Intrusion Detection Based on Decision Tree Using Key Attributes of Network Traffic

Ritu Bala and Ritu Nagpal

Abstract As computer usage is increasing, network security is becoming a huge problem. As time passes, attacks are also increasing on the network, these attacks are nothing else, but it is like the intrusions that are causing **a lot of** much damage to our entire system. IDS is used to protect our data and network from these attacks and to secure our systems from intruders. The technology of data mining is used a lot in order to examine and analyze enormous network data. Data mining is an efficient method that is applied to IDS to detect a large amount of network data, and reduces the pressure of compilation done by humans. This paper compares the different techniques of data mining used to implement IDS. Information gain and rankers' algorithm are used for attribute selection and J48 and random forest classify the data for NIDS and the dataset used is KDDCup99. We have selected 9 attributes from the KDDCup dataset and the experiment is done on the WEKA tool. Moreover, the results show that the detection accuracy with only 9 attributes is almost the same as it was with all 41 attributes.

Keywords NIDS · WEKA · KDDCup99 · J48 · Random forest · Information gain · Rankers algorithm

1 Introduction

Due to the excessive use of computer networks, data's safety and security problem is increasing day by day. Attacks are sometimes called intrusion to cause much damage to the system and our data. IDS helps us in this work and keeps our system free from attacks. Earlier, the methods used were encryption, firewalls, virtual private networks, etc. Now-a-days these cannot be trusted entirely. These techniques do not suffice to protect our data, that is why today there is a need for such a technology that can investigate and analyze the wrong and illegal activities occurring in our systems and networks.

R. Bala (✉) · R. Nagpal
Guru Jambheshwar University of Science and Technology, Hisar, India

A. Choudhary et al. (eds.), *Applications of Artificial Intelligence and Machine Learning*,
Lecture Notes in Electrical Engineering 778,
https://doi.org/10.1007/978-981-16-3067-5_43

Hence a dynamic approach has been designed for the security of our network which is called an intrusion detection system (IDS). It is of two types: Signature-based IDS and anomaly-based IDS [1]. It is further categorized as host-based and network based depending on the situation of the environment. Host-based invigilate the conduct of the host system and network-based invigilate the behavior of the network.

Lee et al. [1] in 1999 constructed the classification model with 41 attributes from the unprocessed traffic that was collected at MIT Lincoln Laboratory [2] called KDD Cup 1999 (The 1999 Knowledge Discovery and Data Mining Tool).

2 Data Mining in IDS

Data mining is a broad concept that is most commonly used in computer science. It is the process of extracting out new valid, meaningful and significant information from the large database. In the last few years, data mining techniques like classification, clustering and association rules have successfully found intruders. Business organizations and commercial accountants mainly use it, but now it is increasingly used in research to extract valuable information during experiments and observations [3]. Because of the excessive use of computer networks, there is a large volume of existing and newly arrived data on the network that need to be processed. That is why data mining-based IDS has gained attention in research. Data mining-based idea is used to examine the covered pattern of intrusion and the relationship hidden in the data [4]. It is used for the detection of variants, control false alarm and improve efficiency [5].

3 WEKA Tool

This tool is used to perform data mining and machine learning tasks. It is a group of many machine learning and data mining algorithms, especially classification, data preprocessing, regression, feature selection and visualization. Its programming is done in the JAVA language. It is used extensively in research. It has 49 data preprocessing tools, in addition to this has 76 classification algorithms, 15 attribute evaluators and 10 search algorithms for the selection of features. All the files must be in ARFF format to be run in it.

4 KDDCup 1999

KDDCup99 is the oldest and the most commonly used dataset which is used to access anomaly detection. It identifies good connections called normal and bad connections

called intruders. Nine weeks of data is collected for training and test data. It contains 490,000 instances and every instance has 41 features labeled as either attack or normal [6]. These features are categorized into three groups as basic, traffic and content. There are 24 training attack types which are grouped as

- Denial of Service: In this attack, the attacker send so much request on the server and make the memory and other resources too busy that the request of genuine user to access the machine is denied, e.g., smurf attack, land attack, etc.
- R2L: In this type of attack, the attacker finds a way to get access to the machine through negotiating the security-like guessing the password, etc.
- U2R: In this attack, the attacker has the benefit of local access to the system and makes an effort to access to the administration, e.g., buffer overflow attack.
- Probe: In this attack, the attacker benefits by collecting the information about the victim machine, he gains this information by taking advantage of their weaknesses, e.g., Port scanning.

