Classification of Intrusion Detection Using Data Mining Techniques

Roma Sahani, Shatabdinalini, Chinmayee Rout, J. Chandrakanta Badajena, Ajay Kumar Jena and Himansu Das

Abstract Nowadays, Internet became a common way for communication as well as a key path for business. Due to the rapid use of Internet, its security aspect is turn more important day by day for which various network intrusion detection systems (NIDSs) are used to protect network data as well as protect the overall network from various attacks. Various intrusion detection systems (IDSs) are placed in different positions of network to protect it. There are various ways by which intrusion detection system can be implemented from which decision tree approach is most commonly used. It provides the easiest way to identify the most corrected field to select, manage, and make proper decision about their identification from a large dataset. This paper focuses to identify normal and attack data present in the network with the help of C4.5 algorithm which is one of the decisions tree techniques, and also it helps to improve the IDS system to identify the type of attacks present in a network. Experimentation is performed on KDD-99 dataset having number of features and different class of normal and attack type data.

R. Sahani · Shatabdinalini · J. Chandrakanta Badajena (✉)
Department of Information Technology, College of Engineering & Technology, Bhubaneswar, Odisha, India
e-mail: j.chandrakantbadajena@gmail.com

R. Sahani
e-mail: romasahani5@gmail.com

Shatabdinalini
e-mail: shatabdikisd@gmail.com

C. Rout
Department of Computer Science and Engineering, Ajay Binay Institute of Technology, Cuttack, Odisha, India
e-mail: chinu123.abit@gmail.com

A. K. Jena · H. Das
School of Computer Engineering, KIIT Deemed to be University, Bhubaneswar, Odisha, India
e-mail: ajay.bbs.in@gmail.com

H. Das
e-mail: das.himansu2007@gmail.com

P. K. Pattnaik et al. (eds.), *Progress in Computing, Analytics and Networking*, Advances in Intelligent Systems and Computing 710,
https://doi.org/10.1007/978-981-10-7871-2_72

Keywords NIDS · Decision tree · C4.5 · KDD-99

1 Introduction

Nowadays, every organization and institution use Internet for their communication as well as one of the business medium to reach to the customer. As much as the use of Internet increased, the growth of network attacks also increased accordingly [1] for which high confidence on network systems connectivity and their resources has generally increased the potential damage due to the presence of attacks which are launched against the systems from remote resources. It is quite very difficult to prevent all the type of attacks by using firewalls because every time different attacks contain unknown weaknesses or bugs [2]. Therefore, real-time intrusion detection systems are used to detect attacks and also used to stop an attack in progress; it gives an alarm signal to the authorized user or network administrator about the presence of malicious activity or the presence of attacks.

Intrusion detection includes a lot of tools and techniques such as machine learning, statistics, data mining, and so on for the identification of an attack [1, 3]. In recent years, data mining method for network intrusion detection system has been giving high accuracy and good detection on different types of attacks [4]. Decision tree technique is one of the intuitionist and frank classification methods in data mining which can be used for this purpose. It has a great advantage in extracting features and rules. So, the decision tree gives a greater significance to intrusion detection. The tree is constructed by identifying attributes and their connected values which will be used to examine the input data at each intermediary node of the tree. After the tree is formed, it can advise newly coming data by traversing, initial from a root node to the leaf node by visiting all the internal nodes in the path depending upon the test environment of the attributes at each node. The main issue in constructing decision tree is which value is chosen for splitting the node of the tree.

In this paper, an improved version of C4.5 algorithm is proposed from the basic concept of C4.5. The detection of intrusion components undergoes two stages. In the first stage, the algorithm evaluates the KDD-99 dataset [5] and constructs the decision tree for detecting the class type as ‘Normal’ or ‘Attack’ type of data in the leaf node in the tree. In the second stage, the classification of attack type is done which will show the attack type. We have considered four types of attacks such as DOS, R2L, U2R, and PROBE [6].

The rest of the paper is organized as follows. Section 2 provides the basic concepts about intrusion detection system with their attack types and also an idea about the decision tree. Section 3 provides some related work for IDS using DT. The proposed model for intrusion detection with respect to attack classification is presented in Sect. 4. In Sect. 5, we illustrated the results and experimental analysis of the proposed model with result comparisons. Finally, Sect. 6 concludes the work.

