ZZSC5836 Hexamester 3, 2022
Sze Chan, z5061978
Assessment 3, Part B: Report

Tree and Ensemble Learning

Table of Contents

Abstract

The focus of this report is the deep unique analysis of how decision trees and ensemble learning can be used in the classification of abalone within a large data set. For this option, the number of rings was the value to predict, and this is a major contribution to the field as it helps streamline an otherwise time consuming and dull task which involved cutting through the shell to ultimately count the number of rings, making this work a significant addition to the literature. The abalone features that were extracted and used for the task included, but not limited to, those more convenient to measure such as weight and length. The classification and regression tree (CART) method was used to generate the model in the classification task and the Tree Visualisation provided in the report. The model was improved further with use of the post-pruning method on the tree.

Introduction

Decision trees have been a very popular choice for learning algorithms, particularly for learning classification models with the abalone dataset in this case, with their relatively simple-to-implement nature (Chandra, 2022). Despite their applications in important classification and regression tasks, decision trees do have their major limitations. Overfitting can occur when the decision tree algorithm selects only a subset of attributes and ignores the rest of the attributes (Chandra, 2022). Other challenges decision trees face include instability, where small changes to incoming data can cause large changes in the tree, potential complexity, using the same data but getting different models, and occasionally there is information overload (Bright Hub PM, 2011).

The motivation behind this project is not only because it is part of the course, but also it is applied to real life biological studies, which greatly interests me personally being a person invested in marine biology. Moreover, the idea of contributing a paper to help automate time consuming tasks with predictions induces a sense of achievement, especially if it is for rather tedious tasks like counting rings in abalone.

The rest of the paper is as follows. In section 2, the data is visualised, and the methodology is described in detail of how the work was set up, including description of the data, overview of the methods, a workflow diagram, the software used and the experiment setting. In section 3, the results of the experiment are thoroughly discussed. In section 4, a discussion covers the results and limitations of the methodology, and finally section 6 concludes the paper with major contributions and directions for future research.

Background

Decision trees are used to break down datasets into smaller subsets (Sayad, 2010), and at the same time other trees are gradually created starting with the root node, then working down the decision nodes and finally ending at the leaf nodes where the nodes do not split any further and are where classes get assigned by majority vote (Galarnyk, 2019).

Applications wise, we see they are good with classifying decisions for relatively straightforward events such as the decision to play tennis or not based on the weather (Chandra, 2022), or golf (Sayad, 2010) and customer relationship management (what-when-how, n.d.). For the purposes of this assignment, CART method has been used as the classification algorithm combined with the Gini impurity index method, which at a high-level means searching for the best homogeneity for the child nodes with the help of Gini Index which measures the purity of the classification (Dutta, 2021).

Decision trees to face their own downfalls as described above, however using regularisation where we set hyperparameters to reduce degrees of freedom (Chandra, 2022), as well as decision tree post- pruning (Chandra, 2022) can help reduce the effects of overfitting, information overload, instability, and complexity.

Methodology

Data

The abalone dataset was sourced from (Warwick et al., 1994) and has 8 attributes with 4177 instances. No missing or abnormal values were detected in the set, and I noticed all features are properly documented. Data types include nominal, continuous and integer.

To facilitate the cleaning of the dataset and later processing of data, the Numpy and Pandas libraries were imported, and in this early part to process the replacement of the gender of the abalone to numbers for more efficient processing later. It is to be noted that infant was chosen as -1 to not confuse it as adult, which were positive integers.

Also, upon inspection of the dataset, I decided properly documenting which column is for which characteristic is best for ease of understanding later, and categorising age groups would help make the processing more efficient as well. The first age group being ages 0 to 7 years denoted with a 1, second age group being ages 8 to 10 years denoted with a 2, third age group being ages 11 to 15 years denoted with a 3 and finally the fourth age group being more than 15 years.

The most important visualisation of the data was the heatmap to check the correlation between all the abalone characteristics (source) to get a better understanding of the relations (Kumar, 2022). It is to be noted the Seaborn open-source library was used to generate the images throughout this report.

