Rethinking Uncertainty Measures: Replacing Entropy with Extropy in Learning Algorithms

S. Abolhosseini, M. Khorashadizadeh and M. Chahkandi
Dept. of Statistics, University of Birjand

Abstract: This paper investigates replacing entropy with extropy as the impurity measure in decision tree algorithms. Decision trees are widely used and interpretable machine learning models that recursively partition data based on features to create hierarchical decision rules. Impurity measures such as entropy and the Gini index play a key role in selecting the best feature and split point at each node. Extropy, a newer measure focusing on the probability of error rather than uncertainty, has been proposed as an alternative to entropy.

Using the Iris dataset, the performance of decision trees constructed with entropy, extropy, and the Gini index was compared. Results showed that the entropy-based tree achieved the highest accuracy of approximately 97.78%, while extropy- and Gini-based trees both attained similar accuracies around 93.33%. The performance difference between entropy and Gini is generally small, and due to its computational efficiency, the Gini index is often preferred in practical applications. Extropy, with its focus on error probability, demonstrated competitive performance close to Gini and can serve as a suitable alternative in scenarios where error-focused impurity assessment is desirable.

Keywords: Decision tree, Extropy, Entropy, Gini Index.

1- Introduction

In recent years, machine learning algorithms have attracted significant attention due to their ability to analyze complex data and extract meaningful patterns. Among these algorithms, decision trees stand out as one of the most widely used and interpretable models. A decision tree is a hierarchical structure that recursively partitions data into subsets based on feature values, resulting in simple and understandable decision rules. Due to their interpretability, computational efficiency, and broad applicability in classification and regression tasks, decision trees have been extensively employed in various fields such as business analytics, medicine, and social sciences.

Impurity measures play a crucial role in the construction of decision trees, as they guide the selection of the best feature and split point at each node. The most commonly used impurity measures include **Entropy** and the **Gini Index**, both of which quantify the degree of disorder or uncertainty in the data. Information Gain, based on entropy, has been the primary splitting criterion in classical algorithms like ID3 [3].

However, a relatively less explored impurity measure called **Extropy** [2] has recently been proposed. Extropy focuses on the probability of error and can serve as an alternative splitting criterion. Given the structural similarities between extropy and the Gini index, as well as the differences with entropy, a thorough comparison of these three measures can provide deeper insights and potentially improve decision tree performance.

The objective of this paper is to replace the traditional information gain with extropy gain in decision tree algorithms and to conduct a comprehensive comparison among extropy, entropy, and the Gini index. Using the Iris dataset, we evaluate the impact of these impurity measures on tree construction, classification accuracy, and other performance metrics.

2. Literature Review and Theoretical Background

Decision trees are one of the most popular supervised machine learning algorithms used for classification and regression tasks. The model is structured as a hierarchical tree consisting of a root node, internal nodes (decision nodes), branches, and leaf nodes (terminal nodes). Each internal node represents a test on a feature, branches correspond to the outcome of the test, and leaf nodes represent class labels or predicted values.

A key objective in decision tree construction is to create subsets of data that are as homogeneous as possible with respect to the target variable. This homogeneity ensures clear decision boundaries and improves classification accuracy. To achieve this, impurity measures are used to evaluate how well a feature splits the data at each node

2.1 Impurity Measures

The most commonly used impurity measures include:

- **Entropy:** Measures the level of uncertainty or disorder in the data. It is defined as

$$Entropy = -\sum_{i=1}^{n} P_i \log P_i$$

where P_i is the proportion of samples belonging to class i. Lower entropy indicates more homogeneous subsets.

- **Gini Index:** Measures the probability of incorrectly classifying a randomly chosen element if it was labeled according to the class distribution in the subset. It is calculated as

$$Gini = 1 - \sum_{j} P_j^2$$

Values closer to 0 indicate purer nodes [1].

- **Extropy:** A less explored impurity measure defined as

$$Extropy = -\sum_{i=1}^{n} (1 - P_i) \log(1 - P_i)$$

which focuses on the probability of error and offers an alternative perspective on impurity.

2.2 Feature Selection and Splitting Criteria

At each node, the decision tree algorithm selects the feature and split point that maximize the reduction in impurity. For entropy, this reduction is quantified by **Information Gain**, which measures the decrease in entropy after the split. Similarly, for Gini and Extropy, analogous gain measures can be defined to guide the splitting process.

2.3 Related Work

Classical algorithms such as ID3 and C4.5 [2] primarily use entropy-based information gain for splitting, while CART uses the Gini index.

3. Methodology

In this study, we propose a novel approach to decision tree construction by replacing the traditional information gain based on entropy with a new impurity measure called **extropy**. The goal is to evaluate whether extropy can serve as a more effective splitting criterion compared to entropy and the widely-used Gini index.