2 Basic Concepts

In this section, we discussed about IDS with its classification methods and about various attack classes. It also focuses how decision tree is constructed and various techniques in decision tree for making proper decision.

2.1 Intrusion Detection System

Intrusion detection system is the software systems which are basically designed to identify and helps to prevent the malicious activities and security strategy violations. Intrusion detection systems (IDSs) were first introduced by James P. Anderson [7]. These systems are placed at the choke point such as organization's connection to a stem line or can be placed on each of the hosts that are being monitored to defend from intrusion which is classified as analysis approach and placement of IDS [8].

Based on analysis approach, it can be either misuse detection or anomaly detection. Misuse detection approach is also known as signature-based approach which is basically used for detecting known type of attacks. In case of anomaly detection monitors the network traffic and compares it with established normal traffic profile. It makes decisions on the basis of network behaviors with the help of some statistical techniques. This approach is able to identify unknown attacks.

According to placement approach of IDS, it can be classified as host-based and network-based systems [9]. In host-based IDS, it is present on each host that needs for monitoring. It is able to determine if an attempted attack is successful and can detect local attacks. In Network Based Systems is monitored the network traffic from unauthorized access by which the hosts are makes secure connection with host systems. This mechanism takes less cost for deployment, and it is also promising for identifying attacks to and from multiple hosts.

Many researchers have proposed and implemented IDS according to which network attacks are having four major categories. Every attack on a network can be one of these attack type [6, 10].

Denial-of-Service Attack (DoS): Attacker makes the systems as busy as possible and also makes system or network resource unavailable to its actual users.

Remote to Local Attack (R2L): Attacker targets to access one or no of network systems which main purpose is to view or steal data illegally and introduce different type malicious software to network system.

User to Root Attack (U2R): This attack type starts working out with access to a normal user account on the system and is able to exploit some vulnerability to gain root access to the system.

Probing Attack: Probing is an attack in which the attackers scan the network systems in order to find out the weaknesses of the network system that may later be exploited so as to compromise the system.

2.2 Decision Tree

Data mining is the procedure for discovering knowledge from huge datasets with the help of statistics and artificial intelligence technique to solve complex real life problems [11–19]. The decision tree (DT) is an important classification method in data mining classification. A decision tree is defined as a flowchart-like or tree-like structure from different verities of data. In DT, each inner node represents a test on an attribute, whereas each stem represents the outcome of the test and each leaf node represent a class label. The path from root node to leaf node represents the classification rules [20]. From an intrusion detection perspective, classification algorithms can distinguish network data as attacks, benign, scanning, or any other category of interest using information like source/destination ports, IP addresses, and the number of bytes sent during a connection [21].

A decision tree classifier has a simple form which can be compactly stored and that efficiently classifies new data. This classifier consists of various algorithms like CART, ID3, C4.5 [22–24].

CART: Classification and regression tree (CART) was proposed by Leo et al. [25] as an umbrella term. It constructs a binary tree model which means a node in a dataset can only be divided into two groups. CART can handle any type of data like both categorical and numerical data. CART uses Gini index for selecting attribute. The attribute with the largest reduction in impurity is used for splitting the nodes of the dataset. It uses cost-complexity pruning and also generates regression trees [26].

ID3: In 1980, a machine researcher named J. Ross Quinlan developed a predictive modeling tool at the University of Sydney which is known as Iterative Dichotomiser 3 (ID3) [26]. This algorithm was designed based on the principles of Occam's razor, with the idea of creating the smallest, most efficient decision tree. ID3 uses information gain of each attribute for construction decision tree. The features having the highest gain can select for the splitting of data records. ID3 algorithm has some drawbacks, such as for a while, data may be over-classified, only one attribute at a time is considered for making decision tree. Only one attribute at a time is tested for making a decision, and it does not handle continuous attribute as well as missing value for making tree.

C4.5: C4.5 is the extension of ID3 algorithm and an arithmetical classifier. It overcomes the problems associated with ID3 algorithm like handling continuous data and missing values. It follows the same procedure as ID3 for categorical data and uses split ratio technique for numerical type of data. This approach is better

classifier in comparison with ID3. It can easily handle the missing values of the dataset.

3 Related Work

Many researchers proposed various anomaly detection mechanisms for IDS with the help of decision tree of which some are discussed below.