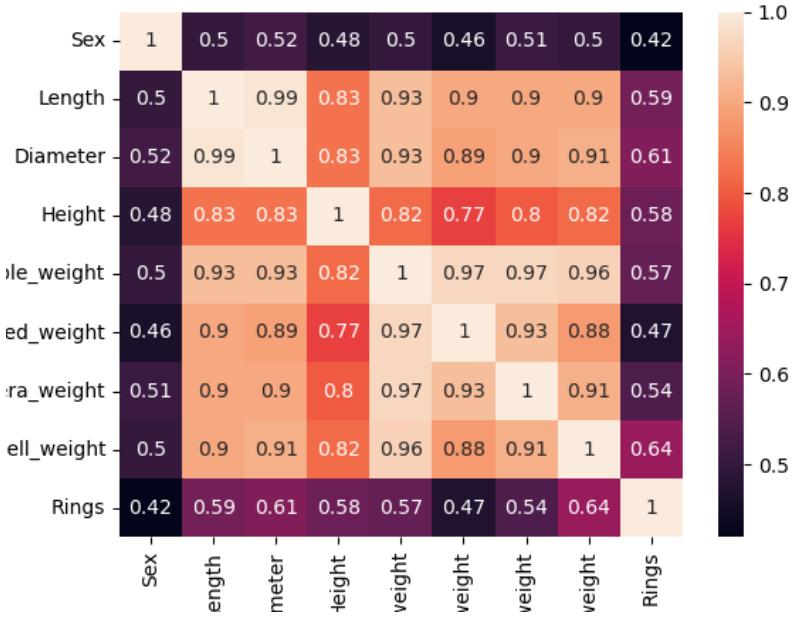

Figure 1 Heatmap

We see in this correlation heatmap that sex has correlation values that are lowest and has lowest linear trend with other variables, as seen by the colour of its squares. Other variables appear to have far higher linear trends with each other, being closer to the colour corresponding with 1.

Overview of the methods: Tree Visualisation

For this assignment, decision tree algorithm has been employed to compute the abalone classification problem. The libraries used for this classification task was Scikit-learn, notably Sklearn, and the data set has been split randomly such that 60% for training set and 40% for testing set. To assist the algorithm in choosing the feature to use at each of the decision nodes, and for the purposes of this assessment, the CART decision tree algorithm has been used within Scikit Learn, although there are other algorithms including ID3, C4.5 and C5.0 which all have their own merits (Chandra, 2022). For this CART method, the Gini impurity measure was used, which measures how intensely each specification directly affects in the resultant case (Patadiya, 2019). The absolute depth of the tree is controlled by a hyperparameter, else cannot be further divided.

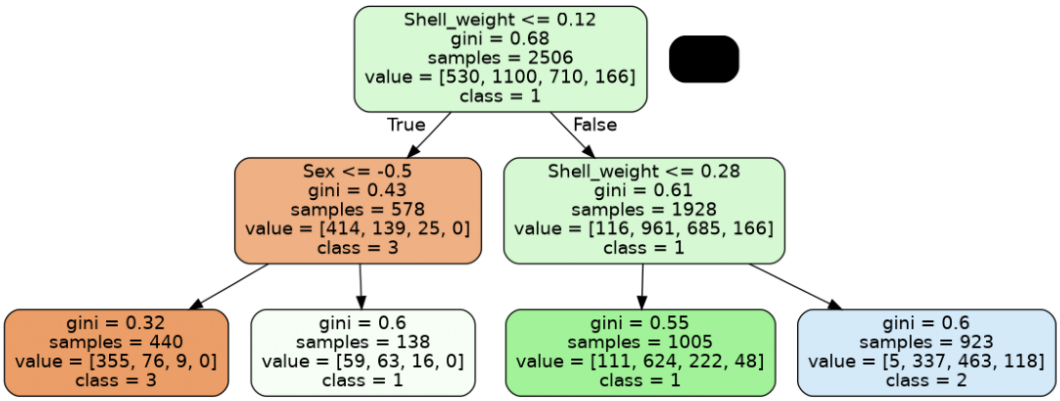

Figure 2 Decision Tree 1

Figure 2 is a decision tree with random state 25.

