Artificial Intelligence Lab Report 7

1st Dipean Dasgupta
202151188
BTech CSE
IIIT,Vadodara

2nd Shobhit Gupta
202151149
BTech CSE
IIIT,Vadodara

3rd Rahul Rathore
202151126
BTech CSE
IIIT,Vadodara

4th Rohan Deshpande
202151133
BTech CSE
IIIT,Vadodara

***Abstract*—In this task, we delve into the examination of decision tree classifiers utilizing a dataset pertaining to automobiles. Our focus lies in evaluating these classifiers by varying the sizes of training data (60%, 70%, and 80%) and employing different methods for attribute selection (entropy versus Gini index). We gauge classification accuracy by employing confusion matrices and F-scores across 20 repetitions. Furthermore, we elucidate the concept of overfitting and provide an illustrative example. The findings provide valuable perspectives on the performance of decision trees and shed light on potential challenges in practical scenarios.**

I. Introduction

A decision tree exhibits a tree-like structure akin to a flowchart, where internal nodes symbolize features, branches signify rules, and terminal nodes denote the algorithm's output. It stands as a versatile supervised machine learning technique applicable to both regression and classification tasks. Renowned for its efficacy, decision trees serve as a foundational algorithm, further leveraged by Random Forests through training on diverse subsets of data, thus solidifying its status as a formidable machine learning approach.

II. Decision Tree

A. Methodology

The methodology entails the following steps:

1) Identification of the optimal attribute for data partitioning.
2) Construction of a hierarchical tree structure based on these partitions.
3) Termination of the process upon data classification or fulfillment of predefined criteria.

B. Entropy

Entropy quantifies the level of uncertainty or unpredictability within a dataset. In the context of classifications, it gauges randomness by analyzing the distribution of class labels. The formula for entropy computation is as follows:

$$H(S) = -p\log_2(p) - (1-p)\log_2((1-p))$$

C. Gini Index

The Gini Index assesses the accuracy of a split among categorized groups. It yields a numerical measure of impurity, with a Gini Impurity score of 0 indicating all observations in a single class and a score of 1 suggesting random distribution within classes. Minimizing the Gini index score is the objective:

$$Gini(S) = 1 - \sum_{i=1}^{n} p_i^2$$

D. Information Gain

Information gain denotes the reduction in entropy or variance achieved by partitioning a dataset using specific values. Enhanced information gain signifies the increased predictive value of the feature with regard to the target variable:

$$\text{Information Gain} = \text{Entropy}(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|}\text{Entropy}(S_i)$$

III. Problem Statement and Solutions

The car dataset available in the laboratory work folder is utilized for training and testing purposes, employing different test sizes and criteria.

A. Problem 1

The task involves randomly selecting 60% of labeled data from each class to form the training set, with the remaining 40% reserved for testing purposes. The objective is to evaluate the accuracy of the classification tree using a confusion matrix and F-score, employing attribute selection based on entropy. This process is repeated 20 times to compute the average accuracy.

The evaluation metrics including confusion matrices, training F1 score and accuracy, and testing F1 score and accuracy using entropy as the criteria are presented below:

```
Training F1 Score: 1.0

Training Accuracy: 1.0

Test Confusion Matrix:
     0    1   3   2
0  475    6   3   0
1    4  147   2   1
3    0    0  27   1
2    0    3   0  23

Test F1 Score: 0.9713422609731355

Test Accuracy: 0.9710982658959537
```

Fig. 1. Metrics

The average accuracy over 20 iteration is 0.9649.

```
Iteration 1: Accuracy = 0.976878612716763
Iteration 2: Accuracy = 0.9552023121387283
Iteration 3: Accuracy = 0.976878612716763
Iteration 4: Accuracy = 0.9725433526011561
Iteration 5: Accuracy = 0.9566473988439307
Iteration 6: Accuracy = 0.9523121387283237
Iteration 7: Accuracy = 0.9725433526011561
Iteration 8: Accuracy = 0.9653179190751445
Iteration 9: Accuracy = 0.9667630057803468
Iteration 10: Accuracy = 0.976878612716763
Iteration 11: Accuracy = 0.9739884393063584
Iteration 12: Accuracy = 0.9638728323699421
Iteration 13: Accuracy = 0.9421965317919075
Iteration 14: Accuracy = 0.9696531791907514
Iteration 15: Accuracy = 0.9725433526011561
Iteration 16: Accuracy = 0.9494219653179191
Iteration 17: Accuracy = 0.9609826589595376
Iteration 18: Accuracy = 0.9552023121387283
Iteration 19: Accuracy = 0.9725433526011561
Iteration 20: Accuracy = 0.9653179190751445
Average Accuracy over 20 iterations: 0.9648843930635838
```

Fig. 2. Avg. Accuracy over 20 Iterations(60-40)

B. Problem 2

Repeat the Prob. 1 using Gini index as criteria.

