International Journal of Advanced Research and Multidisciplinary Trends (IJARMT)
An International Open Access, Peer-Reviewed Refereed Journal
Impact Factor: 6.4 Website: https://ijarmt.com ISSN No.: 3048-9458

Enhancing Classification Efficiency Using the J48 Decision Tree Algorithm

Name: Shaifali Prasad

Supervisor Name: Prof. Mohd. Arif

Department of Computer Science and Engineering

College Name: Rajshree Institute of Management & Technology, Bareilly (U.P)

Abstract

The J48 decision tree algorithm, derived from the C4.5 methodology, is a powerful and widely used tool for classification tasks due to its efficiency and interpretability. This algorithm employs a systematic approach to analyze datasets, beginning with preprocessing steps to address missing values and discretize continuous attributes when necessary. By leveraging Entropy to measure data uncertainty and Information Gain to evaluate attribute significance, J48 recursively splits datasets into subsets, creating decision nodes and leaf nodes for effective classification. The algorithm continues this process until all data is classified or specified stopping criteria are met, such as a minimum number of instances per leaf. To enhance model simplicity and prevent overfitting, J48 incorporates pruning techniques that replace less informative branches with leaf nodes, improving generalization. Its ability to handle mixed data types, work efficiently with large datasets, and generate interpretable decision trees makes J48 a versatile and robust tool for diverse classification applications. This paper discusses the methodology, advantages, and practical applications of the J48 algorithm in enhancing classification efficiency across various domains.

Introduction

Classification is a critical task in data analysis, enabling the categorization of data into predefined classes based on patterns and relationships within a dataset. Decision tree algorithms are widely utilized for their simplicity, interpretability, and efficiency in handling complex classification problems. Among these, the J48 algorithm, an open-source implementation of the C4.5 algorithm, has emerged as a robust tool for constructing decision trees that offer high accuracy and comprehensibility.

The J48 algorithm operates by recursively partitioning the dataset based on attributes that maximize Information Gain, a measure derived from Information Theory. This process begins with preprocessing the dataset to handle missing values and discretize continuous attributes

where necessary. At each step, the algorithm calculates Entropy to evaluate the uncertainty in the dataset and determines the attribute with the highest Information Gain as the splitting criterion. Subsets formed from this split are used to grow the tree, with decision nodes representing tests on attributes and leaf nodes corresponding to class labels.

A key feature of J48 is its ability to handle mixed attribute types (nominal and continuous) and datasets with missing values, ensuring versatility across diverse datasets. Additionally, J48 includes pruning techniques to simplify the resulting decision tree and reduce overfitting, enhancing the model's generalizability to unseen data. By replacing less significant branches with leaf nodes, pruning ensures a balance between model complexity and classification accuracy.

Research Methodology

The J48 Decision Tree Construction Algorithm

The J48 algorithm, an open-source implementation of the C4.5 decision tree algorithm, is a classification method that constructs decision trees by recursively splitting a dataset based on attributes to maximize information gain. The process begins with data preprocessing to handle missing values and continuous attributes. At each node, the algorithm calculates Entropy to measure uncertainty and uses Information Gain to select the best attribute for splitting. The dataset is divided into subsets based on attribute values, and decision nodes and leaf nodes are created accordingly. The process continues recursively until all data is classified or stopping criteria, such as minimum instances per leaf, are met. To prevent overfitting, pruning techniques simplify the tree by replacing less significant branches with leaf nodes. J48 is efficient, interpretable, and handles mixed data types, making it a versatile tool for classification tasks.

Proposed Method

The primary focus of this investigation is to reduce the input space of a dataset, minimize processing time, and enhance classification accuracy. To achieve these goals, the study employs a well-known measure from Information Theory—Entropy. Entropy measures the average uncertainty or disorder within a dataset, making it instrumental in identifying the key aspects of the data.

Once the critical elements are determined using Entropy, the correlation coefficient is applied to select significant attributes from the dataset. This step ensures that only the most relevant features are retained, reducing dimensionality and enhancing computational efficiency.