We see here that the If-Then rule for the tree is:

If Shell_weight <= 0.12 then
If Sex <= -0.5 then
Class = 3
Else class = 1

Else:
If Shell_weight<=0.4 then
Class = 1
Else class = 2

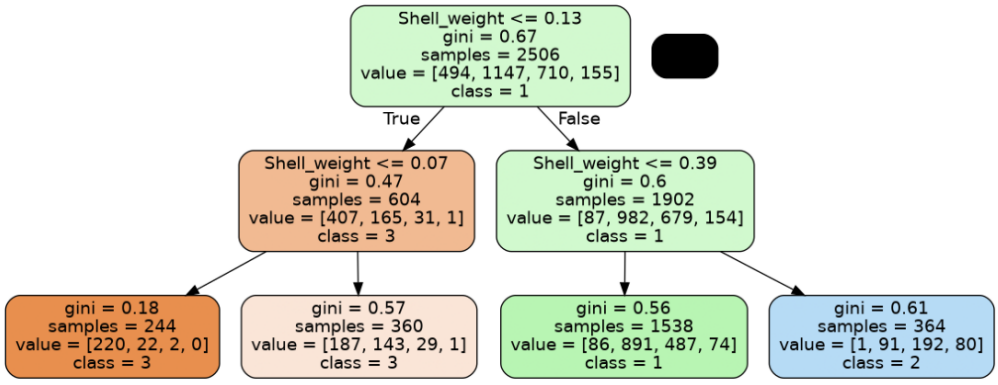

Figure 3 Decision Tree 2

Figure 3 is a decision tree with random state 10 for the new training set and test set. A different decision tree will result. Note that Figure 2 has only 3 results whilst the correct number should be 4 classes; having more layers and nodes would help the classification better.

Improving the performance further

It should be noted that overfitting is a common issue in decision trees when it fits all the samples of the training set. To avoid this, post-pruning of the tree is necessary, starting with a tree with no restrictions, then gradually removing the nodes that are not statistically significant to the accuracy of the prediction (Chandra, 2022), that is they have very little information gain at the decision node (Chandra, 2022).

The chi-square test is a good starting point in determining the significance of the split in the decision node (Singh, 2021), with those of no significance removed along with its children. To do this, the effective alpha is used as a limit to see whether the value of loss function at each node is lower than the limit, if so, it is removed. This is post-pruning the tree and a good method to mitigate the overfitting issue inherent in large decision trees.

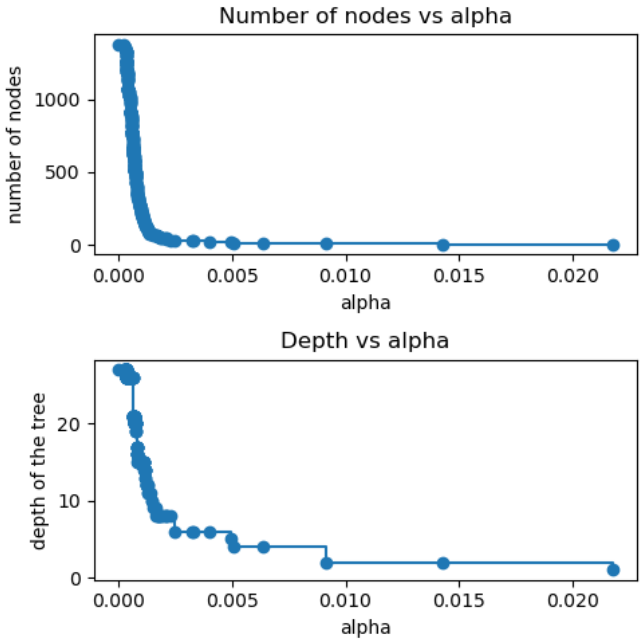

Figure 4 Nodes vs Alpha & Depth vs Alpha

In figure 4 we see that as alpha increases, the number of nodes quickly decreases, as too the depth of the tree. In essence, the node with the lowest effective alpha is first to go. At this stage, the highest effective alpha node is used to train the decision tree further and have a decision tree that generalises better.

In this figure below, we finally see that with the initial condition the training set has accuracy of 1.0, implying overfitting, however the training set has accuracy of 0.1 resulting in underfitting. After analysing the below figure further, the best accuracy to solve overfitting and underfitting is reached when the alpha is approximately 0.003 resulting in the depth of the tree being 5.

