Evaluating the Impact of Pruning on Decision Tree Performance Across Classification and Regression Tasks

Srikanta Mehta SMEHTA44@JH.EDU
Department of Computer Science
Johns Hopkins University
Baltimore, MD 21205, USA

Abstract

This study explores the performance of decision trees before and after pruning across a range of machine learning tasks, focusing on classification and regression. By employing datasets from the UCI Machine Learning Repository, we examine the performance of these models under a variety of conditions, utilizing metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), R-squared (R2), and Pearson score for regression, as well as zero-one loss, precision, recall, and F1 score for classification. Our experimental results reveal insights into the impact of pruning on decision tree performance. Specifically, we observe that Pruned DT often outperforms standard DT in terms of accuracy. However, the degree of performance enhancement varies across datasets, demonstrating the influence of dataset characteristics such as size, feature diversity, and noise levels. For both regression and classification tasks, the benefits of pruning are variable, with certain datasets showing substantial improvements in MSE and zero-one loss scores, while others exhibit negligible differences or even slight deterioration post-pruning. Our findings support the hypothesis that while pruning can enhance model performance, its effectiveness is dependent upon specific dataset attributes and the nature of the task at hand. The study concludes with a discussion of the implications of these findings for model selection in machine learning applications and suggests avenues for future research, including the exploration of advanced techniques.

Keywords: Decision Trees, Reduced Error Pruning, Classification, Regression

1. Introduction

Among the diverse selection of machine learning models, decision trees (DT) stand out for their interpretability and straightforward decision-making process. Stemming from the foundational work by Ross Quinlan in the 1980s on the ID3 (Iterative Dichotomiser 3) algorithm (Quinlan, 1986), decision trees have undergone several modifications aimed at improving their predictive accuracy and computational efficiency.

One significant advancement in this area is the introduction of pruning strategies, designed to address the overfitting issue by trimming down the decision tree structure. Quinlan explored these advancements in the development of the C4.5 algorithm (Quinlan, 1993). Pruned decision trees (Pruned DT) are intended to provide better generalization capabilities compared to their full-grown counterparts by eliminating branches that contribute minimally to the model's predictive power.

This investigation is focused on the hypothesis that Pruned DT will generally exhibit superior performance over standard decision trees, though this improvement may be affected by specific factors such as dataset complexity, the diversity of features, and the presence of noise. This hypothesis suggests that there are further complexities involved in pruning

strategies and decision tree performance, indicating that the advantages of Pruned DT might not uniformly apply across all classification and regression scenarios.

To evaluate this hypothesis, the structure of this report is as follows: Section 2, *Algorithms and Experimental Methods* outlines the decision tree framework, pruning techniques, and the experimental setup used for performance assessment. Section 3, *Results,* offers a detailed analysis comparing standard and Pruned DT across a variety of datasets, utilizing measures such as Mean Squared Error (MSE), and accuracy metrics. In Section 4, *Discussion,* the implications of the findings are explored, particularly how dataset attributes and other considerations influence the differential performance between standard and Pruned DT. The report concludes in Section 5 with a conclusion and propositions for further research efforts.

2. Algorithms and Experimental Methods

The objective of this project was to evaluate the performance of decision trees and their pruned versions for both classification and regression tasks. We developed an implementation that allows for flexibility in handling different types of data and adapting to various experimental settings based on the dataset characteristics. This section will offer a concise review of the datasets selected from the UCI Machine Learning Repository, outline the specifics of the decision tree construction and pruning techniques, and describe the experimental approach employed to investigate our research hypothesis.

2.1 Data Sets and Preprocessing

Our study leveraged six datasets from the UCI Machine Learning Repository, divided equally for classification and regression tasks, to assess decision trees and their pruned variants. Minimal preprocessing was applied, focusing on missing value imputation and the removal of unique identifiers, leveraging the inherent flexibility of decision trees towards numerical and categorical data.

2.1.1 Abalone Dataset

A regression dataset aiming to predict abalone age from physical measurements, the Abalone dataset comprises 4177 instances with 8 attributes, predominantly continuous, except for the nominal Sex attribute.