International Journal of Advanced Research and Multidisciplinary Trends (IJARMT)
An International Open Access, Peer-Reviewed Refereed Journal
Impact Factor: 6.4 Website: https://ijarmt.com ISSN No.: 3048-9458

Subsequently, the J48 algorithm is applied to these selected attributes to build an effective decision tree.

A detailed discussion on Entropy and three types of correlation coefficients follows in subsequent sections, alongside Fig. 1, which outlines the proposed methodology. This structured approach leverages Entropy for feature importance, correlation coefficients for attribute selection, and J48 for classification to achieve optimal results.

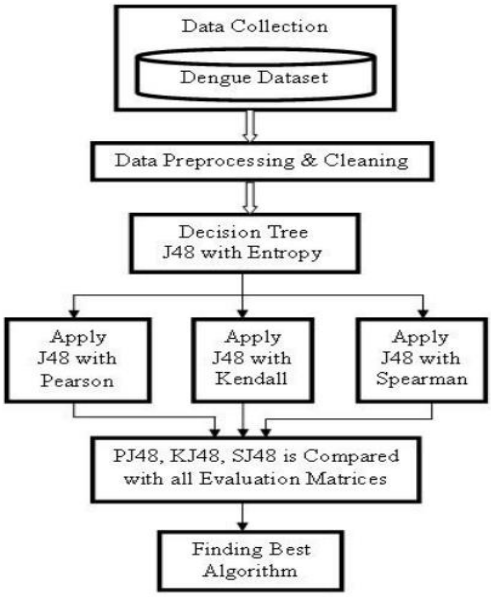

Fig. 1 Proposed Methodology

4.1.1 Model Training and Evaluation

For training and evaluation we used Java based environment using JDK, Visual Studio Code IDE and WEKA 3.8.8 lib. For the Intrusion Detection task, the J48 algorithm, which is based on the C4.5 decision tree framework, has been employed to classify attack types. This approach involves utilizing various hyperparameters to optimize the model's performance and enhance its predictive accuracy. By systematically varying these hyperparameters, we can effectively compare the performance of different J48 models, identifying the configurations that yield the best results for classifying attack types. This methodology not only strengthens the model's reliability but also contributes to more effective intrusion detection in cybersecurity applications.

Data Collection

International Journal of Advanced Research and Multidisciplinary Trends (IJARMT)
An International Open Access, Peer-Reviewed Refereed Journal
Impact Factor: 6.4 Website: https://ijarmt.com ISSN No.: 3048-9458

Since the work is related to IDS hence we are utilizing the Network packets data with Label variable recording type of DOS attack and other independent variable port, timestamp, flow, protocol, duration, etc. The Data has been collected form open source platforms and is CIC IDS 2018.

https://www.kaggle.com/datasets/solarmainframe/ids-intrusion-csv

	Dst Port	Protocol	Timestamp	Flow Duration	Tot Fwd Pkts	Tot Bwd Pkts	TotLen Fwd Pkts	TotLen Bwd Pkts	Fwd Pkt Len Max	Fwd Pkt Len Min	...	Fwd Seg Size Min	Active Mean	Active Std	Active Max	Active Min	Idle Mean	Idle Std	Idle Max	Idle Min	Label
0	0.0	0.0	b'14/02/2018 08:31:01'	112641719.0	3.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	56320859.5	139.300036	56320958.0	56320761.0	b'Benign'
1	0.0	0.0	b'14/02/2018 08:33:50'	112641466.0	3.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	56320733.0	114.551299	56320814.0	56320652.0	b'Benign'
2	0.0	0.0	b'14/02/2018 08:36:39'	112638623.0	3.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	56319311.5	301.934596	56319525.0	56319098.0	b'Benign'
3	22.0	6.0	b'14/02/2018 08:40:13'	6453966.0	15.0	10.0	1239.0	2273.0	744.0	0.0	...	32.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	b'Benign'
4	22.0	6.0	b'14/02/2018 08:40:23'	8804066.0	14.0	11.0	1143.0	2209.0	744.0	0.0	...	32.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	b'Benign'

Figure 2: Data Preview

Simulation Result

Preprocessing: For preprocessing we have used the python notebook environment.