3.1 Extropy as an Impurity Measure

Extropy is defined as follows for a node containing samples from C classes, where P_i denotes the proportion of samples belonging to class i:

$$Extropy = -\sum_{i=1}^{n}(1 - P_i)\log(1 - P_i)$$

Unlike entropy, which measures uncertainty, extropy focuses on the probability of error, providing an alternative perspective on impurity.

3.2 Extropy Gain

Similar to information gain in entropy-based trees, we define **extropy gain** as the reduction in extropy after splitting a node based on a feature and a threshold. Given a dataset D, splitting it into subsets D_1 and D_2, the extropy gain is computed as:

$$Gain_{extropy} = Extropy(D) - \frac{|D_1|}{|D|}Extropy(D_1) - \frac{|D_2|}{|D|}Extropy(D_2)$$

The feature and split point that maximize Gainextropy are selected at each node.

3.3 Decision Tree Construction

The decision tree is constructed recursively as follows:

1. At each node, calculate the extropy gain for all candidate features and their possible split points.
2. Select the feature and split point with the highest extropy gain.
3. Partition the data into left and right subsets based on the selected split.

4. Repeat the process recursively for each child node until a stopping criterion is met, such as:
    - Maximum tree depth
    - Minimum number of samples in a node
    - No further gain in extropy
5. Assign the majority class label to leaf nodes.

3.4 Dataset and Experimental Setup

We use the well-known Iris dataset [4] for evaluation. The dataset contains 150 samples with 4 features each, classified into 3 classes.

The dataset is randomly split into training (70%) and testing (30%) subsets. The decision tree models are trained separately using extropy, entropy (information gain), and Gini index as splitting criteria.

3.5 Performance Metrics

To evaluate and compare the models, the following metrics are used:

- **Accuracy:** The proportion of correctly classified samples.
- **Precision:** The ratio of true positives to the sum of true and false positives.
- **Recall:** The ratio of true positives to the sum of true positives and false negatives.
- **F1-score:** The harmonic mean of precision and recall.

Additionally, confusion matrices and graphical representations of the decision trees are generated for qualitative analysis [1].

4. Experimental Results

In this section, we present the experimental evaluation and performance comparison of decision tree classifiers constructed using three different impurity measures: Extropy, Entropy, and the Gini Index. The experiments were conducted on the Iris dataset, a widely used benchmark in machine learning, and the models were evaluated using four key metrics: Accuracy, Precision, Recall, and F1-score.

4.1 Performance Metrics Comparison

Metric	Extropy-Based Tree	Entropy-Based Tree	Gini-Based Tree
Accuracy	0.9333	0.9778	0.9333
Precision	0.9184	0.9794	0.9444
Recall	0.9167	0.9778	0.9333
F1-score	0.9165	0.9777	0.9327

The entropy-based decision tree achieved the highest performance across all metrics, with an accuracy of approximately 97.78%. Both extropy- and gini-based trees showed similar results, with accuracy around 93.33%. Precision and recall values followed the same trend, indicating that entropy slightly outperforms the other two criteria in this particular dataset.

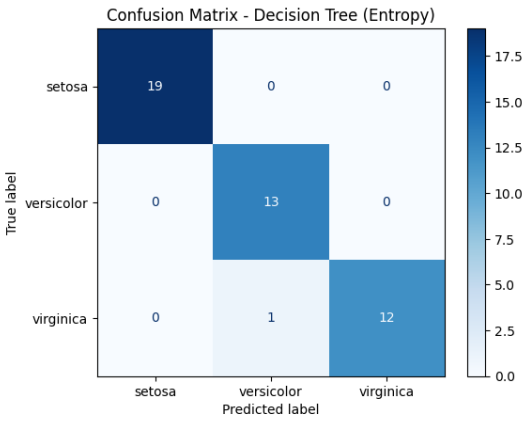
4.2 Discussion

These results align with findings in the literature where entropy often provides marginally better classification performance due to its comprehensive measure of uncertainty and information gain. However, the difference between entropy and gini is generally small, and gini is preferred in many practical applications due to its computational efficiency

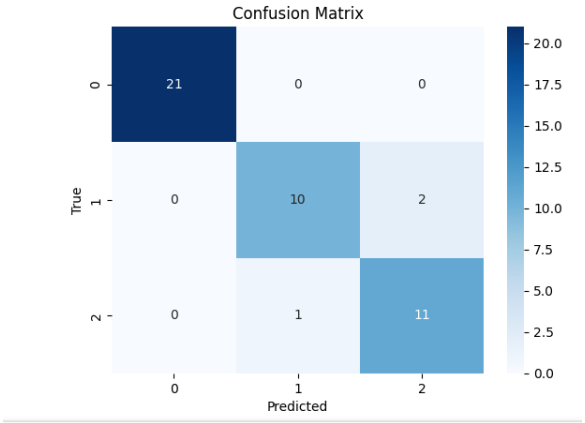
Extropy, as a newer impurity measure focusing on the probability of error rather than uncertainty, demonstrated competitive performance close to gini. This suggests that extropy can be a viable alternative impurity measure in decision tree learning, especially in scenarios where error-focused impurity assessment is desirable.

