JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Decision Tree C4.5 Machine Learning Approach based Network Intrusion Detection for IoT Security Application

[1]Amaresh Kumar, [2]Dr. Ashish Kumar Khare
[1]Research Scholar, [2]Professor & Head
Department of Computer Science Engineering,
Lakshmi Narain College of Technology & Science, Bhopal, India

Abstract : Intrusion detection is one of the important security problems in today's cyber world. A significant number of techniques have been developed which are based on machine learning approaches. So for identifying the intrusion we have designed the machine learning algorithms. By using the algorithm we find out intrusion and we can identify the attacker's details also. IDS are mainly two types: Host based and Network based. A Network based Intrusion Detection System (NIDS) is usually placed at network points such as a gateway and routers to check for intrusions in the network traffic. This paper presents the C4.5 decision tree algorithm for classification. The C4.5 algorithm is used in Data Mining as a Decision Tree Classifier which can be employed to generate a decision, based on a certain sample of data. The simulation results shows that the proposed approach gives the significant good results in term of the precision, recall, F1-Score, Error Rate and accuracy. The overall achieved accuracy is 96.3% or approx 97% with the 3% error rate.

***IndexTerms* - NIDS, C4.5, Decision Tree, accuracy, IoT, IDS, Cyber, Attack, Security.**

I. INTRODUCTION

A cyber attack can be employed by sovereign states, individuals, groups, society, or organizations, and it may originate from an anonymous source. A product that facilitates a cyber attack is sometimes called a cyber weapon. A cyber attack may steal, alter, or destroy a specified target by hacking into a susceptible system. Cyber attacks can range from installing spyware on a personal computer to attempting to destroy the infrastructure of entire nations. Legal experts are seeking to limit the use of the term to incidents causing physical damage, distinguishing it from the more routine data breaches and broader hacking activities.

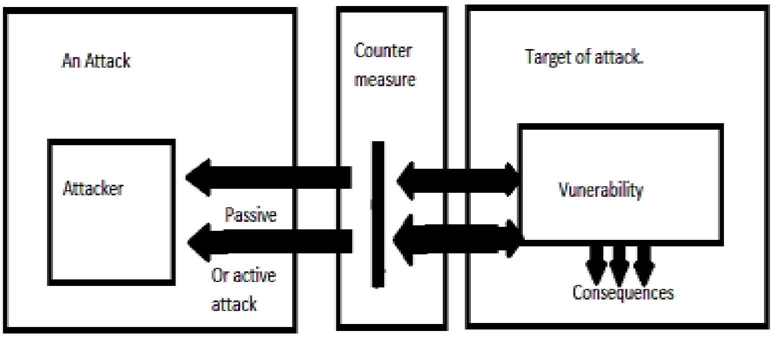


Figure 1: Attack system

Things have evolved due to the convergence of multiple technologies, real-time analytics, machine learning, ubiquitous computing, commodity sensors, and embedded systems. Traditional fields of embedded systems, wireless sensor networks, control systems, automation (including home and building automation), and others all contribute to enabling the Internet of things. In the consumer market, IoT technology is most synonymous with products pertaining to the concept of the "smart home", including devices and appliances (such as lighting fixtures, thermostats, home security systems and cameras, and other home appliances) that support one or more common ecosystems, and can be controlled via devices associated with that ecosystem, such as smart phones and smart speakers. The IoT can also be used in healthcare systems.

II. METHODOLOGY

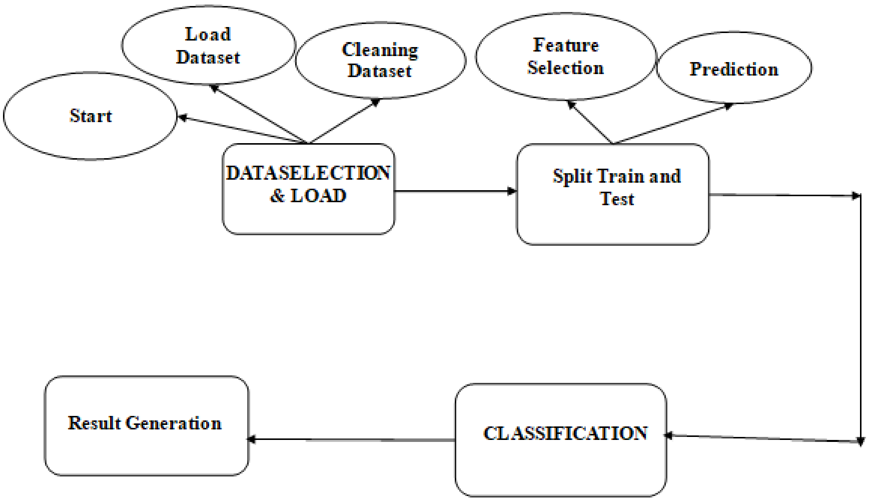


Figure 2: Flow Chart

Steps-

- Firstly, download the dataset from KDD dataset kaggle website, which is a large dataset provider company for research.
- Now preprocessing of the data, here handing the missing dataset. Remove the null value or replace from common 1 or 0 value.
- Now apply the classification method based on the machine learning approach. The Decision Tree (DT) with C4.5 machine learning method is applied.
- Now check and calculate the performance parameters in terms of the precision, recall, F_measure, accuracy and error rate.

