Analysis of KDDCUP99 and NSL-KDD using Various Classification Algorithms

RITU BALA and Dr. RITU NAGPAL
Department of Computer Science,
Guru Jambheshwar University of Science & Technology,
HISAR.

Abstract:

Intrusion detection system is a very important part of our defense system which is used to identify the abnormal activities. As the number of attacks occurring in our computer network is increasing day by day and this is why It has become very important to detect the intrusion, intercept and analyze a large number of network data. Normally, traditional IDS relies on the extensive information of our security experts. To reduce this dependency, different machine learning and data mining techniques are used. In this paper we will compare the performance of different learning algorithm like decision tree, naïve bayes, multilayer perception using different datasets such as KDDCup99, NSL-KDD and by different scenarios identify between normal and abnormal connections on different data sets

Keywords: intrusion detection system, data mining, KDDCup, NSL-KDD, naïve bayes, machine learning

Introduction

For the safety and security of the network and computer system IDS plays very important role. It collects the information from certain region of the network and computer, examining it and finding out what breaches its security [1]. The faster the use of computer network is increasing, the faster its security problem is also increasing. Security means that providing a kind of protection to the network system and authentication, confidential and integrity is the most important goal of the security [2]. Any type of attack on the network is called intrusion. Intrusion means any illegal program that steals very important information. IDS help the system to prevent infiltration from outside. IDS collects the data running on the network, examine it and then separate it into normal and abnormal data and the result is produced to system administrator [3].

IDS monitors both the internal and external activities of the network and if it seems any intrusion trying to steal any information or harm the system, it generates an alarm which alert the system administrator. It is basically prepared so that it can protect the supreme information of any organization from intrusion and attacks.

IDS collects information from different sources and compares it with attack signature's data base to test it. IDS is a device that helps to find illegally penetrated intruders wherever they are. It makes an alarm when an intrusion occurs, it detects the attacks that have taken place.

There are some terms which are used to identify normal and abnormal behavior of traffic
True Positive: gives the correct result means if there is an attack, identify it correctly.
True Negative: gives the correct result means if the data is normal, identify it correctly
False Positive: gives the wrong result means it detect a normal data as attack.
False Negative: detect attacked data as normal.

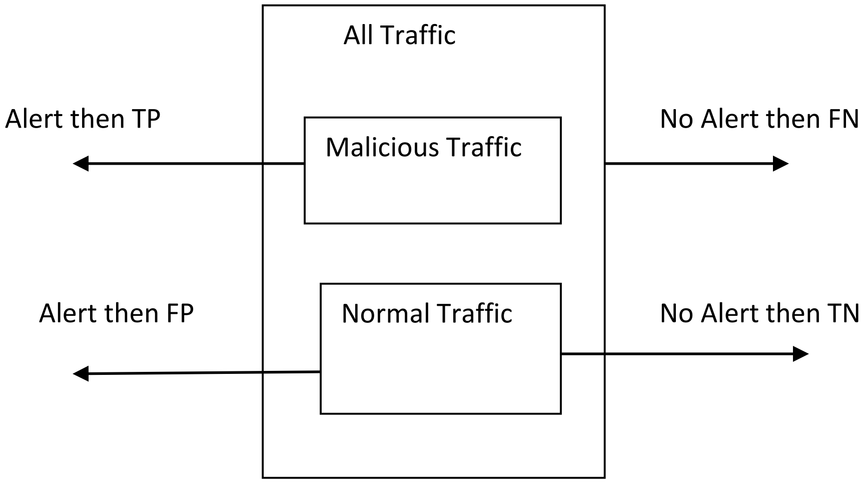

Figure1. Anomaly Detection Process

Features of intrusion detection system are:

1. It investigates and keep an eye on the system and user's aspect
2. It examines the system configuration and vulnerability.
3. It examines file and system integrity
4. It has the ability to recognize pattern of attack
5. It detects the abnormal activities
6. It also keeps an eye on the violation of user's policy [4]

IDS are deliberately launched on the network to go over with the packets and identify threats. IDS do this by collecting the data from different sources, examine it and if it finds any abnormal activity the take corrective measures by generating alarm and reporting to system administrator [5].