Rai et al. [27] used C4.5 decision tree for making taking decision. They have taken into consideration the two important issues, feature selection and split value for constructing the decision tree. In this paper, the algorithm was designed to address these two issues. This paper implements its algorithm by selecting the attributes from different levels of decision tree nodes, and then, it calculates the gain ratio for every attribute. Then select the attribute with the largest gain ratio to decide the root node of the decision tree. This paper used a novel approach for selecting the best attribute to split the dataset on each iteration. This work gets good accuracy with even less number of features selected using information gain.

Swamy and Vijaya Lakshmi [4] used two techniques like voting criteria and attribute selection method for selecting best attribute for splitting attribute for reaching in the leaf node or class node. In this work, firstly, all the attributes ware was selected and then attributes ware was shorted according to min-max principle. If all attributes ware belongs to the same class, then mark it as leaf node otherwise by using attribute selection method a best splitting attribute was taken. According to splitting criteria, the classification is carried out until it does not reach the leaf node. If all attribute does not belong to the same class then make a major voting for selecting the leaf level. It detects different types of attacks with high accuracy and less error-prone.

Phutane et al. [28] used data mining algorithms as improved C4.5 in order to detect the different types of attacks with high accuracy and less error-prone as well as it helps to increase performance of the system. In this approach, every time the input which is coming from the client system is stored in a database. If incoming data is similar to older one, then no need to go through the apriori algorithm, simply test the type of data which is already defined. If not, then apriori algorithm is applied which consists of associate rules in which all the data are collected. After that, all the frequent item set should be found by applying minimum support mechanism. Then, find the subsets which are common to at least a minimum number constant of the item sets. This would continue until there is no further extension or comparison is found. Then test the leaf data to the defined data type like attack types or normal data. It uses apriori algorithm for making decision tree, and minimum support mechanism gives the way for splitting of attributes or data. This proposed work overcomes the limitations of ID3 algorithm and also increases the system performance and better result in case of large database.

Shon Nadiammai and Hemalatha [29] gives four solutions for different IDS problems, they included the problem of data categorization, high level of personal effort, unlabeled data, and circulated denial-of-service attack efficacy. They solved the

first problem (classification of data) using efficient data adapted decision tree (EDADT). The aim of this method was to reduce the dimensionality of model by feature extraction of significant features to every type of attack. The authors compared the proposed algorithm to other methods like C4.5, SVM, and others. The results they obtained show that their algorithm achieved the highest accuracy rate.

By studying various aforesaid algorithms, it is obtained that for making decision tree selecting correct attribute and way of selection is more important and classifying attack is much more important factor in IDS. For these reasons, our main objective is to select the best attribute which has the highest gain ratio for constructing the decision tree which will give the result whether the data are a normal data or an attack data. And our work also focuses to classify the data which are attack type corresponding to its class type (DoS, R2L, U2R, and Probing).

4 Proposed Model

This system is the process of identifying the normal and attack data in the network.

For our work, we have used KDD99 dataset which was used in the third International Knowledge Discovery and Data Mining Tools for building a network intrusion detector [30]. It consists of 42 features including class column having normal and attack type data. The proposed model undergoes two stages. In the first stage, the network dataset is preprocessed in which dataset having discrete type of data is converted to numeric data and then decision tree is constructed with the help of preprocessed dataset which is capable of distinguishing the record of normal data or attack data in the leaf node of the tree. The second stage is the detection phase which identifies the attacks corresponding to its class type with their number of occurrences and identifies the noisy data or missing-valued data named as unknown attack. Figure 1 represents the proposed model of our approach.