2.1.2 Forest Fires Dataset

This regression dataset, consisting of 517 instances and 13 attributes, predicts forest fire areas from meteorological data. It presents a regression challenge, especially due to data skew towards smaller fires.

2.1.3 Computer Hardware Dataset

Featuring 209 instances and 10 attributes, this dataset aims at modeling relative CPU performance from hardware specifications. Attributes include various hardware specifications, with Vendor Name and Model attributes dropped for non-predictiveness.

2.1.4 Congressional Vote Dataset

A classification dataset based on the 1984 US Congressional Voting Records. It includes 435 instances with 16 votes each, aimed at predicting party affiliation (Democrat or Republican) from voting patterns. Missing votes were treated as abstentions.

2.1.5 Car Evaluation Dataset

This classification dataset assesses car acceptability over 1728 instances based on six attributes, categorizing cars into four acceptability classes. It evaluates cars on criteria such as price, maintenance cost, and safety, among others.

2.1.6 Breast Cancer Dataset

A classification dataset with 699 instances, each with 10 attributes. it predicts breast cancer malignancy from diagnostic imaging. Attributes cover various characteristics of cell nuclei from breast mass biopsies, classifying tumors into benign or malignant.

2.2 Algorithms

2.2.1 Building Decision Trees

Decision Trees (DT) serve as foundational models in machine learning, suitable for addressing both classification and regression challenges. By learning simple decision rules inferred from the data features, these models recursively partition the dataset. This results in a hierarchical structure where internal nodes represent decision points based on features, and leaf nodes signify predicted outcomes.

Node Representation A node within a decision tree is described by:

- **Feature** (f_i): The attribute that splits the data at the node.
- **Threshold** (θ): For numerical features, the value that divides the data into different branches.
- **Value** ($\hat{y}$): In a leaf node, this is either the predicted class label (classification) or the continuous value (regression).
- **Children**: This is a dictionary linking outcomes of the split to the corresponding child nodes.

Splitting Criterion Determining the optimal point for data division relies on:

- **Classification**: The gain ratio is used to select features, it is calculated as:

$$gainRatio(f_i) = \frac{gain(f_i)}{IV(f_i)}.$$

Here, $gain(f_i)$ represents the information gain from dividing the dataset based on feature f_i, expressing the entropy reduction achieved by the split. It is calculated as:

$$gain(f_i) = H(D) - \sum_{j=1}^{m_i} \frac{|D_j|}{|D|} H(D_j),$$

where $H(D)$ is the entropy of the dataset before the split, m_i is the count of unique outcomes for feature f_i, $|D_j|$ is the size of subset j post-split, and $H(D_j)$ is the entropy of subset j. The intrinsic value ($IV(f_i)$), is defined as:

$$IV(f_i) = -\sum_{j=1}^{m_i} \frac{|D_j|}{|D|} \log_2 \left(\frac{|D_j|}{|D|} \right).$$

- **Regression**: Focuses on minimizing the Mean Squared Error (MSE), a measure of variance within the node:

$$MSE = \frac{1}{|D|} \sum_{i=1}^{|D|} (y_i - \hat{y})^2,$$

where $|D|$ represents the number of samples in the node, y_i is the actual target value for sample i, and $\hat{y}$ is the mean target value for samples within the node.

Recursive Tree Construction The building of a decision tree from training data is a recursive process that iteratively refines the decision-making path from the root to the leaf nodes. This procedure has the following steps:

1. **Optimize Split:** Evaluate each feature f_i to find the split (feature and threshold) that maximizes information gain for classification or minimizes mean squared error (MSE) for regression.
2. **Partition Dataset:** Split the dataset into subsets based on the chosen feature's value(s), with binary splits for numerical features and multi-way splits for categorical features.
3. **Recursive Splitting:** Apply steps 1 and 2 recursively to each subset, stopping when a predefined criterion like maximum depth or minimum subset size is met.
4. **Determine Leaf Value:** For leaf nodes, assign a prediction value—mode of target values for classification or mean for regression.