- CSV to ARFF: The original data was in CSV format then we use WEKA to convert it into ARFF format.
- Consistency: Identify and remove duplicate and missing values.
- Class balancing: The imbalance in Label class being handled using SMOTE.
- Scaling: Data has been resampled and scaled using Standard Scaler.

EDA: Following graphs visualize the various insights of data:

International Journal of Advanced Research and Multidisciplinary Trends (IJARMT)
An International Open Access, Peer-Reviewed Refereed Journal
Impact Factor: 6.4 Website: https://ijarmt.com ISSN No.: 3048-9458

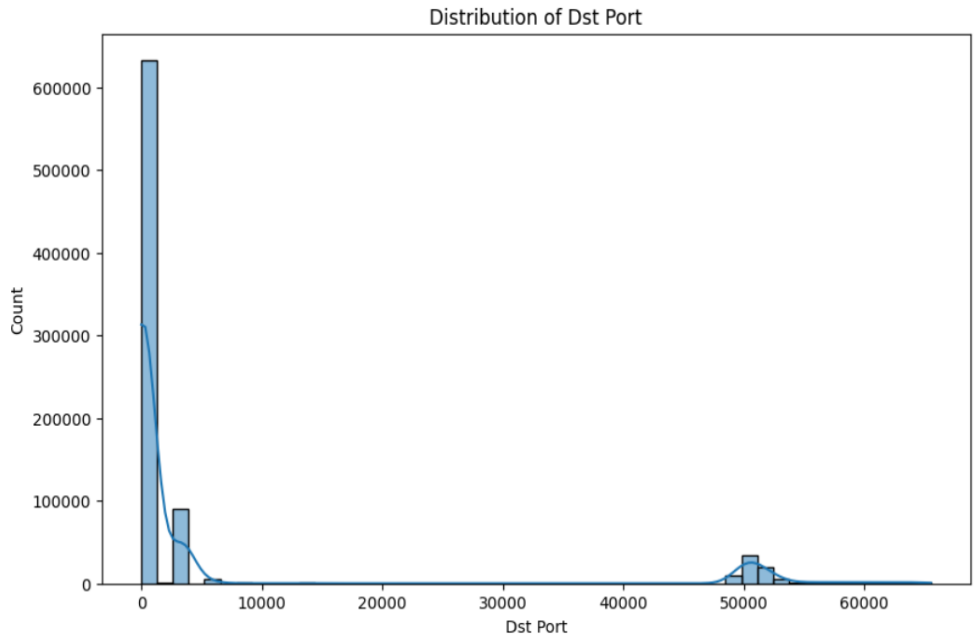

Figure 3: Distribution of Dst Port

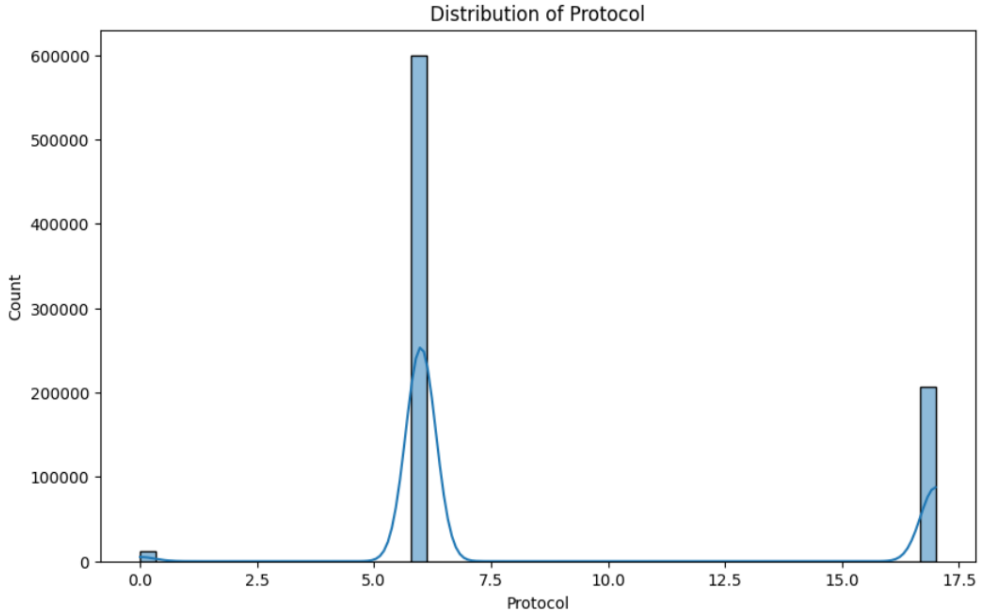

Figure 4: Distribution of Protocol

International Journal of Advanced Research and Multidisciplinary Trends (IJARMT)
An International Open Access, Peer-Reviewed Refereed Journal
Impact Factor: 6.4 Website: https://ijarmt.com ISSN No.: 3048-9458

