Fruit Classification Using Decision Trees and AdaBoost Ensemble Learning

Boosting

Md. Emran Khan [194063], Anamul Haque [194110], Md. Muntasir Mahammud[194119]

DUET, Gazipur

October 2023

Table of Contents

Abstract

This project employs decision tree classification augmented by AdaBoost to classify fruits based on their weight and mass attributes. The dataset encompasses a diverse range of fruits, with emphasis on these two features. The decision tree provides initial classification insights, while AdaBoost iteratively refines the model for enhanced accuracy. Experimental results demonstrate a substantial improvement in classification accuracy. This approach holds practical significance in automated fruit sorting systems, emphasizing the importance of weight and mass attributes in accurate classification.

Introduction

Accurate fruit classification is pivotal in modern agricultural and food processing industries, where efficient sorting and distribution processes are paramount. Among the key determinants for distinguishing fruits, attributes like weight and mass hold particular significance. These metrics offer valuable insights into the physical properties and composition of fruits, making them crucial factors in automated sorting systems.

Introduction

Traditional decision tree classification methods have proven effective in fruit categorization, providing interpretable models for differentiation. However, their performance can be further amplified through the incorporation of ensemble learning techniques. AdaBoost, a powerful ensemble learning algorithm, presents an opportunity to refine decision tree classifiers by iteratively focusing on misclassified samples. This iterative learning process proves especially advantageous when dealing with fruits that may share similar characteristics but vary subtly in weight or mass.

Introduction

This study aims to synergize decision tree classification with AdaBoost to improve fruit classification accuracy, leveraging the collective strengths of these methodologies. By emphasizing the importance of weight and mass attributes, we seek to enhance the efficiency of automated fruit sorting systems. The implications of this research extend to various industries, positively impacting quality control measures and streamlining distribution processes.

Related Work

Existing research has established the effectiveness of decision tree classification in fruit categorization. While AdaBoost has demonstrated significant improvements in classification accuracy across diverse domains, there is limited exploration into its integration with weight and mass attributes for fruit classification, distinguishing it from other boosting techniques like Gradient Boosting and XGBoost.

Backgound

AdaBoost, introduced in 1996, is a prominent ensemble learning algorithm. It combines weak learners, emphasizing misclassified samples in iterations for robust model creation. Widely used, it excels in various domains due to its adaptability and resistance to overfitting.

System Details

What Is the AdaBoost Algorithm?

There are many machine learning algorithms to choose from for your problem statements. One of these algorithms for predictive modeling is called AdaBoost. AdaBoost algorithm, short for Adaptive Boosting, is a Boosting technique used as an Ensemble Method in Machine Learning. It is called Adaptive Boosting as the weights are re-assigned to each instance, with higher weights assigned to incorrectly classified instances.

System Details

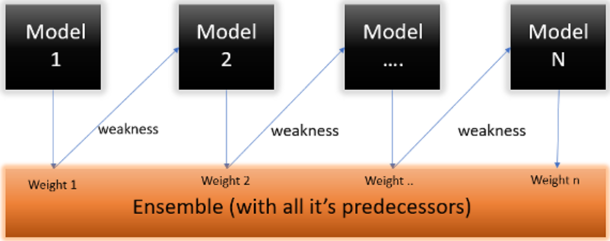


Figure: Boosting

Understanding the Working of the AdaBoost Algorithm

The formula to calculate the sample weights is:

$$w(x_i, y_i) = \frac{1}{N}, \; i = 1, 2, \ldots . n$$

Where N is the total number of data points
Here since we have 5 data points, the sample weights assigned will be 1/5

Understanding the Working of the AdaBoost Algorithm

Calculate the Influence We'll now calculate the 1 or 2 for this classifier in classifying the data points using this formula:

$$\frac{1}{2}\log\frac{1 - Total\ Error}{Total\ Error}$$

The total error is nothing but the summation of all the sample weights of misclassified data points.