5 Related Work

The most important project in network security is to make an effective NIDS. Experts have done a lot of work in this field. The work is divided into 3 parts. First, find the important and relevant features from the attribute set, then upgrade learning algorithms and then assess the performance on any dataset. Li et al. [7] gave an ideal IDS in which they sorted the required features by applying the information gain method and chi-square method. The outcome of this work indicates that the accuracy of the detection is still maintained even by using some selected attributes. In [8], the author used PCA to select features and the features which had higher Eigenvalues were retained. In [9], the author explains that by using adequate training parameters and selecting the right features, high accuracy can be achieved. The performance of an IDS can also be improved. In [10] writers used function rating set of rules to lessen the function area through the usage of three rating set of rules primarily based totally on support vector machine (SVM), multivariate adaptive regression splines (MARS) and linear genetic programs (LGP). In [11], creators propose "Enhanced Support Vector Decision Function" for function selection that is primarily based totally on essential factors. First, the function's rank, and second, the correlation among the features were adopted. In [12], writers advise an automated function choice process primarily based totally on correlation—primarily based feature selection (CFS).

6 Algorithms Applied

6.1 J48

J48 is an open-source algorithm in WEKA used to build the decision tree and made known by Ross Quinlan. This algorithm works on supervised learning. The decision trees produced are used for classification. During decision tree construction, first of all find out the instances belonging to the same class. If found then, the tree is represented by a leaf and labeled by same class. Second, the information gain calculates every attribute and then results in the best attributes for branching the tree [13–15].

Entropy is the term used to calculate information gain [16]. The formula of entropy (E) is

$$E = -\sum_{i=1}^{n} P_i \log_2 P_i \tag{1}$$

Gain (S, X) of attribute X w.r.t. total sample (S) is

$$\text{Gain}(S, x) = E(S) - \sum_{j \in \text{values}(X)} \frac{\text{mod}(Sj)}{\text{mod}(S)} E(S_j) \tag{2}$$

Information Gain can be calculated as

$$\text{Splitinfo}(S, X) = -\sum_{k=1}^{c} \frac{\text{mod}(Sk)}{\text{mod}(S)} \log_2 \frac{\text{mod}(Sk)}{\text{mod}(S)} \tag{3}$$

$$\text{Gain Ratio} = \frac{\text{Gain}(S, X)}{\text{Splitinfo}(S, X)} \tag{4}$$

6.2 Random Forest

Supervised learning algorithm that is used in both classification and regression, but most commonly used in the classification. This algorithm builds the decision tree using each data sample and takes the prediction of each decision tree. Then, voting is performed on every predicted result and selects that prediction that gets maximum votes and gives the best results. This algorithm also reduces the problem of over fitting to some extent. This algorithm works efficiently with large amount of data. Its accuracy is very high even after missing a large proportion of data.

Table 1 Comparison of J48 and random forest

Algorithm used	Number of attributes	TP	FP	Precision	Time (s)
J48	42	0.996	0.004	0.996	2.54
	12	0.997	0.005	0.997	1.26
	10	0.997	0.006	0.995	0.33
	09	0.997	0.005	0.995	0.33
Random forest	42	0.999	0.003	0.999	9.31
	12	0.999	0.004	0.997	4.46
	10	0.998	0.004	0.997	3.4
	09	0.998	0.004	0.997	3.43

7 Experimental Result

7.1 Parameter

True positive (TP) gives the correct result means if there is an attack, identify it correctly.

False positive (FP): means there is actually an attack, but it predicts it as normal.

Precision: tell the correct percentage of positive prediction.

Time: Gives how much time it takes to build the model.

Accuracy = TP + TN/TP + TN + FP + FN * 100.

Precision = TP/TP + FP * 100.

7.2 Results

Table 1 shows the comparison of J48 and random forest with varying numbers of attributes. As shown in the table, if we reduce the attributes, there seems no significant difference between TP, FP, and precision results, but there is a big difference in the time it takes to build a model. If we get almost identical or we can say better results with a smaller number of attributes then it makes no sense that we have to use complete 42 attributes (Figs. 1, 2, 3 and 4).

8 Conclusion

In this paper, we have used information gain for the extraction of attributes for intrusion detection. To classify these extracted attributes, we have compared J48 and Random forest decision tree algorithms. The accuracy of the parameters has shown some progress by taking only 08 attributes compared to using 41 attributes.