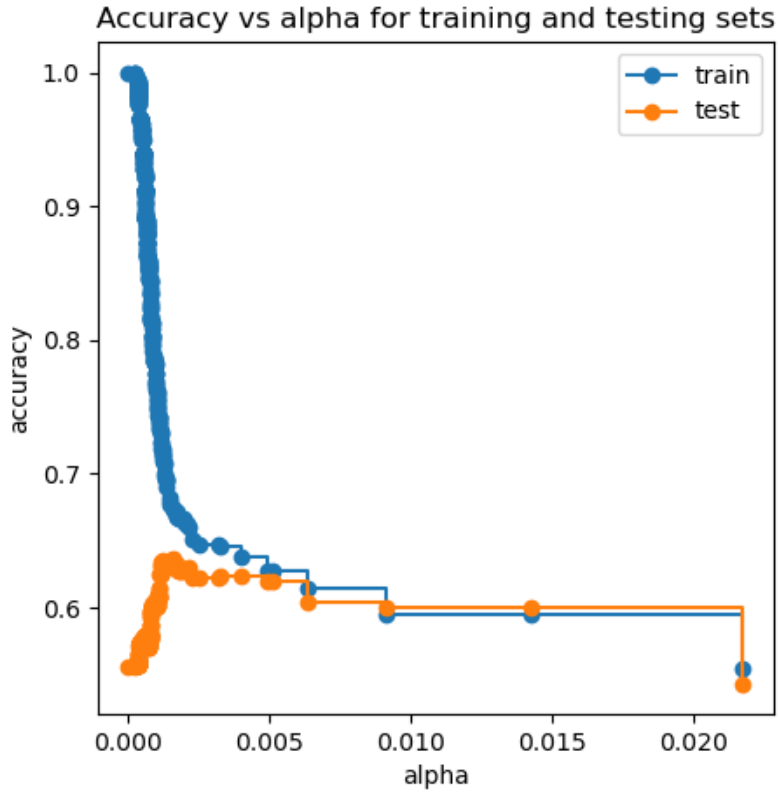

Figure 5 Measure of accuracy against alpha for training and testing sets

We thus decide to prune the layers beyond the 5th to achieve the optimal accuracy for both training and testing sets. The below figure shows the 5 layered decision tree that has best accuracy.

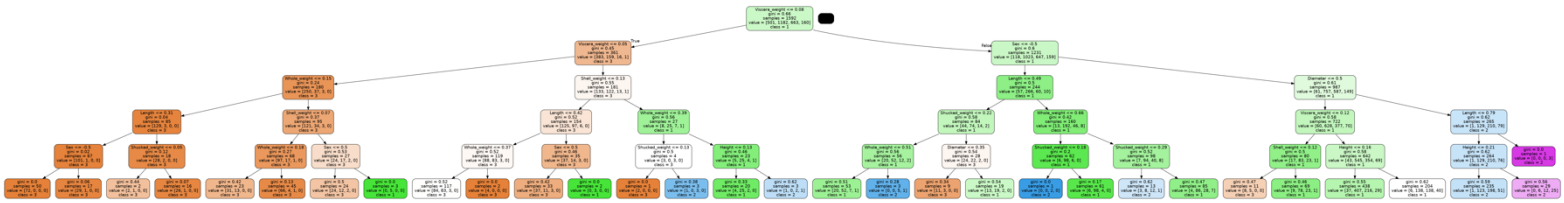

Figure 6 Decision tree with 5 layers

Results

In this section, the results of the above methodology are further discussed. The first figure depicts the correlation between abalone characteristics, summarised in a neat heatmap

with the darker colours symoblising a weaker correlation and the lighter colours symbolising a stronger correlation. It was noted that the gender had the lowest correlation with the other characteristics. The second figure depicts a simple 3-layer decision tree which was found to be inadequate, and post-pruning the decision tree was required to increase the accuracy, with figure 3 showing improvements with 4 classes instead of 3 at the child nodes. The fourth figure showed the number of decision nodes decreasing as the effective alpha increased, which in essence meant the node with the lowest effective alpha was first to be removed during the post-pruning. However, it was realised that the highest effective alpha resulted in a useless one node decision tree, and that had to be removed as well. Figure then shows the optimal effective alpha for optimal accuracy that solves the overfitting and underfitting issues inherent in decision tree algorithms and eventually the decision tree generalised better. The optimal depth of tree was found to be 5 layers and visualised in Figure 6.

Discussion

The tree visualisations and statistical graphs reveal the importance of post-pruning decision trees to improve their performance, as alluded to early in this report. However, there are a few limitations that restricted the scope of this experiment. First, the accuracy can be further improved with a larger dataset, environmental factors should be included in the set as they may well have a greater impact on the number of rings, and finally perhaps finding other characteristics to measure of the abalone would be beneficial and interesting, for example, the internal organs.

Conclusion

This paper makes significant contribution to the application of decision tree classification algorithms in predicting the number of rings on abalone, which essentially means automating a very boring and repetitive task in which the time could otherwise be used to measure other environmental factors or for other duties in general. It would be great for

future research to be directed at exploring environmental factors and recording them, then integrating the new data into the current model to enhance the classification performance.

