International Journal of Technology Research and Management
ISSN (Online): 2348-9006
Vol 6 Issue 4 April 2019

# An Efficient Intrusion Data Classification using Improved Decision Tree Algorithm

Priyanka Pawar[1], Harish Patidar[2]
Research Scholar (MTech), Department of Computer Science Engineering, Lakshmi Narain College of Technology Indore M.P, India[1]
Associate Professor, Department of Computer Science Engineering, Lakshmi Narain College of Technology Indore M.P, India[2]
it.priyankapawar@gmail.com[1], harish.patidar@gmail.com[2]

***Abstract:*** *Intrusion detection is to detect or analyze attacks against a computer system and reports the attacks to network administrator, its plays an important role in network security. It is useful in business sector as well as an active area of research. In Information Security, intrusion detection is the method of detecting actions that attempt to compromise the confidentiality, integrity or availability of a resources . The objective of this research work is to analyze the features of intrusion datasets using dimensionality reduction by using information gain ratio and Classification of the reduced feature subsets using the better version or improved Decision tree classifiers with optimization that improve the accuracy of intrusion detection classification. The proposed decision tree based classifier is compared with the existing classifiers in terms of accuracy and efficiency.*

***Keywords:*** *IDS, Feature Selection, Classification, Decision tree, Information gain, Attacks, weka, NSL-KDD Dataset.*

## 1. INTRODUCTION

Intrusion detection is very important in identifying the harmful activity or intruders who break into the system. To protect the network against a various network attacks and malicious activity we need a security mechanism.

An IDS is a device or software application that monitors the system or network for harmful or malicious activities, or any security violation and generates alert or reports to network administrator, where it is analyzed for further prevention and detection [3]. An intrusion detection system simply scans or monitors network traffic and alert the network administrator of any harmful activity. Intrusion detection is of two types Host based intrusion detection system (HIDS) and network based intrusion detection system (NIDS) according to the place they used [1].

IDS are classified as signature based or misuse based and Anomaly based. Signature based IDS uses various previously stored patterns or attacks in the database, also called signatures database for detection. Anomaly based IDS detects the network activity which are deviates from the normal system behavior stored in database. Signature based intrusion detection they identify attacks with more accuracy and less false alarms. They are easy to implement and configure [2].

Data mining is a process of finding patterns in large data Sets and transform it into understandable form [3]. Classification maps the data into predefined classes. Classification inputs dataset and outputs classifiers which classify the new data in future.

Machine learning uses data mining techniques and other learning algorithms to build models to predict future outcomes [6]. Machine learning algorithms like supervised and unsupervised learning techniques.

## 2. RELATED WORK

Venkata et al. [11] as the cost of the data processing and Internet accessibility increases, more and more organizations are becoming vulnerable of cyber threats. Most current offline intrusion detection systems are focused on

unsupervised and supervised machine learning approaches. In this system, Information Gain (IG) and Triangle Area based KNN are used for selecting more discriminative features by combining Greedy k- means clustering algorithm and SVM classifier to detect Network attacks. This system achieves high accuracy detection rate and less error rate.

Esh et al. [12] The unsupervised learning techniques using the machine learning for intrusion detection datasets, we know that Clustering is the best techniques on the efficient data mining for intrusion detection. This algorithm K-means is widely used for intrusion detection, because it gives efficient results.

Deepika et al. [13] Intrusion detection is an active area of research in a current scenario. In their work the object is to affect a method for intrusion detection using KNN classification and Dempster theory of evidence.

Prabhu et al. [14]. The proposed system is based on the adaboost algorithm with Naive Bayes classifier to detect network intrusions low false-alarm rates and high detection rates. This results shows low error rates and less computational complexity

Nagarajan et al. [15] IDS which are increasingly a key part of system defense are used to identify abnormal activities in a computer system. In general, the traditional intrusion detection relies on the extensive knowledge of security experts, in particular, on their familiarity with the computer system to be protected

Nasser et al. [16] the rapid growth of Internet malicious activities has become a major concern to network forensics and security community. With the increasing use of internet there is a need for stronger intrusion detection mechanisms

Debdutta et al. [17] in multi-hop wireless systems, the need for cooperation among nodes to relay each other's packets exposes them to a wide range of security attacks. In a wormhole attack, where a malicious node records and control traffic at one location and passes it to another compromised node, some remote place, which plays it locally.