Understanding the Working of the AdaBoost Algorithm

Calculate TE and Performance and Update weights

$$New\ sample\ weight\ =\ old\ weight\ *\ e^{\pm Amount\ of\ say\ (\alpha)}$$

The amount of, say (alpha) will be negative when the sample is correctly classified.
The amount of, say (alpha) will be positive when the sample is miss-classified.

Experiment

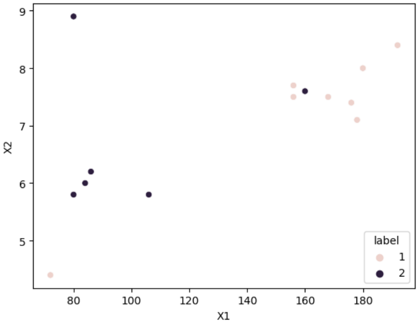


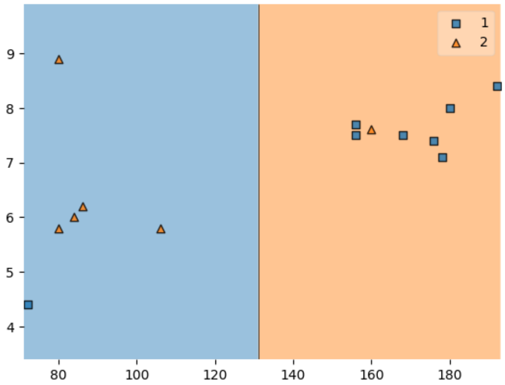
Figure: Dataset for our project

Experiment

For this project, we've used dataset from Kaggle. We took fruit properties such as fruit mass and fruit width.

```
df['label'] = [1,1,1,2,2,2,2,2,1,1,1,2,1,1]
df['X1'] = [192,180,176,86,84,80,80,106,178,72,156,160,156,168] #mass
df['X2'] = [8.4,8,7.4,6.2,6.0,5.8,8.9,5.8,7.1,4.4,7.7,7.6,7.5,7.5] #width
```

Experiment



data set

Experiment

We utilized fruit mass and width as features for classification. AdaBoost's weight update mechanism was pivotal in refining the model.

Experiment

	label	X1	X2	weights
6	2	80	8.9	0.071429
9	1	72	4.4	0.071429
7	2	106	5.8	0.071429
12	1	156	7.5	0.071429
5	2	80	5.8	0.071429
6	2	80	8.9	0.071429
2	1	176	7.4	0.071429
9	1	72	4.4	0.071429
5	2	80	5.8	0.071429
2	1	176	7.4	0.071429
9	1	72	4.4	0.071429
2	1	176	7.4	0.071429
11	2	160	7.6	0.071429
8	1	178	7.1	0.071429

Experiment

Misclassified samples were assigned higher weights in each iteration, allowing subsequent weak learners to focus on challenging instances. This iterative process significantly improved classification accuracy, showcasing the effectiveness of AdaBoost in handling nuanced variations in fruit attributes.

Experiment

	label	X1	X2	weights	y_pred
6	2	80	8.9	0.071429	2
9	1	72	4.4	0.071429	1
7	2	106	5.8	0.071429	1
12	1	156	7.5	0.071429	1
5	2	80	5.8	0.071429	1
6	2	80	8.9	0.071429	2
2	1	176	7.4	0.071429	1
9	1	72	4.4	0.071429	1
5	2	80	5.8	0.071429	1
2	1	176	7.4	0.071429	1
9	1	72	4.4	0.071429	1
2	1	176	7.4	0.071429	1
11	2	160	7.6	0.071429	2
8	1	178	7.1	0.071429	1

Conclusion

With this project, we've examined the effectiveness of the AdaBoost algorithm. The weight update mechanism proved instrumental in refining the model's accuracy. Through iterative adjustments, AdaBoost effectively fixed subtle variations in fruit attributes, which led to a substantial improvement in classification performance.