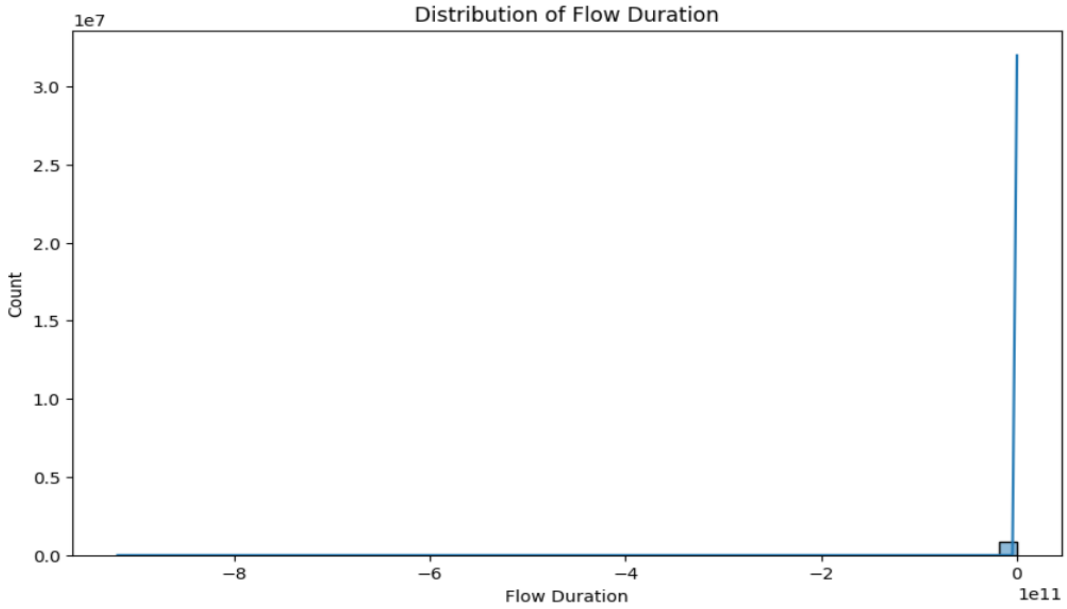

Figure 5: Distribution of Flow Duration

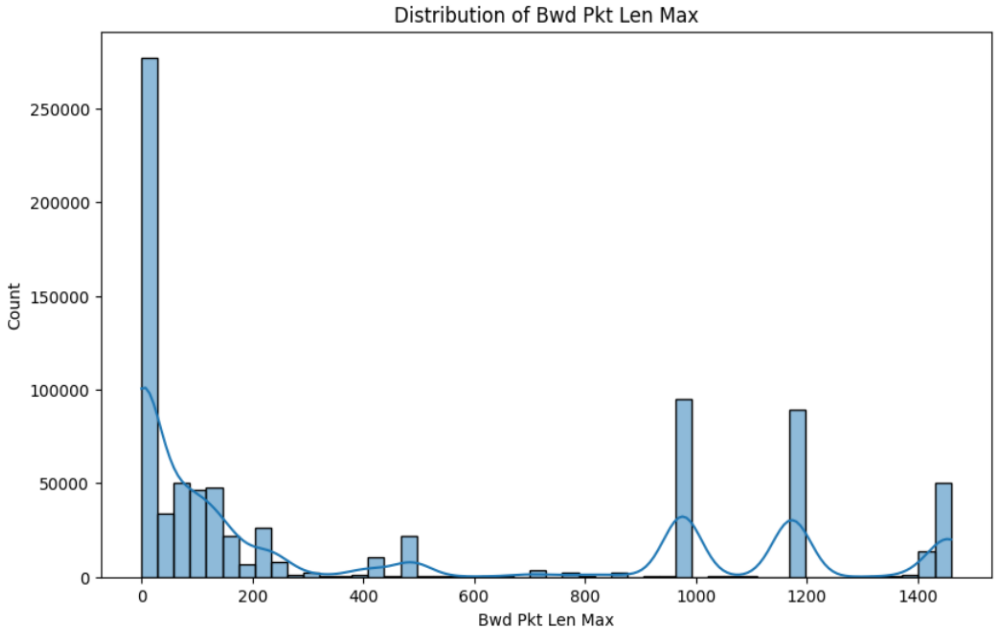

Figure 6: Packet Len Max

Table 1: Model 1: Default Parameters

Parameter	Value
Confidence Factor	0.25
Unpruned Tree	False
MinNumObj	2
Number of Leaves	50

International Journal of Advanced Research and Multidisciplinary Trends (IJARMT)
An International Open Access, Peer-Reviewed Refereed Journal
Impact Factor: 6.4 Website: https://ijarmt.com ISSN No.: 3048-9458

Tree Size	99
Accuracy	0.92

Results:

Correctly Classified Instances 1048543 99.9969 %

Incorrectly Classified Instances 32 0.0031 %

Kappa statistic 0.9999

Mean absolute error 0

Root mean squared error 0.0045

Relative absolute error 0.0115 %

Root relative squared error 1.075 %

Total Number of Instances 1048575

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Benign
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	FTP-BruteForce
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	SSH-Bruteforce
Weighted Avg.	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	