This approach ensures that each decision point in the tree is data-driven. By doing so, the decision tree becomes capable of capturing complex relationships between features and the target variable, while maintaining interpretability through its hierarchical decision structure.

2.2.2 Pruning Decision Trees

Pruning is a process in the development of decision trees aimed at reducing the complexity of the model and improving its generalization capability. The purpose of pruning is to remove parts of the tree that do not contribute significantly to the model's predictive power, reducing overfitting and improving the model's performance on unseen data.

Pruning Methodology The pruning process involves evaluating each node of the tree to decide whether converting it into a leaf node would result in a lower error on a validation dataset not used during the training phase. This evaluation is performed recursively from the leaves up to the root of the tree, ensuring that each decision to prune is based on the most current structure of the tree.

Error Evaluation for Pruning The decision to prune a node is based on a comparison of the error before and after the pruning operation. The error is calculated using mean squared error (MSE) for regression tasks and zero-one loss for classification tasks.

Pruning Steps

1. Starting at the leaf nodes, evaluate the potential error reduction by converting each node into a leaf with the most common class (for classification) or the mean target value (for regression) of the samples reaching that node.

2. Calculate the error on the validation dataset both before and after simulating the pruning of the node. The simulation involves temporarily treating the node as a leaf.

3. If the error post-pruning is less than or equal to the error pre-pruning, the node is permanently converted into a leaf. Otherwise, the node retains its structure.

4. Repeat this process recursively for each node moving upward until the root node is evaluated.

This pruning strategy ensures that only splits that improve performance are retained, simplifying the model and enhancing its ability to generalize from the training data to new, unseen datasets.

2.3 Prediction Methodology

The prediction process in decision trees involves traversing from the root node to a leaf node based on the attributes of the test instance. This traversal determines the outcome of the prediction, whether it is a class label in classification problems or a continuous value in regression problems. Below is an overview of the steps involved:

1. **Start at the Root**: The prediction process begins at the root node of the decision tree.

2. **Traverse the Tree**: At each node, the value of the feature that the node represents is compared against the corresponding value in the test instance.

    - For **numerical features**, if the test instance's feature value is less than or equal to the node's threshold, the traversal moves to the left child of the node; otherwise, it moves to the right child.
    - For **categorical features**, the traversal follows the path corresponding to the feature's value in the test instance.

3. **Reach a Leaf Node**: The process continues until it reaches a leaf node.

4. **Make a Prediction**: The value associated with the leaf node is returned as the prediction. This value is the most common class among the training instances that fall into this leaf (for classification tasks) or the average of the target values of these instances (for regression tasks).

2.4 Experiment Design

The experiment is structured to measure the performance of decision trees before and after pruning, as compared to a null model, under the hypothesis that pruning enhances model performance by reducing overfitting. The experimental procedure is outlined as follows:

1. The dataset is partitioned into an 80% training set and a 20% validation set. The validation set is reserved exclusively for pruning the decision tree.

2. We use a 5x2 cross-validation strategy for model evaluation, which includes:

   (a) Dividing the 80% training portion into two halves for each fold. Stratification is applied in classification tasks to preserve class distribution.

   (b) Training a decision tree on one half and testing it on the other half and vice versa, to complete one iteration of cross-validation.

   (c) Utilizing the 20% validation set to prune the decision tree.

3. Performance metrics are collected for each fold and averaged across all runs to calculate the model's performance. These include mean squared error (MSE), mean absolute error (MAE), r-squared (R2), and Pearson's correlation for regression tasks. For classification tasks, we measure 0-1 Loss, precision, recall, and f1 score.

This design allows for a detailed comparison between the unpruned and pruned decision trees that takes into account variability induced by random splitting by averaging performance.