The Confusion matrix , Training F1 score and accuracy and Testing F1 score and accuracy using Gini index as criteria can be seen below:

```
Training F1 Score: 1.0

Training Accuracy: 1.0

Test Confusion Matrix:
     0    1   3   2
0  473    9   2   0
1    4  145   2   3
3    0    1  27   0
2    0    1   1  24

Test F1 Score: 0.9671344730451169

Test Accuracy: 0.9667630057803468
```

Fig. 3. Metrics

The average accuracy over 20 iteration is 0.9614.

```
Iteration 1: Accuracy = 0.9624277456647399
Iteration 2: Accuracy = 0.9552023121387283
Iteration 3: Accuracy = 0.9754335260115607
Iteration 4: Accuracy = 0.9725433526011561
Iteration 5: Accuracy = 0.9653179190751445
Iteration 6: Accuracy = 0.958092485549133
Iteration 7: Accuracy = 0.9667630057803468
Iteration 8: Accuracy = 0.9609826589595376
Iteration 9: Accuracy = 0.9494219653179191
Iteration 10: Accuracy = 0.9523121387283237
Iteration 11: Accuracy = 0.9638728323699421
Iteration 12: Accuracy = 0.9739884393063584
Iteration 13: Accuracy = 0.9667630057803468
Iteration 14: Accuracy = 0.9667630057803468
Iteration 15: Accuracy = 0.953757225433526
Iteration 16: Accuracy = 0.9638728323699421
Iteration 17: Accuracy = 0.9508670520231214
Iteration 18: Accuracy = 0.9552023121387283
Iteration 19: Accuracy = 0.9566473988439307
Iteration 20: Accuracy = 0.958092485549133
Average Accuracy over 20 iterations: 0.9614161849710984
```

Fig. 4. Avg. Accuracy over 20 Iterations(60-40)

C. Problem 3

Repeat above problems for 70% and 80% training data.

The Confusion matrix , Training F1 score and accuracy and Testing F1 score and accuracy using entropy index as criteria can be seen below:

```
Test Size: 0.3

 Confusion Matrix:
     0    1   3   2
0  359    2   2   0
1    4  110   1   0
3    0    1  18   2
2    0    0   1  19

 F1 Score: 0.9751115471972883

 Accuracy: 0.9749518304431599

Test Size: 0.2

 Confusion Matrix:
     0   1   3   2
0  241   1   0   0
1    1  74   1   1
3    1   3  10   0
2    0   0   0  13

 F1 Score: 0.9761392830843497

 Accuracy: 0.976878612716763
```

Fig. 5. Metrics (70%-80%) using Entropy criteria

The Confusion matrix , Training F1 score and accuracy and Testing F1 score and accuracy using Gini index as criteria can be seen below:

```
Test Size: 0.3

 Confusion Matrix:
     0    1   3   2
0  360    3   0   0
1    6  102   7   0
3    0    5  16   0
2    0    0   0  20

 F1 Score: 0.9596491569516618

 Accuracy: 0.9595375722543352

Test Size: 0.2

 Confusion Matrix:
     0   1   3   2
0  242   0   0   0
1    4  73   0   0
3    0   0  14   0
2    0   0   2  11

 F1 Score: 0.9825040272908178

 Accuracy: 0.9826589595375722
```

Fig. 6. Metrics (70%-80%) using Gini index criteria

The average accuracy over 20 iteration is for 0.3 and 0.2 test size using entropy as criteria can be seen below:

```
Test Size: 0.3
Average Accuracy over 20 iterations: 0.9694605009633909

Test Size: 0.2
Average Accuracy over 20 iterations: 0.9712186897880539
```

Fig. 7. Avg. Accuracy over 20 Iterations(0.3-0.2)

The average accuracy over 20 iteration is for 0.3 and 0.2 test size using gini index as criteria can be seen below:

```
Test Size: 0.3
Average Accuracy over 20 iterations: 0.9648362235067435

Test Size: 0.2
Average Accuracy over 20 iterations: 0.9662331406551059
```

Fig. 8. Avg. Accuracy over 20 Iterations(0.3-0.2)

D. Problem 4: Overfitting

Overfitting occurs when a machine learning model becomes overly tailored to the training data, leading to excellent performance on training data but poor performance on unseen data.

Illustratively, when examining the training F1 score and accuracy of models trained with entropy and Gini index using 60% of the data, both metrics display perfect scores of 1.0. However, discrepancies arise in the test confusion matrix, indicating imperfect performance on unobserved data. This discrepancy suggests potential overfitting, where the model may have memorized training data rather than learning underlying patterns.

Notably, increasing the size of the training data does not necessarily resolve the issue of overfitting, as observed throughout our investigation.

CONCLUSION

In conclusion, our study delved into decision tree classifiers using a car dataset, experimenting with different training data sizes and attribute selection criteria. While both entropy and Gini index were effective, we observed persistent challenges with overfitting. Despite increasing the training data size, overfitting remained a concern, highlighting the importance of robust model evaluation and addressing overfitting in machine learning applications.

REFERENCES

[1] Scikit-learn, "Decision Trees," Scikit-learn Documentation. [Online]. Available: https://scikit-learn.org/stable/modules/tree.html. [Accessed: May 1, 2024].

[2] GeeksforGeeks, "Decision Tree," GeeksforGeeks. [Online]. Available: https://www.geeksforgeeks.org/decision-tree/. [Accessed: May 1, 2024].

[3] Dr. Pratik Shah, "Car Dataset," 2024.