In the case is Decision Trees, it is essential that the node are aligned as such that the entropy decreases with splitting downwards. This basically means that the more splitting is done appropriately; coming to a definite decision becomes easier.
So, we check every node against every splitting possibility. Information Gain Ratio is the ratio of observations to the total number of observations (m/N = p) and (n/N = q) where m+n=Nm+n=N and p+q=1p+q=1. After splitting if the entropy of the next node is lesser than the entropy before splitting and if this value is the least as compared to all possible test-cases for splitting, then the node is split into its purest constituents.

1. Check for the above base cases.
2. For each attribute a, find the normalised information gain ratio from splitting on a.
3. Let a_best be the attribute with the highest normalized information gain.
4. Create a decision node that splits on a_best.
5. Recur on the sublists obtained by splitting on a_best, and add those nodes as children of node.

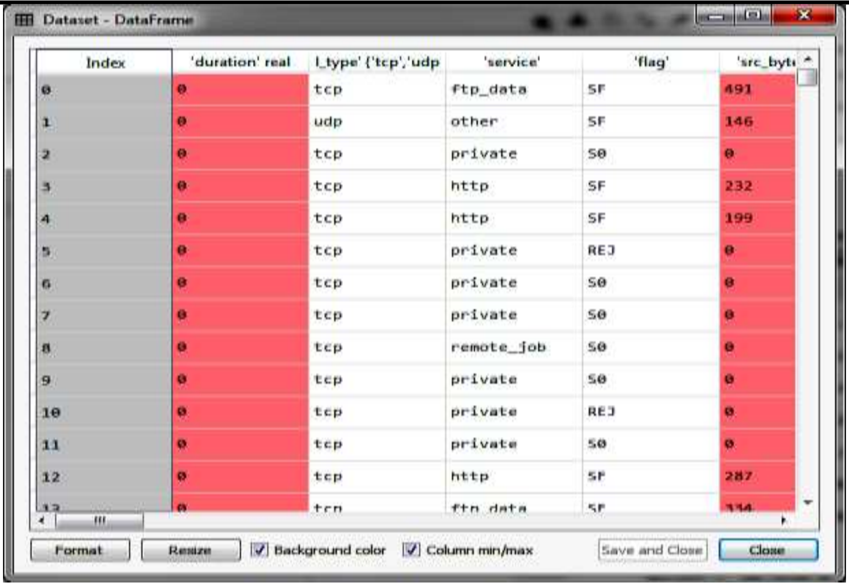
Advantages of C4.5 over other Decision Tree systems:

1. The algorithm inherently employs Single Pass Pruning Process to mitigate over fitting.
2. It can work with both Discrete and Continuous Data
3. C4.5 can handle the issue of incomplete data very well

We should also keep in mind that C4.5 is not the best algorithm out there but it does certainly prove to be useful in certain cases.

III. SIMULATION RESULTS

The implementation of the proposed algorithm is done over python spyder 3.7. The sklearn, numpy, pandas, matplotlib, pyplot, seaborn, os library helps us to use the functions available in spyder environment for various methods like decision tree, random forest, naive bayes etc.