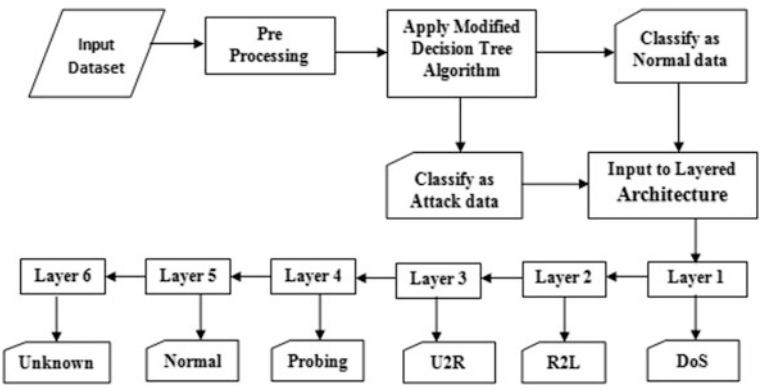

Fig. 1 Flow of proposed model

4.1 Proposed Modeling Framework

This proposed model is an improved version of C4.5 algorithm which handles both continuous and categorical data simultaneously for classifying dataset as normal and attack at leaf level. In basic C4.5, the dataset requires a shorted format, and it handles categorical and continuous data separately which is a time-consuming process and also selecting split value is an important factor for making decision tree. By focusing such scenarios, we have modified some cases in our proposed model, like instead of handling both categorical and continuous data separately, we simply convert the categorical data to continuous data in preprocessing and without any shoring the dataset we directly apply the algorithm for classification. For splitting purpose, we have applied geometric mean which helps to give a better decision tree result. This model makes the decision tree for identification of attack and normal data present in the dataset. The steps of the algorithm are as follows:

Algorithm

Input: Any Dataset with
Number of samples in dataset (RW), Number of unique elements in class column (UC), Column in Dataset (D), Number of distinct values presents in D column (V), D_j: Number of each element in that D column (D_j), Number of unique element in column (T_i)
Output: Classified data

Begin

1. If input dataset having same class type, then
   Leaf ← class name.
2. If single class is present in input data
   Leaf ← Histogram (class column)
3. Entropy of dataset$(\mathrm{Entp(RW)}) = \sum_{i=1}^{UC} \frac{\mathrm{freq\ (Ti,RW)}}{|\mathrm{RW}|} * \log \frac{\mathrm{freq\ (Ti,RW)}}{|\mathrm{RW}|}$
4. Information of each attributes $(\mathrm{INFO_{Att}(D)}) = \sum_{j=1}^{V} |\frac{D_j}{RW}| * \mathrm{Entp}(D_j)$
5. Information Gain (IG (D)) = Entp (RW) - INFOAtt (D)
6. Split information $(\mathrm{Split\_info(D)}) = \sum_{j=1}^{V} |\frac{D_j}{RW}| * \log |\frac{D_j}{RW}|$
7. Gain Ratio $(\mathrm{Gain_{Ratio\ (D)}}) = \frac{\mathrm{IG\,(D)}}{\mathrm{Split\ \_info\ (D)}}$
8. Decision node (a_best) ← highest gain ratio attribute
9. Split value ← means (a_best attribute's values)
   Left subset ← (dataset < split value)
   Right subset ← (dataset > split value)
10. Repeat steps from 1 to 9 on each subsets produced by dividing the set on attribute 'a_best' and insert those nodes as descendant of parent node.

End

After successfully construction of decision tree, we have to classify the attacks according to its classification, such as either DoS or U2R, R2L or PROB type. For

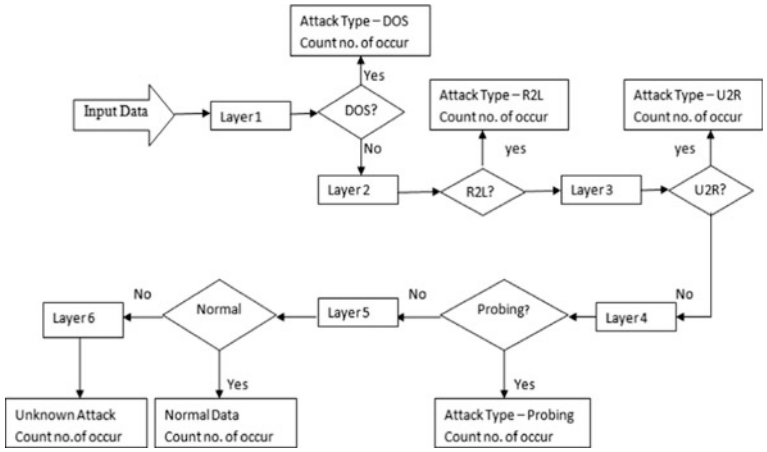

Fig. 2 Flow of attack classification

this reason, the dataset goes to the layered architecture to finding attack type with its number of occurrences.

The input dataset samples undergo six stages in which at each stage the samples are compared against the attack class type. Each attack class consists of its attack name which belongs to that category. In each layer, the test result shows the number of samples that belong to its category. Suppose a sample goes to the layered architecture, then first it checks whether it is a DoS type if yes the simple discard that sample and make DoS type as one, if no then it checked with R2L type if yes then increase the no of occurrence of R2L type if no then check whether it is U2R type and then Probing type. If the sample does not belong to these four attack types then check whether it is a normal type data or not, if yes then increase the number of occurrence of normal type, and if it does not belong to any one of these categories, then make it as unknown attack type and note the number of occurrences. Finally, the result shows the number of samples presents in each class type. In every layer, the data are filtered to its appropriate class type. The classification of the flow of attack is shown in Fig. 2.

