

PERFORMANCE OF DECISION TREE ALGORITHM WITH RANDOMLY DISTRIBUTED FEATURE SPACE

Nurul Radziah binti Ajis, Muhamad Hasbullah bin Mohd Razali

Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Perlis Branch, Malaysia

e-mail: nurulraziawork01@gmail.com, hasbullah782@uitm.edu.my

Abstract

This study investigates the performance of well-known tree-based algorithms, namely the J48 and Random Forest (RF), on data sets with various feature space distributions. The feature space was simulated using logistic regression due to its simplicity in linearly separable scenarios mainly to reflect a bell-shaped (Normal), right (Gamma) and left (Beta) skewness of real data sets. Both algorithms were compared through accuracy. Simulation results indicate that with normally distributed feature space of balanced class data sets (Sim1, Sim2), J48 performed less accurately than RF. Meanwhile, both algorithms were found to be equally accurate in rather randomly distributed feature space with imbalanced class. This is expected since most algorithms were designed to detect the majority groups for high classification rate. Nevertheless, the simulation procedure highlights the importance of investigating the feature space structure in model performance particularly in unstructured or noisy environments due to reliance on feature space conditions. Future works would be to consider the overall structure of the data sets including class distributions and number of splits as decision trees rely heavily on the branches for classification tasks.

Keywords: *Decision Tree, Logistic Regression, Classification Accuracy, Feature Space Distribution*

1. Introduction

This study evaluates the performance of the Decision Tree algorithm on datasets with randomly distributed feature spaces, using Logistic Regression as a baseline. While Decision Trees are valued for their interpretability and ability to handle mixed data types, their performance can decline with unstructured or noisy data. In contrast, Logistic Regression remains robust in linearly separable cases. The study aims to analyse how random feature distributions impact classification performance and focuses on accuracy, precision, recall, and F-measure. It also explores the challenges random distributions pose to Decision Trees and compares their robustness against Logistic Regression in such conditions.

2. Methodology

This study uses a simulation-based approach with synthetic datasets generated from normal, beta, and gamma distributions to represent different random feature spaces. Logistic Regression establishes the ground truth for classification. The models tested include Decision Trees (J48), Random Forests, and Logistic Regression as a baseline, all implemented using the WEKA software. The models are evaluated using accuracy, precision, recall, and F-measure. A comparative analysis is conducted to assess the consistency and robustness of each model under varying random feature conditions, allowing a systematic investigation of how randomness in feature distributions affects classification performance.

3. Results / Findings

The simulation results show that Decision Trees, particularly the J48 algorithm, struggle with highly randomized feature spaces, resulting in lower accuracy and unstable outcomes. Random Forests consistently outperform J48, offering higher accuracy and better adaptability to noisy or unstructured data. Logistic Regression maintains stable performance in linearly separable cases but declines in

complex, non-linear feature spaces. Overall, Random Forests demonstrate the best performance among the tested models, followed by Decision Trees and Logistic Regression. Both Decision Trees and Random Forests improve with more data, though Random Forests achieve stable, reliable performance more quickly.

4. Novelty of Research / Product / Project

This research provides new insights into how random feature distributions impact Decision Tree performance, an area rarely explored. The project's key novelty lies in systematically simulating randomized feature spaces using various probability distributions to examine their effects on classification. It also offers a comprehensive product by comparing tree-based models with Logistic Regression under these conditions. The analysis identifies Decision Trees' weaknesses in unstructured feature spaces and highlights Random Forests' superior robustness. This study delivers practical guidance for data scientists in selecting suitable algorithms for noisy or random datasets, improving understanding of model behaviour in challenging data environments.

5. Conclusion or Future Work

This study finds Random Forest to be the most reliable classifier for randomly distributed feature spaces, offering high accuracy and consistent performance. Decision Trees, though interpretable, show instability and lower accuracy, while Logistic Regression performs well in simple, linearly separable cases but struggles with complex feature interactions.

REFERENCE LIST

- Quinlan, J. R. (2014). *C4.5: Programs for Machine Learning*. Elsevier.
- Muslim, M. A., Nurzahputra, A., & Prasetyo, B. (2018). Improving accuracy of C4.5 algorithm using split feature reduction model and bagging ensemble for credit card risk prediction. 2018 International Conference on Information and Communications Technology (ICOIACT), 4, 141–145. <https://doi.org/10.1109/icoiact.2018.8350753>
- Hui, J. (2022, March 19). Probability distributions in machine learning & deep learning. Medium. <https://jonathan-hui.medium.com/probability-distributions-in-machine-learning-deep-learning-b0203de88bdf>
- Sharma, A. (2024, December 12). 4 Simple ways to split a decision tree in machine learning (Updated 2025). Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2020/06/4-ways-split-decision-tree/#h-what-is-node-splitting-in-a-decision-tree-and-why-is-it-done>
- SkillCamper. (2025, April 9). Random Forest: Why ensemble learning outperforms individual models. SkillCamper. <https://www.skillcamper.com/blog/random-forest-why-ensemble-learning-outperforms-individual-models>