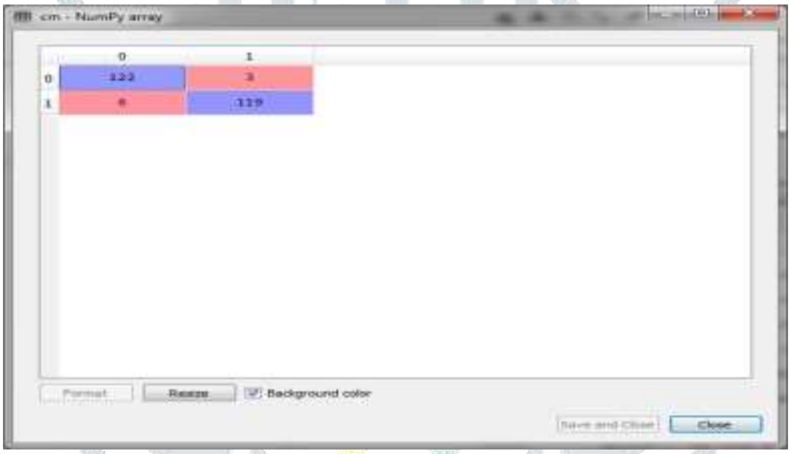


Dataset - DataFrame

Index	'duration' real	l_type' {'tcp','udp	'service'	'flag'	'src_byt
0	0	tcp	ftp_data	SF	491
1	0	udp	other	SF	146
2	0	tcp	private	S0	0
3	0	tcp	http	SF	232
4	0	tcp	http	SF	199
5	0	tcp	private	REJ	0
6	0	tcp	private	S0	0
7	0	tcp	private	S0	0
8	0	tcp	remote_job	S0	0
9	0	tcp	private	S0	0
10	0	tcp	private	REJ	0
11	0	tcp	private	S0	0
12	0	tcp	http	SF	287
13	0	tcp	ftp_data	SF	334

Figure 3: Dataset

Figure 3 is showing the KDD data set. This dataset contain the total 999 datas with 42 coloum features like 'duration' real 'protocol_type' {'tcp','udp', 'icmp'} 'service' 'flag' 'src_bytes' real 'dst_bytes' real 'land' {'0', '1'} 'wrong_fragment' real 'urgent' real 'hot' etc.



cm - NumPy array

	0	1
0	122	3
1	6	119

Figure 4: Confusion Matrix

Figure 4 is showing the confusion matrix. A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing.

TP: True Positive: Predicted values correctly predicted as actual positive

FP: Predicted values incorrectly predicted an actual positive. i.e., Negative values predicted as positive

FN: False Negative: Positive values predicted as negative

TN: True Negative: Predicted values correctly predicted as an actual negative

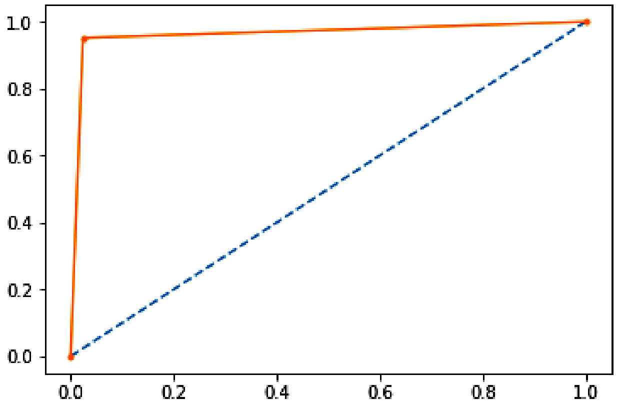


Figure 5: ROC

The figure 5 is showing the Receiver Operating Characteristic Curves (ROC) is a plot of signal (True Positive Rate) against noise (False Positive Rate).

Table 1: Simulation Result of DT with C4.5

Sr. No.	Parameters	Proposed Method (%)
1	Precision	97.6
2	Recall	95.3
3	F-measure	96.4
4	Accuracy	96.3
5	Error Rate	3.5

Table 2: Comparison of proposed work with previous work

Sr. No.	Parameters	Previous work [1]	Previous work [2]	Proposed Work
1	Methodology	LSTM	SMOTE	DT with C4.5
2	Precision (%)	63.76	85.82	97.6
3	Recall (%)	66.36	84.49	95.3
5	F-measure (%)	65.04	85.14	96.4
6	Accuracy (%)	91.63	83.58	96.3
7	Error Rate(%)	8	16	3.5

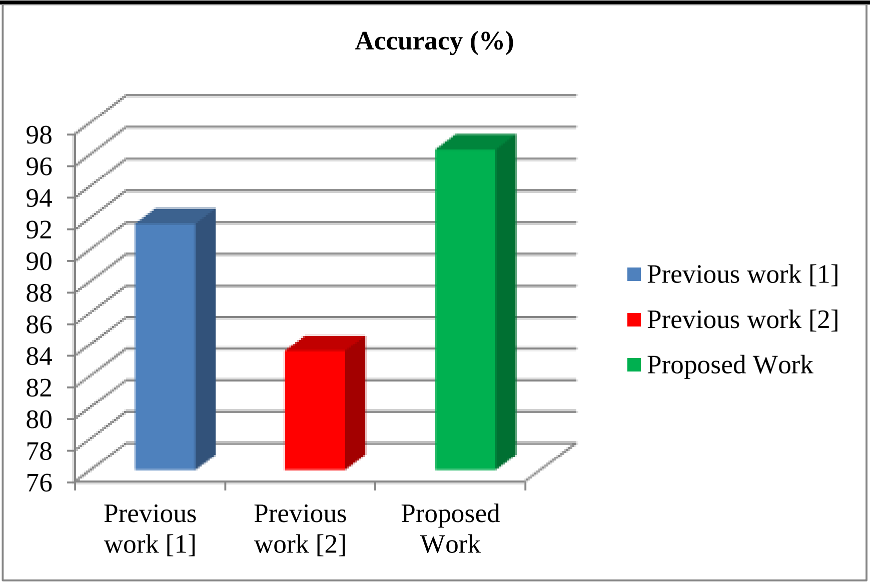


Figure 6: Accuracy comparison

Table 2 is showing the results comparison of the previous and proposed research works. It is clear from the previous and proposed work performance parameters result calculation, the proposed work is achieving significant better results than existing.