4.3 Confusion Matrices

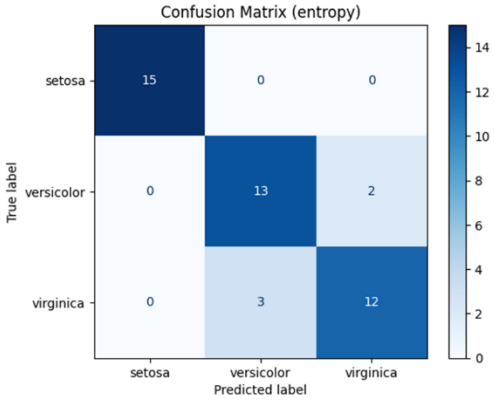
To further analyze the classification performance of the decision tree models based on Extropy, Entropy, and Gini Index, confusion matrices were generated for each model. The confusion matrix is a comprehensive tool that summarizes the number of correct and incorrect predictions broken down by each class. It provides insights into which classes are being confused by the classifier.



a-Entropy



b-Extropy

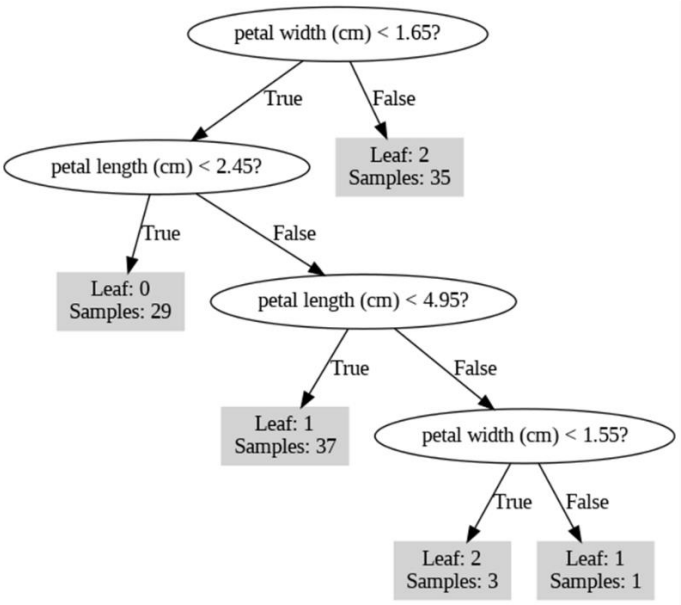


c-Gini Index

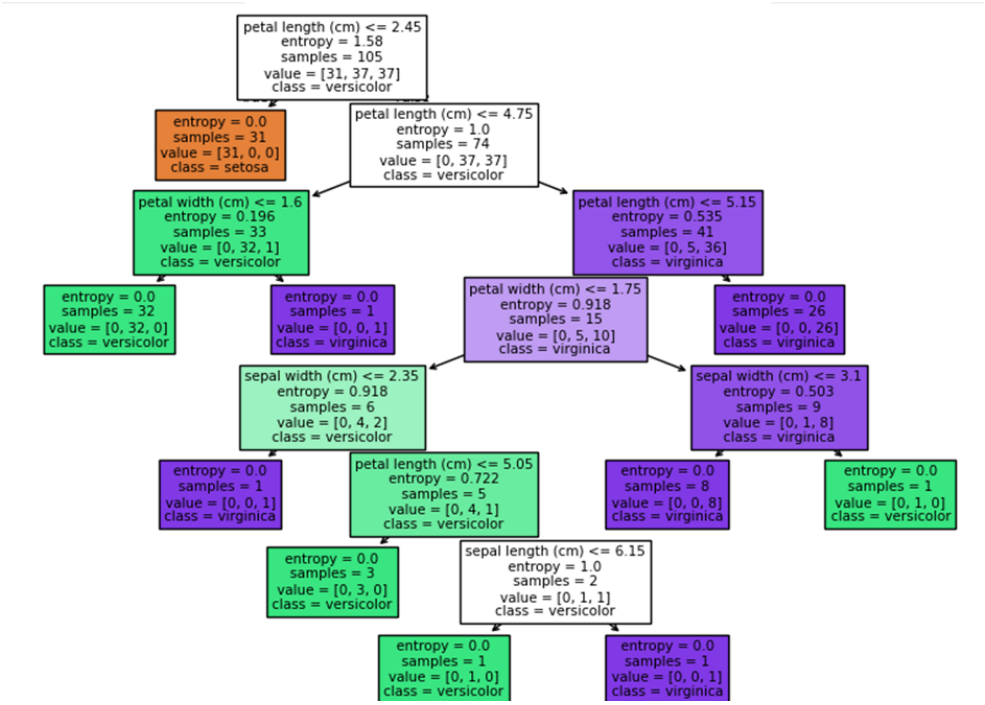
4.4 Decision Tree Visualizations

Graphical representations of the decision trees constructed using Extropy, Entropy, and Gini impurity measures are presented in Figures A, B, and C respectively. These visualizations illustrate the hierarchical decision rules and feature splits used by each model.

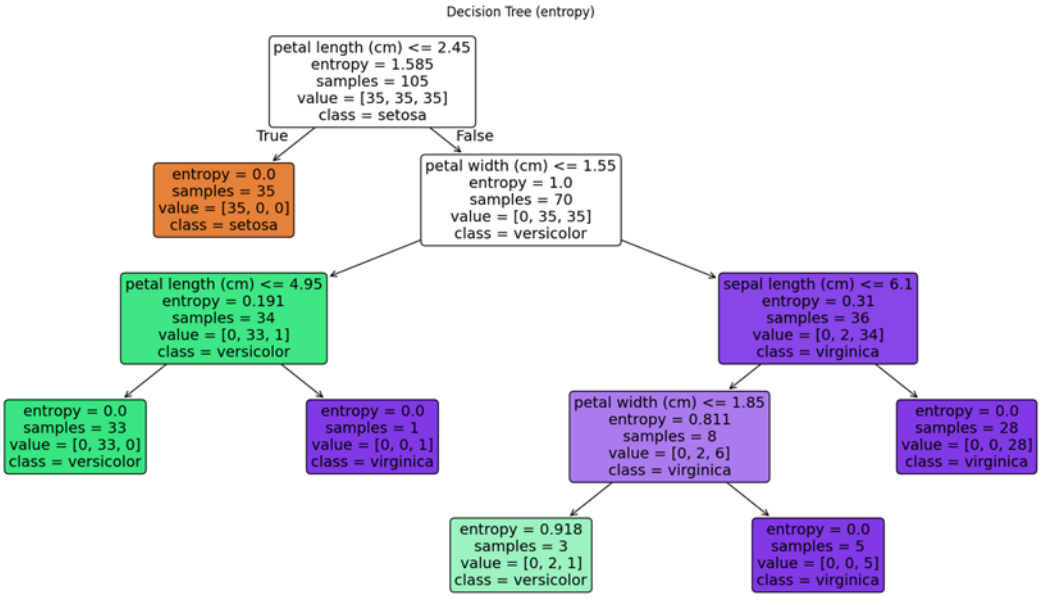
The trees predominantly split on petal length and petal width features, consistent with domain knowledge about the Iris dataset. The structures of the trees are similar, reflecting the comparable performance of the impurity measures, but subtle differences in split points and tree depth can be observed.



A- Extropy



B-Entropy



C-Gini Index

5.Conclusion

This study demonstrated that extropy, as a new impurity measure in decision tree algorithms, offers competitive performance close to the Gini index and can serve as a suitable alternative to entropy in scenarios where focusing on the probability of error is important. Although the entropy-based tree achieved higher accuracy in this research, the performance difference with extropy and Gini was negligible, and extropy may provide computat5.ional advantages. Therefore, the choice of impurity measure should be based on the specific needs of the problem and the trade-off between accuracy and computational efficiency. It is also recommended that

further extensive experiments be conducted on more diverse datasets to better evaluate the generalizability and benefits of extropy.

6. References

[1] Han, J., Kamber, M., & Pei, J. (2012). Data mining: Concepts and. *Techniques, Waltham: Morgan Kaufmann Publishers*.

[2] Lad, F., Sanfilippo, G., & Agro, G. (2015). Extropy: Complementary dual of entropy.

[3] Quinlan, J. R. (1986). *Induction of decision trees*. Machine Learning, 1, 81-106.

[4] Quinlan, J. R. (1996). *Improved use of continuous attributes in C4.5*. Journal of Artificial Intelligence Research, 4, 77-90.

[5] Unwin, A., & Kleinman, K. (2021). The iris data set: In search of the source of virginica. *Significance*, *18*(6), 26-29.