Dianbo et al. [18] Neural Networks approach is an advanced methodology used for intrusion detection. As a type of Neural Network, Self-organizing Maps (SOM) is getting more attention in the field of intrusion detection.

Hazem et al. [19] E-government is an important issue which integrates existing local area networks into a global network that provide many services to the nation citizens This network requires a strong security infrastructure to guarantee the confidentiality of data and the availability of services.

Todd Heberlein [20] proposed an intrusion detection system called network system monitor. This system is based on the concept of analyzing network instead of the system log entry.

Teng, Chen, And Lu [21], proposed time based inductive machine to capture or store user behavior. Inductive generalization is also a part of the process.

Anderson D, Lunt TF, Javitz H, Tamaru A, Valdes [22], proposed a network intrusion detection expert system. This system learns from the training data and predicts the test data. Lee W. and Stolfo S. and Mok [23] propose a novel data mining based framework for intrusion detection. This model is based on the concept of the utilizing the contents of the audited programs.

Debar, H., Dacier, M., And Wespi [24] proposes taxonomy of the intrusion detection systems. This classification is done according to the property of the intrusion detection system.

## 3. PROPOSED METHODOLOGY

The proposed methodology consist of data preprocessing Data classification and pruning the tree classifier for better accuracy.

### 3.1 Data Pre-processing

Data pre-processing is the first step in which document conversion from text to Csv or Arff file format [29], Feature selection and feature weighting which eliminates the irrelevant and redundant feature by relevant feature selection method using Weka tool [29].

### 3.2 Decision tree

Decision tree is one of the classification techniques in data mining. Decision tree builds classification or regression models in the form of tree structure where internal node denotes a test on an attribute, branch represents an outcome of the test and leaf node represents a class label[8]. The result is a decision tree with decision nodes and leaf nodes. The classic decision tree algorithm, C4.5 was proposed by Quinlan [8]. Majority of the research works in decision trees concerned with the improvement in the performance using optimization techniques.

C4.5 algorithm:

1. If all the instances of S belong to the same target class, the leaf node is created labeling with the same class name.
2. The information is calculated for every attribute, by a test on the attribute using gain ratio and information gain:

Entropy is calculated by:

$$\text{Ent}(S, A) = - \sum_{j=1}^{\text{class\_num}} \frac{\text{freq}(Lj,S)}{|S|} * \log_2 \left(\frac{\text{freq}(Lj,S}{|S|}\right) \quad (1)$$

Information Gain:

$$IG(S, A) = \text{Ent.}(S) - \sum_{v \in \text{values}(A)} \frac{|S\_v|}{|S|} * \text{Ent.}(S\_v) \quad (2)$$

Then the gain ratio is calculated by dividing the information gain of an attribute with split value of that attribute.

$$\text{Gain ratio}(A) = \frac{IG(a)}{\text{Split}(a)} \quad (3)$$

3. Select the best attribute using Selection measures like Gain ratio, Information gain and Gini Index to split the records.

4. Make that best attribute a decision node and breaks the dataset into smaller subsets.

5. Starts building the tree by repeating this process recursively for each child until one of the condition will match:

a. All the tuples belong to the same attribute value.

b. No more remaining attributes.

C. No more instances.

3.3. Pruning

It is found that the present algorithm [5] constructs empty branches containing zero values. It is difficult to take decision that how deeply to grow the decision tree and its result in an unbalanced tree. It is also difficult to choose an best attribute selection measure and manage training data with missing attribute values. That all is resulting in less accuracy and also taking additional time in model construction and search. Tree Pruning /Optimization.

1. By using above pseuocode the decision tree constructed.
2. All non leaf nodes of the tree are evaluated in bottom up Fashion.
3. All the nodes that do not have any impact on correctness of tree are eliminated.

## 4. IMPLEMENTATION

The experiments are performed on a 64-bit windows 10 pro operating system with 4 GB of RAM and a Intel (R) core Processor with CPU speed of 1.90GHz .For designing and implementation we use Machine learning tool Weka and Java Net beans . NSL-KDD dataset is used for network intrusion detection. The proposed algorithm is compared with the existing classifiers like Naive Bayes, CART, SVM, and C4.5.

