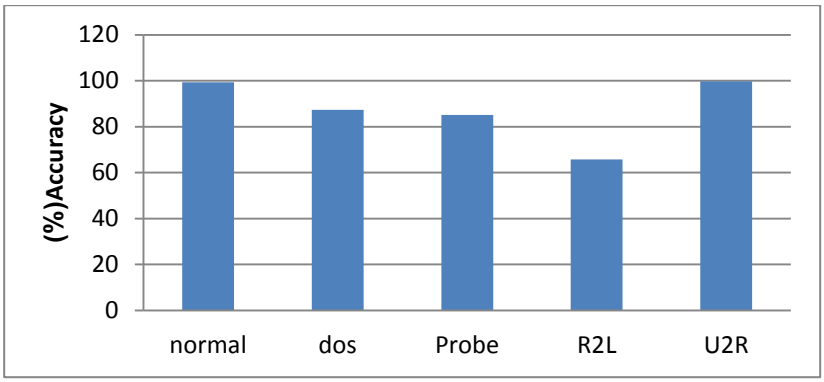

Fig 3. Comparison of system accuracy in normal and attacks datasets

5. Conclusions and recommendation for future works

According to spread communication and connection of different information networks, can be access to many information easily. On other hand, the protection of sensitive information against unauthorized access has become critical. Hence information systems need the intrusion detection system to detecting network attacks.

In this study, we presented an intrusion detection system based on support vector machine; that use BN-KDD data set for training for the first time. In this system, the performance results show that this system has precision 80 percent and recall 90 percent in attack detection, which is good result for an intrusion detection system.

References

Depren, O.,Topallar, M.,Anarim, E. &Ciliz, M. K. (2005).*An Intelligent Intrusion Detection System (IDS) for Anomaly and Misuse Detection in Computer Networks*, Expert Systems with Applications, vol29, 713-722.

BonifácioJr, J. M., Moreira, E. S., Cansian, A. M. &Carvalho A. C. P. L. F. (1998).*An Adaptive Intrusion Detection System Using Neural Networks*, Proceedings of the 14th Int. Information Security Conference, Vienna/Budapest, Austria/Hungary,

Sung, A.H.&Mukkamala, S. (2003).*Identifying Important Features for Intrusion Detection Using Support Vector Machines and Neural Networks*, International Symposium on Applications and the Internet Technology.

Bivens, A.,Palagiri, C., Smith, R. & Szymanski, B. (2002).*Network-Based Intrusion Detection Using Neural Networks, in Proceedings of the Intelligent Engineering Systems Through Artificial Neural Networks",* St.Louis,Vol12, 579-584, New York.

Moradi, M. &Zulkernine, M. (2004).*A Neural Network Based System forIntrusion Detection and Classification of Attacks*, IEEE InternationalConference on Advances in Intelligent Systems - Theory and Applications,Luxembourg-Kirchberg, Luxembourg, November 15-18,2004.

Endorf, C.F, Schultz, E.,&Mellander, J. (2004).*Intrusion detection and prevention*publisher: Brandon A. Nordin, California: McGraw-Hill,

Bashah, N,BharanidharanShanmugam, I. & Ahmed, A. (2005).*Hybrid Intelligent Intrusion Detection System*, Transactions on Engineering, Computing and Technology, vol. 6, 291-294.

Vokorokos, L., Baláž, A. &Chovanec, M. (2006) *Intrusion Detection System Using Self Organizing Map*, ActaElectrotechnica et Informatica, No. 1, Vol. 6.

Linda, O., Vollmer, T. &Manic, M. (2009).*Neural Network Based Intrusion Detection System for Critical Infrastructures*, IJCNN 2009, Int. Joint INNS-IEEE Conf. on Neural Networks, Atlanta, Georgia, June 14-19.

Ahmad, I., Abdullah, A.B. &Alghamdi, A. (2009).*Application of Artificial Neural Network in Detection of DOS Attacks*, Proceedings of the 2nd international conference on Security of information and networks, 229-234.

Mhammad T.J.M. &Mehrotra, M. (2010).*Design Network Intrusion Detection System using hybrid Fuzzy-Neural Network*", International Journal of Computer Science and Security, Volume (4) , Issue (3), 285-294.

Horng, and et al. (2011).*A novel intrusion detection system based on hierarchical clustering and support vector machines***,** Expert Systems with Applications: An International Journal, Volume 38 Issue 1, 306-313, January.

Visumathi, J.,Shanmuganathan, K.L.&MuhamedJunaid, K.A. (2012).*Misuse and Anomaly-based Network Intrusion Detection System using Fuzzy and Genetic Classification Algorithms*, InternationalConference on Computing and Control Engineering (ICCCE 2012).