5 Results and Experimental Analysis

In order to evaluate the performance of proposed algorithm for network intrusion detection, we work on five different class classifications with the help of KDD-99 dataset for both training and testing purpose. For the training purpose, 21606 number of random samples is taken which consists of 20 number of class types as one normal and 19 attack types. These 19 attacks belong to four major attack categories with some undefined attack category mark it as unknown attack type.

Table 1 Result of proposed model for data classification

Number of samples		Number of samples
Normal		10221
Attack	DoS	9083
	R2L	1877
	U2R	12
	Probing	300
	Unknown	113

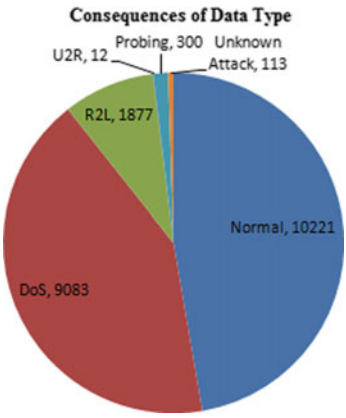

Fig. 3 Result of types of attacks and non-attack

The purposed model successfully builds DT with proper identification of attack type. From the total 21606 number of samples, verities of data which are identified from the proposed model are given in Table 1.

The graphical representation of the result table is represented in Fig. 3.

The output scenario of proposed model is compared with other decision tree algorithm like CART, existing C4.5 which is an extension of ID3 algorithm with respect to performance time, accuracy, and error percentage which gives an idea that proposed model gives better result when compared to other two presented in Table 2 and graphically represented in Fig. 4.

Table 2 Result comparison with other classifiers

Classification techniques	Performance time (min)	Accuracy rate (%)	Error rate (%)
CART	1.3	86.42	13.58
ID3 extension (C4.5)	16.26	99.079	0.91
Proposed model	4.5	99.79	0.51

Fig. 4 Comparison of performance, accuracy, and error rate in different techniques with proposed model

6 Conclusion

Decision tree is one of the most effective and popular technique for intrusion detection system. It can make proper decision of whether the incoming network traffic data are either an attack or normal data. The proposed model builds the decision tree with the help of gain ratio and geometric mean for splitting the dataset. It also successfully identifies different type of attacks present in the dataset with identification of unknown data. The result of the proposed model is compared with other DT techniques like CART and ID3 extension (C4.5) with the help of KDDCUP-99 dataset, and proposed model gives 99% accuracy for attack identification with lesser time. The advantages of proposed model over C4.5 are the probability to achieve high detection rate over dissimilar types of attacks with less error rate and time. The future work is to test the performance of this model over a large dataset and also to handle the classification of unknown attack in an automatic control system.

References

1. Barbara, Daniel, et al.: ADAM: Detecting intrusions by data mining. In Proceedings of the IEEE Workshop on Information Assurance and Security. (2001): 11–16.
2. Swamy, K.V.R., and K.S. Vijaya Lakshmi: Network intrusion detection using improved decision tree algorithm. International Journal of Computer Science and Information Security 10.8 (2012): 4971–4975.
3. Farid, Dewan Md, et al.: "Attacks classification in adaptive intrusion detection using decision tree." World Academy of Science, Engineering and Technology 63 (2010): 86–90.