4.1 Dataset

The NSL-KDD is the updated and improved version of KDD 99 dataset. It consists of 41 features and a class labeled as normal and attack. The attacks are grouped into four classes: DOS, U2R, R2L and Probe[23]. The experiments were performed on training data set. The first step is pre-processing ,to obtain reduced dataset, for efficient and accurate classification .we have to perform training and testing on these reduced dataset on proposed decision tree classifier to train , build and evaluate the model on attacked dataset.

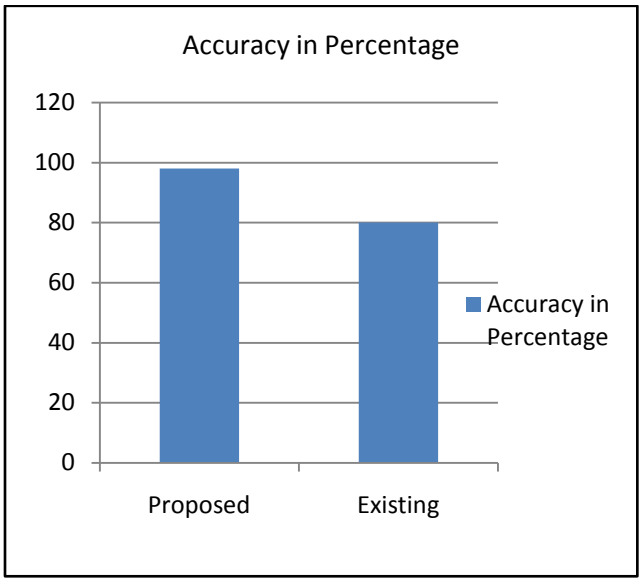
The following types of attacks classes present in NSL-KDD dataset grouped into four categories [23]:

1. DOS: A Denial of service is a security attack where the attacker tries to prevent the legitimate user from accessing the network service or resources.e.g.-SYN flood, Teardrop attacks etc.
2. U2R: A User to Root attack where the attacker has local access (unauthorized) to the victim machine and tries to gain super user privileges (root).e.g. - Buffer overflow attacks.
3. R2L: Attacker does not have an account on the victim machine, hence tries to gain unauthorized access through remote machine .e.g. - password guessing.
4. Probe: Attacker tries to gain information about the targeted remote victim.E.g.-port scanning.

4.2 Result Analysis

The dataset used in experimental study is NSL-KDD.It has 41 features and is labeled as either normal or an attack, with exactly one specific type- DOS, U2R, R2L and Probe[6]. We have performed classification on selected and relevant attribute which contains particular attack gives better performance. The experiment shows that proposed method is better than existing methods in terms of accuracy and efficiency.

We have used improved C4.5 algorithm for building this model and the resulted tree is optimized as it proved to be more efficient in detection of attacks classification.




## 5. CONCLUSION

It is found that there are many existing methods for classification of IDS data but still we have scope to improve the accuracy of classifier by using different similarity measures. This paper presented an updated decision tree based algorithm for attack data classification. Useless attributes were eliminated from the fully constructed decision tree. This approach is also able to deal with attributes with missing values. Experimental set up has shown that the memory consumption of the proposed method is also less. In future we use this approach with C5.0 decision tree.