List of References

Bottou, L. (1998). *Online Algorithms and Stochastic Approximations.* Cambridge: Cambridge University Press.

Bownlee, J. (2020, February 12). *Bagging and Random Forest for Imbalanced Classification*. Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/bagging-and-random-forest-for-imbalanced-classification/

Bright Hub PM. (2011, September 2). *A Review of Decision Tree Disadvantages*. Retrieved from BrightHubPM: https://www.brighthubpm.com/project-planning/106005-disadvantages-to-using-decision-trees/

Brownlee, J. (2017, July 3). *Gentle Introduction to the Adam Optimization Algorithm for Deep Learning*. Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/

Chandra, R. (2022, May). *Assessment 3: Project, Foundations of Neural Networks*. Retrieved from Edstem: https://edstem.org/au/courses/8454/lessons/21224/slides/150647

Chandra, R. (2022, May). *Decision Trees*. Retrieved from Edstem: https://edstem.org/au/courses/8454/lessons/20313/slides/144655

Chandra, R. (2022, May). *Ensemble Learning*. Retrieved from Edstem: https://edstem.org/au/courses/8454/lessons/21236/slides/150762

Chandra, R. (2022, May). *Overfitting vs Underfitting the data*. Retrieved from Edstem: https://edstem.org/au/courses/8454/lessons/20313/slides/144660

Dutta, B. (2021, July 27). *A Classification and Regression Tree (CART) Algorithm*. Retrieved from AnalyticSteps: https://www.analyticssteps.com/blogs/classification-and-regression-tree-cart-algorithm

Galarnyk, M. (2019, August 1). *Understanding Decision Trees for Classification (Python)*. Retrieved from TowardsDataScience: https://towardsdatascience.com/understanding-decision-trees-for-classification-python-9663d683c952

Jiang, Y.-y. (2010). Selective Ensemble Learning Algorithm. *2010 International Conference on Electrical and Control Engineering*, (pp. 1859-1862). Wuhan.

Kumar, A. (2022, April 16). *Correlation Concepts, Matrix & Heatmap using Seaborn*. Retrieved from VitalFlux: https://vitalflux.com/correlation-heatmap-with-seaborn-pandas/#:~:text=A%20correlation%20heatmap%20is%20a,necessarily%20imply%20a%20causal%20relationship.

Makhijani, C. (2020, October 6). *Advanced Ensemble Learning Techniques*. Retrieved from TowardsDataScience: https://towardsdatascience.com/advanced-ensemble-learning-techniques-bf755e38cbfb

Makhtar, M., Abdullah, F. S., Baba, N., & Awang, M. K. (2015, November). *Current issues in ensemble methods and its applications*. Retrieved from ResearchGate: https://www.researchgate.net/publication/288818144_Current_issues_in_ensemble_methods_and_its_applications

Nagpal, A. (2017, October 18). *Decision Tree Ensembles- Bagging and Boosting*. Retrieved from TowardsDataScience: https://towardsdatascience.com/decision-tree-ensembles-bagging-and-boosting-266a8ba60fd9

Patadiya, R. (2019, July 6). *Gini Index -CART Decision Algorithm in Machine Learning*. Retrieved from Medium: https://medium.com/@riyapatadiya/gini-index-cart-decision-algorithm-in-machine-learning-1a4ed5d6140d

Sayad, S. (2010). *Decision Tree- Classification* . Retrieved from SaedSayad: https://www.saedsayad.com/decision_tree.htm#:~:text=Decision%20tree%20builds%20classification%20or,decision%20nodes%20and%20leaf%20nodes.

Singh, H. (2021, March 25). *How to select Best Split in Decision Trees using Chi-Square*. Retrieved from AnalyticsVidhya: https://www.analyticsvidhya.com/blog/2021/03/how-to-select-best-split-in-decision-trees-using-chi-square/

Warwick, N. J., Sellers, T. L., Talbot, S. R., Andrew, C. J., & Ford, W. B. (1994). *The Population Biology of Abalone (_Haliotis_ species) in Tasmania. I. Blacklip Abalone (_H. rubra_) from the North Coast and Islands of Bass Strait.* Retrieved from UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/datasets/Abalone

what-when-how. (n.d.). *Decision Tree Applications for Data Modelling (Artificial Intelligence)*. Retrieved from what-when-how: http://what-when-how.com/artificial-intelligence/decision-tree-applications-for-data-modelling-artificial-intelligence/