3. Results

This section details the outcomes from the experiments conducted on six distinct datasets retrieved from the UCI Machine Learning Repository. The objective was to assess and compare the performance of unpruned and pruned decision trees across these datasets. Table 1 showcases the performance metrics for the regression tasks. For these analyses, both the unpruned and pruned decision tree models were optimized based on the data characteristics, with particular attention to minimizing the mean squared error (MSE) for regression outcomes. Table 2 presents the results for classification tasks. Similar to the regression tasks, the decision trees were evaluated for their accuracy, precision, recall, and F1 score, with and without the application of pruning techniques. Figure 1 illustrates the percentage improvement in performance when comparing pruned decision trees against their unpruned counterparts across both sets

of tasks. The performance measure used to calculate improvement is MSE for regression tasks and zero-one Loss for classification.

Table 1: Regression Task Performance Comparison

Dataset	**Algorithm**	**MSE**	**MAE**	**R2 Score**	**Pearson Score**
	Null Model	19073.87	86.17	N/A	N/A
	DT	7077.10	38.50	0.6412	0.8206
Machine Data	Pruned DT	7617.83	42.28	0.6128	0.8161
	Null Model	10.281	2.358	N/A	N/A
	DT	8.725	2.089	0.1510	0.5754
Abalone Data	Pruned DT	7.710	1.860	0.2510	0.5748
	Null Model	2161.23	11.25	N/A	N/A
	DT	4954.86	21.56	-3.5144	-0.0102
Forest Fire Data	Pruned DT	3100.31	14.34	-1.6187	N/A

Table 2: Classification Task Performance Comparison

Dataset	**Algorithm**	**0-1 Loss**	**Precision**	**Recall**	**F1 Score**
	Null Model	0.2945	0.4977	0.7055	0.5837
Car Data	DT	0.0957	0.9071	0.9043	0.9047
	Pruned DT	0.1221	0.8897	0.8779	0.8780
	Null Model	0.3937	0.3676	0.6063	0.4577
House Votes Data	DT	0.0701	0.9308	0.9299	0.9300
	Pruned DT	0.0460	0.9551	0.9540	0.9542
	Null Model	0.3506	0.4217	0.6494	0.5113
Breast Cancer Data	DT	0.0669	0.9342	0.9331	0.9329
	Pruned DT	0.0554	0.9456	0.9446	0.9448

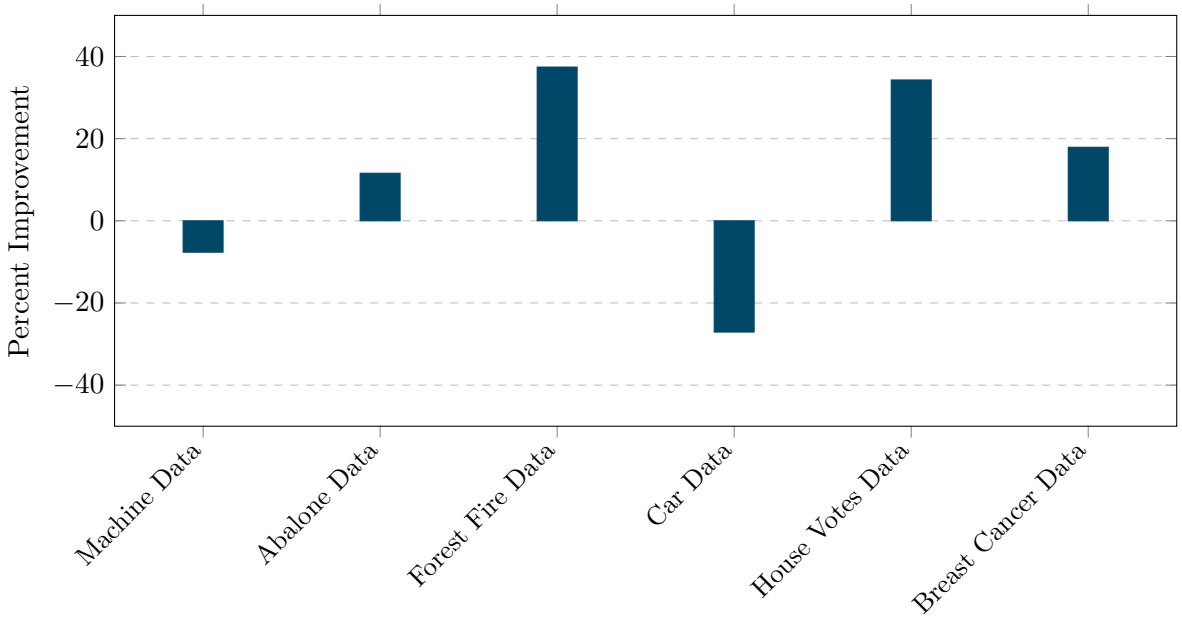

Figure 1: Percent improvement from Decision Tree to Pruned Decision Tree across datasets.