## REFERENCES

[1] K. Kendall, "A Database of Computer Attacks for the Evaluation of Intrusion Detection Systems, "Massachusetts Institute of Technology Master's Thesis, 1998.
[2] Pieter de Boer & Martin Pels, "Host-based Intrusion Detection Systems", Revision 1.10, pp.19-20– February 4, 2005.
[3] K. Asif, Talha A. Khan, Sufyan Yakoob, "Network Intrusion Detection and Its Strategic Importance", IEEE Beiac, P.P 978-1-4673, September 2013.
[4] Mukherjee, Biswanath L, Heberlein T, Levitt KN.Network Intrusion detection system .IEEE Network8 (3):26-41:1994.
[5] K. Rai, M. S. Devi, A. Guleria, "Decision Tree Based Algorithm for Intrusion Detection", Int.J Advanced Networking and Application, 2016.
[6] R. A. Kemmerer and G.Vigna, "Intrusion detection: a brief history and overview," Computer, vol. 35, no.4pp. 27-30, 2002.
[7] Alireza Osareh, Bita Shadgar (Computer Science Department, Faculty of Engineering, Shahid Chamran University, Ahvaz, Iran),"Intrusion Detection in Computer. Networks based on Machine Learning Algorithms", IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.11, November 2008.
[8] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, 1993.
[9] Vangie Beal," Intrusion Detection (IDS) and Prevention (IPS) Systems", posted 2005 [07-15-2005], last updated 2010[08-31-2010].
[10] Aydın M. A., Zaim A. H., Ceylan K. G., A hybrid intrusion detection system design for computer network security, Computers and Electrical Engineering. 35,517-526, 2009.
[11] Venkata SuneethaTakkellapati,G.V.S.N.R.V Prasad," "Network Intrusion Detection system based on Feature Selection and Triangle area Support Vector Machine", International Journal of Engineering Trends and Technology-Volume3, Issue 4, 2012.
[12] Esh Narayan, Pankaj Singh and Gaurav Kumar Tak, "Intrusion Detection System Using Fuzzy C-Means Clustering with Unsupervised Learning via EM Algorithms" VSRD-IJCSIT, Vol. 2 (6), 502-510, 2012.
[13] Deepika Dave, Prof. Vineet Richhariya, "Intrusion detection with KNN classification and DS- theory", IRACST Vol. 2, No.2, April 2012.
[14] P.S. Prabhu, "Network Intrusion Detection Using Enhanced Adaboost Algorithm", International Journal of Communications and Engineering Volume 3, No.3, and Issue: 02 March 2012.
[15] R. Shanmugava divu, Dr. N. Nagarajan, "Network Intrusion Detection System Using Fuzzy Logic" IJCSE Vol. 2 No. 1, 2011.
[16] Nasser S. Abouzakhar And Abu Bakar, "A Chi-Square Testing-Based Intrusion Detection Model", CFET, 2010.
[17] Debdutta Barman Roy, Rituparna Chaki, Nabendu Chaki, "A New Cluster-Based Wormhole Intrusion Detection Algorithm for Mobile Ad-Hoc Networks", IJNSA, Vol 1, No 1, April 2009.
[18] Dianbo Jiang, Yahui Yang, Min Xia, "Research on Intrusion Detection Based on an Improved SOM Neural Network", IEEE 2009.
[19] Hazem M. El-Bakry, Nikos Mastorakis, "A Real-Time Intrusion Detection Algorithm for Network Security", Wseas Transactions on Communications Issue12, Volume 7, December 2008.
[20] Todd, H. L., Gihan V.D., Karl N.L., Biswanath, M., Jeff, W. and David, W. "A network security monitor," in Proceedings of Symposium on Research in Security and Privacy, Oakland, CA, pp. 296–304, 1990.
[21] A. Teng, H., Chen, K. and Lu, S. "Adaptive real time anomaly detection using inductively generated sequential patterns', IEEE Computer Society Symposium on Research in Security and Privacy, California, IEEE Computer Society, pp. 278-84, 1990.

[22] A. Anderson, J.B. and Mohan, S. "Sequential coding algorithms: A survey and cost analysis", IEEE Transactions on Communication, Vol.32, pp. 169-176, 1984.
[23] A. Lee, W., Stolfo, S. and Mok, K. "Adaptive intrusion detection: A data mining approach", Artificial Intelligence Review, Kluwer Academic Publishers, Vol. 14, No.6, pp. 533-S567, 2000.
[24] A. Debar, H., Becker, M. and Siboni, D. "A neural network component for an intrusion detection system," in IEEE Symposium on Research in Computer Security and Privacy, pp. 240-250, 1992.
[25] K. Rai, M. S. Devi, A. Guleria, "Decision Tree Based Algorithm for Intrusion Detection", Int. J Advanced Networking and Application, 2016.
[26] Priyanka pawar and Harish Patidar, "A Survey of various Techniques for Intrusion Detection System", IJTRM Vol 4 Issue11, 2017.
[27] Lior Rokach and Oded Maimon. . "Top-down Induction of Decision Trees Classifiers-A Survey." IEEE Transaction on systems, man, and cybernetics—Part C: Applications and Reviews, VOL. 35, NO. 4, November 2005.
[28] A.S. Galathiya, A. P. Ganatra, and C. K. Bhensdadia. "Improved Decision Tree Induction Algorithm with Feature Selection, Cross Validation, Model Complexity and Reduced Error Pruning." International Journal of Computer Science and Information Technologies 3, no. 2 (2012).
[29] http://www.cs.waikato.ac.nz/ml/weka/documentation.html.