The attack is a threat to computer security. Sensitive and confidential data transfer and information exchange is part of network traffic that gives an open path to attacks. However, we know very well that the dependency on the network is increasing rapidly due to which the network problem has also become very serious and will become more complex in the future. This traffic can cause a huge loss of network system and related resources. Anomaly detection is a method that checks traffic on the network based on traffic patterns and detects malicious and unwanted attacks from it.

There are many datasets which are open source and evaluate the performance of anomaly detection system. The first dataset for the evaluation of IDS was DARPA1999. By processing the tcpdump data of DARPA, KDD99 was created. KDDCup99 has some limitations and to cover up these limitations NSL-KDD was designed.

In this paper different datasets are classified on different classification techniques and the results are compared. Tool which supposed to use is WEKA.

DATASETS

KDDCUP99: The main purpose to design the dataset was to build a predicative model which can differentiate between normal traffic and attacked traffic. This dataset consists variety of intrusion in their database. KDD is built by dealing the tcpdump files of DARPA 98. this is the most accepted dataset for NIDS. Each record contains 41 features and 42nd feature is class. It also has the same problem as it had in DARPA. It has more than 20 attack types which are divided into four parts i.e., Dos, Probe, U2R and R2L. Normal and both records are put together in a simulated environment and the result was a lot of redundant records which affect the accuracy of the result [8,9]. This dataset is very large consist about 494021 instances and to use a full dataset is not practically possible. So, the researchers have carefully drawn the subset of total set. The compact size subset called KDD10%.

NSL-KDD: To come up from the drawbacks of KDDCup, NSL-KDD dataset was designed [8,9]. It is the processed version of KDD99 whose aim to reduce the racism of the classifier. In NSL-KDD dataset, the redundant and duplicate records have been removed which results in better detection and high accuracy. As this dataset consist reasonable number of instances, so there is no need to select the subset. The whole dataset can be run during experiments [10]

Supervised Learning Techniques

In this section different supervised learning techniques are discussed. It uses labeled training data to detect intrusion. Training and testing are the two stages of this approach. In the training phase, the algorithm learns from these data samples by identifying relevant features and classes.

In this technique every record is labeled as either normal or intrusion. Each record has some features like duration, source and destination address, port etc. out of these many features are redundant and to eliminate these unnecessary features, feature selection technique can be applied. After that, classifier is trained to learn the intrinsic association between input and labeled output. In testing phase, this trained model classifies the data into normal or abnormal

There are various supervised learning techniques and each of having some pros and cons. Decision tree, SVM, naïve bayes, k-nearest neighbor, neural network etc. build their classification model by using learning method. The fundamental task of learning algorithm is to build the classification model.

Decision tree: Decision tree has a treelike structure which consist decision node, branch and leaf. Decision node represent features or attributes of dataset, branch represent decision taken based on attributes or features and the leaf represent the class to which a particular instance related. ID3, J48, CART etc. are some decision tree algorithms.

Naïve Bayes: It is frequently used classification algorithm based on Bay's principle. It gives the probability of occurrence of any type of attack by observing the system activities using conditional probability. It is most popular because of its calculation efficiency and easy to use which comes from the conditional independent assumption property [6]. It does not work well for large datasets. Hidden Naïve Bayes can be used for large datasets that have highly interrelated attributes, large dimensions and fast network [7].

Evaluation Metrics

Some metrics have been used to evaluate the performance of IDS such that accuracy, false positive (FP), f-measure and precision. For an acceptable IDS, accuracy and detection rate should be high and false positive (FP) should be very low.