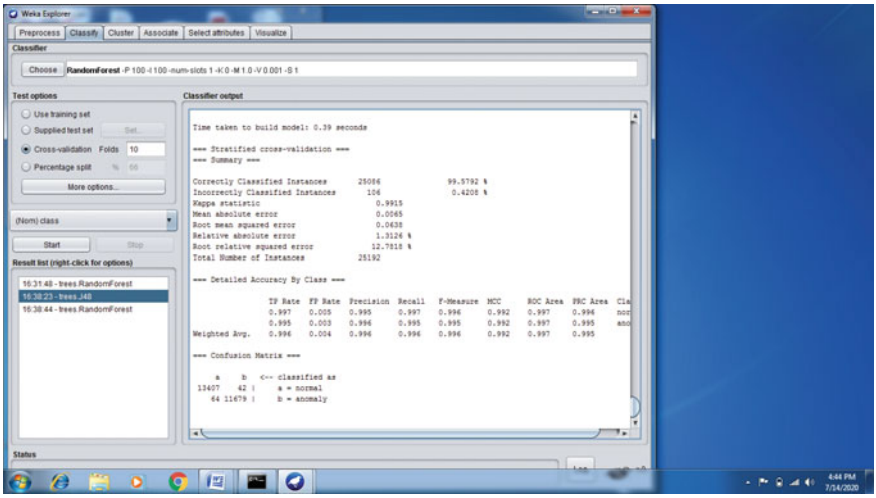

Fig. 1 J48 algorithm with 9 attributes

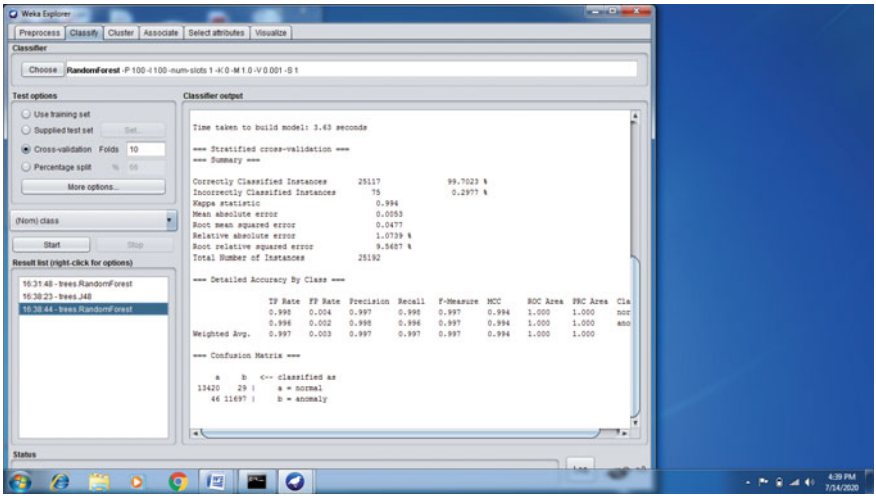

Fig. 2 Random forest algorithm with 9 attributes

The significant difference is in the time to build the model which shows remarkable improvement with fewer attributes. The experiment is performed on the WEKA tool using KDDcup99 Dataset. According to the results, J48 with 09 attributes gives better results. Still, some more work is required in this field. We propose to use other classification algorithms apart from J48 and Random Forest on varied datasets in the near future.