4. Discussion

Our analysis of decision trees (DT) and their pruned versions (Pruned DT) across a variety of datasets for classification and regression tasks has provided an interesting set of results. Compared to null models, decision trees demonstrated superior predictive accuracy, as evidenced by significant reductions in mean squared error (MSE) for regression tasks, and zero-one loss for classification tasks.

Pruning showed a diverse impact across datasets as shown in Figure 1. In the Machine dataset, pruning led to a slight decrease in performance. This was not unexpected as the Machine dataset was the smallest with only 209 instances. In the cross-validation process, the training set used to build the decision tree only contained 83 instances, significantly limiting the model's learning capacity. Such a small sample size for training makes it challenging for the decision tree to capture the full complexity of the dataset, and pruning may exacerbate this issue by further reducing the tree's depth and complexity.

The Forest Fire dataset presented a unique challenge, where both DT and Pruned DT models underachieved compared to the null model as seen in Table 1. This highlights the potential limitations of decision trees in dealing with datasets characterized by skewed distributions or significant outliers, which can adversely affect model accuracy. Despite this, pruning the DT was still able to offer almost a 40% improvement in MSE, demonstrating the value of pruning in reducing the model's complexity to mitigate overfitting. The improvement, although not enough to surpass the null model's performance, indicates that pruning can enhance the ability of decision trees to perform on datasets with challenging characteristics.

The classification results were also interesting. Decision trees significantly outperformed null models across all evaluation metrics as shown in Table 2. The application of pruning further improved model performance, especially in the House Votes and Breast Cancer datasets, where improvements in accuracy, precision, recall, and F1 score were observed. Interestingly, the Car dataset saw the largest drop in performance post-pruning across all 6 datasets. The Car dataset, with its categorical attributes and multi-class classification labels, might have been more sensitive to reductions in tree depth, leading to over-simplification of the model.

In conclusion, our findings support the hypothesis that pruning is a strong tool for improving model performance by addressing overfitting. Nonetheless, the effectiveness of pruning depends on the dataset's unique features, necessitating a thoughtful approach to applying this technique for optimal results.

5. Conclusion

This study has highlighted the capabilities of decision trees (DT) and their pruned counterparts (Pruned DT) in addressing both classification and regression challenges. The experiment design, focused on a comprehensive analysis of six distinct datasets from the UCI Machine Learning Repository, provided a varied collection of contexts to assess the model's adaptability and performance.

Looking forward, there is a lot of opportunity for further exploration into decision tree optimization and application. Future work could explore alternative splitting criteria or methods beyond binary splits for numerical features. For example, using quintiles, mean and standard deviation, or equal-with binning to manage splits. It would also be interesting to experiment with other techniques for reducing overfitting like the integration of decision trees into ensemble methods like Random Forests and Gradient Boosting Machines or early stopping criteria. In a real-world application, it is also important to optimize the data-processing steps for the specific dataset in consideration. This could include data augmentation techniques like SMOTE or feature selection.

In conclusion, this study confirms the value of decision trees and pruned decision trees as versatile and powerful tools in the machine learning toolkit. Their ability to provide interpretable models, coupled with the potential for enhanced performance through pruning, makes them a strong choice for a wide range of tasks.

References

J. Ross Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986. URL https://api.semanticscholar.org/CorpusID:189902138.

J.R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann series in machine learning. Elsevier Science, 1993. ISBN 9781558602380. URL https://books.google.com/books?id=HExncpjbYroC.