Accuracy = TP +TN/ TP+TN+FP+FN
Precision = TP/TP+FP
F-measure = 2TP/2TP+FP+FN

Experimental Result

In this paper, we have tried to test both KDDCup and NSL-KDD data sets by applying different classification algorithms as shown in figure1 and figure2. We have used random forest, j48, naïve bayes and multilayer perception. After applying these algorithms, we have evaluated it above some metric as accuracy, TP, FP, F-measure and Precision and in the result, we have found that random forest give j48 gives more accuracy on NSL-KDD and random forest gives more accuracy on KDDCup dataset. KDDCup has a huge dataset so we used only 10% of the data of KDD dataset. For experiment WEKA tool is used

NSL-KDD

Algorithm	Accuracy	TP	FP	F-measure	Precision
Random Forest	62.9873	0.875	0.425	0.462	0.314
J48	63.9747	0.873	0.412	0.4680	0.320
Naïve Bayes	55.0772	0.678	0.469	0.358	0.243
Multilayer Perception	53.4008	0.663	0.495	0.341	0.229

Figure1

KDDCup99

Algorithm	Accuracy	TP	FP	F-measure	Precision
Random Forest	99.9625	0.999	0.003	1.000	0.995
J48	99.9375	0.996	0.004	1.000	0.996
Naïve Bayes	95.5	0.996	0.000	0.998	0.999
Multilayer Perception	99.8625	1.000	0.000	1.000	1.000

Figure2

Conclusion

This paper presents the analysis of two dataset KDDCup and NSL-KDD using random forest, naïve bayes, j48 and multilayer perception algorithms. The main purpose of this paper to give an idea which classification algorithm is best suited on which benchmark dataset. WEKA tool is used for the experiment. On the basis of result we conclude that j48 gives beset can better result on NSL-KDD and random forest is best suited for KDD. But KDDCup has some redundant records so sometimes its biased result. In future more dataset can be evaluated differently by applying different classification algorithms.

References

1. Ravi Jain and Ajith Abraham “Soft Computing Models for Network Intrusion Detection Systems” School of Information Science, University of South Australia, Australia ravi.jain@unisa.edu.au and Department of Computer Science, Oklahoma State University, USA ajith.abraham@ieee.org
2. Anita John and Deepthy K Denatious “Survey on Data Mining” *Conference on Computer Communication and Informatics (ICCCI-2012)*, 2012, Coimbatore
3. Kai-Fan Cheng, Rung-Ching Chen and Chia-Fen Hsieh, “Using Rough Set and Support Vector Machine for Network Intrusion Detection”, *International Journal of Network Security & Its Applications (IJNSA)*, vol 1, issue 1, April 2009
4. Mr. Suresh Kashyap, Ms. Pooja Agrawal, Mr. Vikas Chandra Pandey and Mr. Suraj Prasad Keshri “Soft Computing Based Classification Technique Using KDD99 Data Set for Intrusion Detection System” *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, vol. 2, issue 4, April 2013.
5. D. Rozenblum “Understanding Intrusion Detection Systems,” *SANS Technology Institute*, vol.29, issue5, pp. 11–15, 2001.
6. X. Yang and Y. L. Tian, "EigenJoints-based action recognition using Nave-Bayes-nearest-neighbor," *in 2012 IEEE computer society conference on computer vision and pattern recognition workshops*, pp. 14–19, 2012

7. L. Koc, T. A. Mazzuchi, and S. Sarkani, "A Network Intrusion Detection System Based on A Hidden Naïve Bayes Multiclass Classifier," *Expert System with Applications*, vol. 39, issue 18, pp. 13492–13500, 2012.
8. John McHugh “Testing Intrusion detection systems: a critique of the 1998 and 1999 DARPA Intrusion Detection System Evaluations as Performed by Lincoln Laboratory” *ACM Transaction on Information and System Security*, vol. 3, issue 4, pp. 262–294, 2000.
9. Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani “A detailed analysis of the KDD CUP 99 data set,” *Proceedings of the IEEE Symposium on Computational Intelligence for Security and Defense ApplicationsICISDA)*, Piscataway, NJ, pp.1–6, 2009.
10. Nachiket Sainis, Durgesh Srivastava and Rajeshwar Singh “Classification of various Dataset for Intrusion Detection System” *International Journal of Emerging Technology and Advanced Engineering*, volume 8, Issue 1, pp. 40-50, January 2018.
11. Ring, M., Wunderlich, S., Grudl, D., Landes, D., Hotho, A.” Flow-based benchmark data sets for intrusion detection” *Proceedings of the 16th European Conference on Cyber Warfare and Security (ECCWS*), ACPI, 2017.