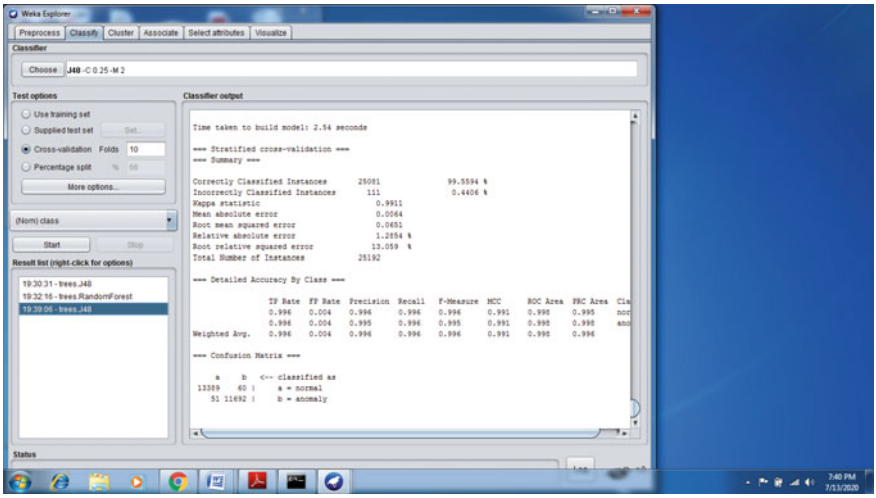

Fig. 3 J48 algorithm with 42 attributes

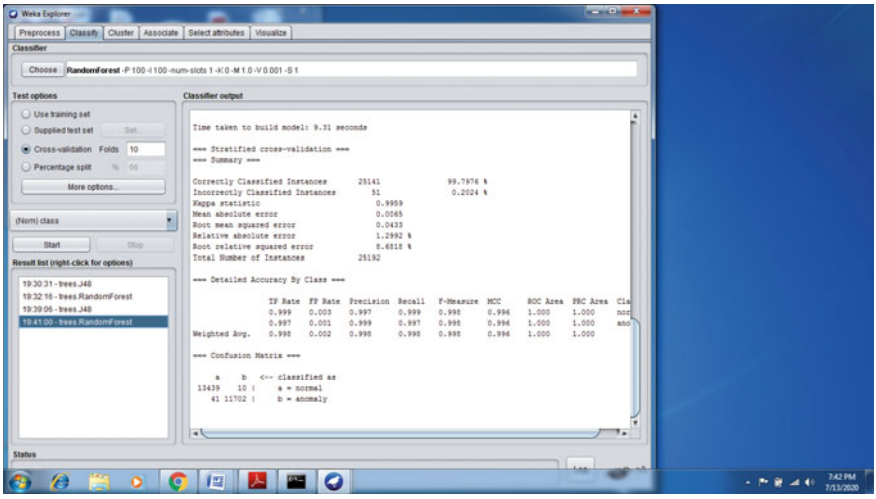

Fig. 4 Random forest with 42 attributes

References

1. Lee W, Stolfo S, Mok K (1999) A data mining framework for adaptive intrusion detection model. In: Proceeding of 1999 IEEE symposium on security and privacy. IEEE, pp 120–132
2. MIT Lincoln Laboratory-DARPA Intrusion Detection Evaluation http://www.ii.mit.edu/ist/ideval/docs/docs_index.html

3. Thuraisingham BM (2011) Data mining for malicious code detection and security applications. In: 2011 European intelligence and security informatics conference, pp 15–18. https://doi.org/10.1109/EISIC.2011.80
4. Dutt I, Borah S (2015) Some studies in intrusion detection using data mining techniques. Int J Innov Res Sci Eng Technol 4(7)
5. Lu T, Boedihardjo AP, Manalwar P (2005) Exploiting efficient data mining techniques to enhance intrusion detection systems. In: IRI-2005 IEEE international conference on infomation reuse and integration, pp 512–517. https://doi.org/10.1109/IRI-05.2005.1506525
6. Modi U, Jain A (2015) A survey of IDS classification using KDDCUP 99 dataset in WEKA. Int J Sci Eng Res 6(11):947–954
7. Li Y, Fang BX, Guo L (2006) A lightweight intrusion detection model based on feature selection and maximun entropy. In: 2006 international conference on digital object identifier, pp 1–4. https://doi.org/10.1109/ICCT.2006.341771
8. Ahmad I, Abdulah AB, Alghamdi AS, Alnfajan K, Hussain M (2011) Feature subset selection for network intrusion detection mechanism using genetic Eigen vectors. Proc CSIT 5
9. Abdulla SM, Najla B, Zakaria O (2010) Identify features and parameters. World Acad Sci Eng Technol 4(10):1553–1557
10. Sung AH, Mockamole S (2004) The feature selection and intrusion detection problems. In: Proceedings of the 9th Asian computing science conference 2004, lecture notes in computer science. Springer, pp 468–482
11. Zaman S, Karray F (2009) Features selection for intrusion detection systems based on support vector machines. In: Proceedings of the 2009 6th IEEE conference on consumer communications and networking conference. https://doi.org/10.1109/CCNC.2009.4784780
12. Nguyen H, Franke H, Petrovic S (2010) Improving effectiveness of intrusion detection by correlation feature selection. In: 2010 international conference on availability, reliability and security, pp 17–24. https://doi.org/10.1109/ARES.2010.70
13. Yu J, Kang H, Park D, Bang H, Kang DW (2013) An in-depth analysis on traffic flooding attacks detection and system using data mining techniques. J Syst Arch 59(10):1005–1012
14. Kim G, Lee S, Kim S (2014) A novel hybrid intrusion detection method integrating anomaly detection with misuse detection. Expert Syst Appl 41(4):1690–1700
15. Mukherjee S, Sharma N (2012) Intrusion detection using Naive Bayes classifier with feature reduction. C3IT-2012 Proc Technol 4:119–128
16. Meena G, Choudhary R (2017) A review paper on IDS classification using KDD99 and NSL-KDD dataset in WEKA. In: 2017 international conference on computer, communication and electronics (comptelix), pp 553–558. https://doi.org/10.1109/COMPTELIX.2017.8004032