4. IDS over Firewall, https://www.scribd.com/document/45263670/Limitations-Of-Firewall. January 2017.
5. Sarkar, Sutapa: High Performance Network Security Using NIDS Approach. International Journal of Information Technology and Computer Science (IJITCS) 6.7 (2014): 47–55.
6. Das, Niva, and Tanmoy Sarkar: Survey on host and network based intrusion Detection System. Int. Journal of Advanced Networking and Applications 6.2 (2014): 2266–2269.
7. KDD99 dataset, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, 2017.
8. Paliwal, Swati, and Ravindra Gupta: Denial-of-service, probing & remote to user (R2L) attack detection using genetic algorithm. International Journal of Computer Applications 60.19 (2012): 57–62.
9. Kumar, Sandeep, and Satbir Jain: "Intrusion detection and classification using Improved ID3 algorithm of data mining." International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) 1.5 (2012): 352–356.
10. Moon, Daesung, et al.: DTB-IDS: An intrusion detection system based on decision Tree using behavior analysis for preventing APT attacks. The Journal of supercomputing (2015): 1–15.
11. P Sarkhel, Himansu Das, and L K Vashishtha, "Task Scheduling Algorithms in Cloud Environment", In 3rd International Conference on Computational Intelligence in Data Mining, Springer India, 2017.
12. I Kar, RNR Parida, Himansu Das, "Energy Aware Scheduling using Genetic Algorithm in Cloud Data Centers" in International Conference on Electrical, Electronics, and Optimization Techniques, IEEE, 2016.
13. Himansu Das, A K Jena, P K Rath, B Muduli, S R Das, "Grid Computing Based Performance Analysis of Power System: A Graph Theoretic Approach", in International Conference on Intelligent Computing, Communication & Devices, Springer India, 2015, pp. 259–266.
14. Himansu Das, G S Panda, B Muduli, and P K Rath. "The Complex Network Analysis of Power Grid: A Case Study of the West Bengal Power Network." In International Conference on Advanced Computing, Springer India, 2014, pp. 17–29.
15. KHK Reddy, Himansu Das, D S Roy, "A Data Aware Scheme for Scheduling Big-Data Applications with SAVANNA Hadoop", in Futures of Network, CRC Press, 2017.
16. Panigrahi, C R, M Tiwary, B Pati, and Himansu Das., "Big Data and Cyber Foraging: Future Scope and Challenges." In Techniques and Environments for Big Data Analysis, Springer India, 2016, pp. 75–100.
17. Himansu Das, D.S.Roy, "A Grid Computing Service for Power System Monitoring," International Journal of Computer Applications (IJCA), 2013, Vol. 62 No. 20, pp 1–7
18. Himansu Das, Bighnaraj Naik, Bibudendu Pati, and Chhabi Rani Panigrahi, "A Survey on Virtual Sensor Networks Framework," International Journal of Grid & Distributed Computing (IJGDC), 2014, Vol. 7 no. 5, pp 121–130
19. Himansu Das, D.S.Roy, "The Topological Structure of the Odisha Power Grid: A Complex Network Analysis", in International Journal of Mechanical Engineering and Computer Applications (IJMCA), 2013, Vol.1 Issue 1, pp 12–18
20. Rathee, Anju, and Robin Prakash Mathur: Survey on decision tree classification algorithms for the evaluation of student performance. International Journal of Computers & Technology 4.2a1 (2013): 244–247.
21. Patel, B.R. and Kushik K.R.: A survey on decision tree algorithm for classification. Int. Journal of Engineering Development and Research 2.1 (2014): 1–5.
22. IDS History, http://csrc.nist.gov/publications/history/ande80.pdf. May 2017.
23. Das, Himansu, Ajay Kumar Jena, Janmenjoy Nayak, Bighnaraj Naik, and H. S. Behera. "A novel PSO based back propagation learning-MLP (PSO-BP-MLP) for classification." In Computational Intelligence in Data Mining-Volume 2, pp. 461–471. Springer, New Delhi, (2015).
24. DARPA Intrusion Detection Evaluation KDD dataset, http://kdd.ics.uci.edu//databases/kddcup98/kddcup98.html. December 2016.
25. CART model, http://www.datasciencecentral.com/profiles/blogs/introduction-to-classification-regression-trees-cart. February 2017.

26. Quinlan, J. Ross: Induction of decision trees. Machine learning 1.1 (1986): 81–106.
27. Recent attack Presents over internet, http://www.internetworldstats.com/stats.htm. May 2017.
28. Rai, Kajal, M. Syamala Devi, and Ajay Guleria: Decision Tree Based Algorithm for Intrusion Detection, Int. Journal of Advanced Networking and Applications 7.4 (2016): 2828–2834.
29. Phutane, Ms Trupti, and Apashabi Pathan: Intrusion detection system using decision tree and apriori algorithm. Journal of Computer Engineering and Technology 6.7 (2015): 09–18.
30. Shon Nadiammai, G.V., and M. Hemalatha: Effective approach toward Intrusion Detection System using data mining techniques. Egyptian Informatics Journal 15.1(2014): 37–50.