IV. CONCLUSION

This paper presents the C4.5 decision tree algorithm for classification. The C4.5 algorithm is used in Data Mining as a Decision Tree Classifier which can be employed to generate a decision, based on a certain sample of data. The dataset is taken from the KDD dataset kaggle. Precision value of existing results is 63.76 and 85.82% while proposed work achieved 97.6%. The recall value achieved by proposed technique is 95.3% while previous achieved is 66.36 and 84.49%. The f measure value is 96.4% and error rate is 3.5% by proposed technique while previous results is 65.04 and 85.14% of Fmeasure and 8 and 16% is error rate. Finally the accuracy achievement is 96.3% by the proposed methodology while previous accuracy value is 91.63 and 83.58%.

REFERENCES

1. H. Hou et al., "Hierarchical Long Short-Term Memory Network for Cyberattack Detection," in IEEE Access, vol. 8, pp. 90907-90913, 2020, doi: 10.1109/ACCESS.2020.2983953.
2. P. Feng, J. Ma, T. Li, X. Ma, N. Xi and D. Lu, "Android Malware Detection Based on Call Graph via Graph Neural Network," 2020 International Conference on Networking and Network Applications (NaNA), 2020, pp. 368-374, doi: 10.1109/NaNA51271.2020.00069.
3. S. Liu, M. Dibaei, Y. Tai, C. Chen, J. Zhang and Y. Xiang, "Cyber Vulnerability Intelligence for Internet of Things Binary," in IEEE Transactions on Industrial Informatics, vol. 16, no. 3, pp. 2154-2163, March 2020, doi: 10.1109/TII.2019.2942800.
4. Y. Jin, M. Tomoishi and N. Yamai, "Anomaly Detection by Monitoring Unintended DNS Traffic on Wireless Network," 2019 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM), 2019, pp. 1-6, doi: 10.1109/PACRIM47961.2019.8985052.
5. B. Peng, Q. Wang, X. Li, J. Cai, J. Fei and W. Chen, "Research on Abnormal Detection Technology of Real-Time Interaction Process in New Energy Network," 2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), 2019, pp. 433-440, doi: 10.1109/iThings/GreenCom/CPSCom/SmartData.2019.00092.
6. W. Bi, K. Zhang, Y. Li, K. Yuan and Y. Wang, "Detection Scheme Against Cyber-Physical Attacks on Load Frequency Control Based on Dynamic Characteristics Analysis," in IEEE Systems Journal, vol. 13, no. 3, pp. 2859-2868, Sept. 2019, doi: 10.1109/JSYST.2019.2911869.
7. K. Liu, Z. Fan, M. Liu and S. Zhang, "Hybrid Intrusion Detection Method Based on K-Means and CNN for Smart Home," 2018 IEEE 8th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER), 2018, pp. 312-317, doi: 10.1109/CYBER.2018.8688271.
8. Y. Jin, K. Kakoi, N. Yamai, N. Kitagawa and M. Tomoishi, "A Client Based Anomaly Traffic Detection and Blocking Mechanism by Monitoring DNS Name Resolution with User Alerting Feature," 2018 International Conference on Cyberworlds (CW), 2018, pp. 351-356, doi: 10.1109/CW.2018.00070.
9. R. Velea and Ş. Drăgan, "CPU/GPU Hybrid Detection for Malware Signatures," 2017 International Conference on Computer and Applications (ICCA), 2017, pp. 85-89, doi: 10.1109/COMAPP.2017.8079736.
10. S. Merat and W. Almuhtadi, "Artificial intelligence application for improving cyber-security acquirement," 2015 IEEE 28th Canadian Conference on Electrical and Computer Engineering (CCECE), 2015, pp. 1445-1450, doi: 10.1109/CCECE.2015.7129493.
11. S. Han, M. Xie, H. Chen and Y. Ling, "Intrusion Detection in Cyber-Physical Systems: Techniques and Challenges," in IEEE Systems Journal, vol. 8, no. 4, pp. 1052-1062, Dec. 2014, doi: 10.1109/JSYST.2013.2257594.
12. M. Bousaaid, T. Ayaou, K. Afdel and P. Estraillier, "Hand gesture detection and recognition in cyber presence interactive system for E-learning," 2014 International Conference on Multimedia Computing and Systems (ICMCS), 2014, pp. 444-447, doi: 10.1109/ICMCS.2014.6911197.