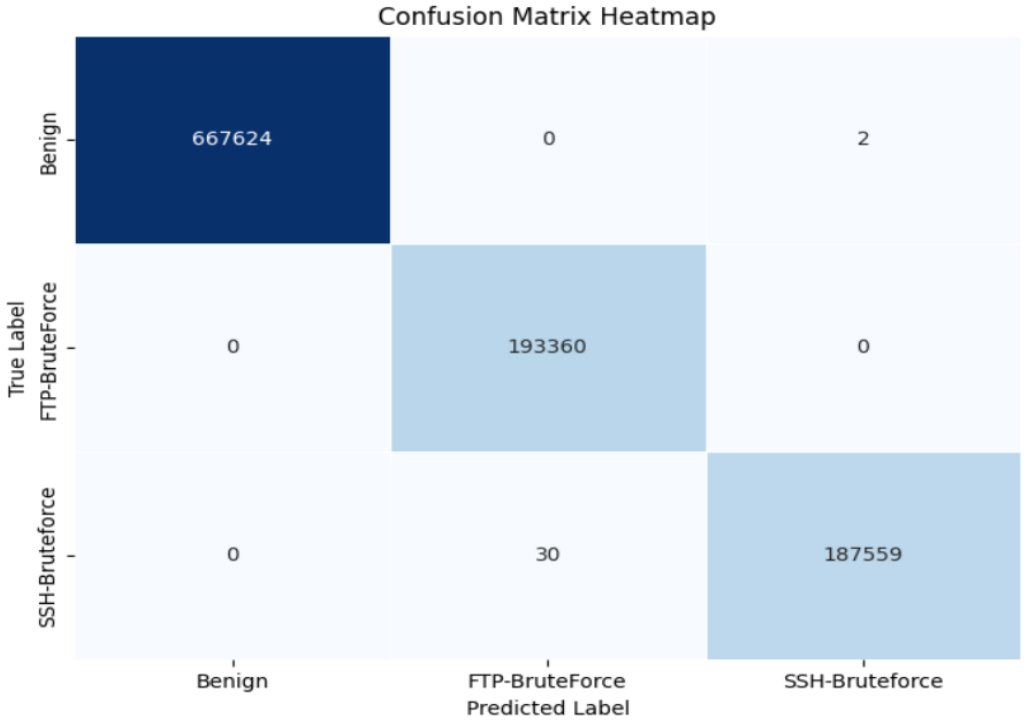

Figure 10: Confusion Matrix for Model

Conclusion

The J48 decision tree algorithm has proven to be an effective and efficient tool for enhancing classification accuracy across a variety of applications, including Intrusion Detection Systems (IDS). By leveraging Information Theory principles, such as Entropy and Information Gain, J48 systematically identifies significant attributes and constructs decision trees that are both interpretable and computationally efficient. Its ability to handle mixed attribute types, address missing values, and incorporate pruning techniques ensures a balance between model complexity and classification accuracy, making it suitable for dynamic and noisy datasets. In the context of IDS, J48 demonstrates significant potential in detecting anomalies, reducing false positives, and adapting to evolving cyber threats. Its scalability and adaptability make it a versatile choice for both small-scale and large-scale datasets. This study highlights the algorithm's methodological strengths, practical applications, and its role in improving classification efficiency while maintaining model interpretability. As cyber threats continue to grow in sophistication, the need for intelligent and adaptive solutions like J48 becomes increasingly vital. Future research could focus on integrating J48 with advanced machine learning frameworks or hybrid approaches to further enhance its performance and applicability in complex environments. The J48 algorithm thus remains a cornerstone in advancing classification techniques and addressing emerging challenges in data-driven domains.

References

[1] P. Porambage, A. Braeken, C. Schmitt, A. Gurtov, M. Ylianttila, and B. Stiller, "Group key establishment for enabling secure multicast communication in wireless sensor networks deployed for IoT applications," IEEE Access, vol. 3, pp. 1503–1511, 2015.

[2] N. Andreadou, M. O. Guardiola, and G. Fulli, "Telecommunication technologies for smart grid projects with focus on smart metering applications," Energies, vol. 9, no. 5, p. 375, 2016.

[3] S. Omar, A. Ngadi, and H. H Jebur, "Machine learning techniques for anomaly detection: an overview," International Journal of Computer Application, vol. 79, no. 2, pp. 33–41, 2013.

[4] Afreen Bhumgara and Anand Pitale, "Detection of Network Intrusion Using Hybrid Intelligent System", IEEE International Conferences on Advances in Information Technology, pp. no. 167-172, Chikmagalur, India 2019.

[5] Ritumbhira Uikey and Dr. Manari Cyanchandani "Survey on Classification Techniques Applied to Intrusion Detection System and its Comparative Analysis", IEEE 4th International Conference on Communication $ Electronics System (ICCES), pp. no. 459-466, Coimbatore, India 2019.

[6] Aditya Phadke, Mohit Kulkarni, Pranav Bhawalkar and Rashmi Bhattad "A Review of Machine Learning Methodologies for Network Intrusion Detection", IEEE 3rd National Conference on Computing Methodologies and Communication (ICCMC), pp. no. 703-709, Erode, India 2019.

[7] S. Sivantham, R.Abirami and R.Gowsalya "Comapring in Anomaly Based Intrusion Detection System for Networks", IEEE International conference on Vision towards Emerging Trends in Communication and Networking (ViTECon), pp. no. 289-293, Coimbatore, India 2019.

[8] Azar Abid Salih and Maiwan Bahjat Abdulrazaq "Combining Best Features selection Using Three Classifiers in Intrusion Detection System", IEEE International Conference on Advanced science and Engineering (ICOASE), pp. no. 453-459, Zakho - Duhok, Iraq 2019.

[9] Lukman Hakim and Rahilla Fatma Novriandi "Influence Analysis of Feature Selection to Network Intrusion Detection System Performance Using NSL-KDD Dataset", IEEE International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE), pp. no. 330-336, Jember, Indonesia 2019.

[10] T. Sree Kala and A. Christy, "An Intrusion Detection System Using Opposition Based Particle Swayam Optimization Algorithm and PNN", IEEE International Conference on Machine Learning, Big Data, Cloud and Parallel Computing, pp. no. 564-569, Coimbatore, India 2019.