Machine Learning Project 3

Anthony Vasquez
Ph:313-920-8877
Email: avasque1@jh.edu

Editor: Anthony Vasquez

Abstract

This document presents findings for classification and regression methods run on several unique datasets. The ID3, CART and tree pruning techniques are used to investigate effectiveness on three each of classification and regression datasets.

Keywords: Classification, Regression, CART, ID3, Early-stopping, Reduced-error-pruning, Entropy, Gain, Information

1 Introduction

This document covers the data used and preprocessing of that data for project three. It then covers the implementation of the ID3 and CART decision tree algorithms that were used to build decision trees using the specified data. Additionally, one pruning technique is discussed for each of the decision tree methods. Finally, the results of the experiments are discussed, and a conclusion composed of lessons learned are explored.

1.1 Data Used

The datasets used for classification in these experiments include breast-cancer-wisconsin, car, and house-votes-84. For regression, abalone, forestfires, and machine datasets were used. See the appendix for more information on each of the datasets.

Measurement	abalone	breast-cancer	car	forestfires	house-votes	machine
Dataset Size	4177	699	1728	517	435	209
Train Size	2672	436	1104	328	148	132
Test Size	668	109	276	82	37	33
Holdout Size	835	136	345	103	46	41
Targets	na	2	4	na	2	na
Attributes	8	10	6	12	16	7
Avg. Train Time Per Set (s)	46	20	4.12	13.4	3.9	2.3
K-folds	5	5	5	5	5	5

Table 1. Dataset statistics are shown for all of the datasets used in project 3.

1.2 Data Preprocessing

All datasets underwent standardization by using the equation $z(x) = \frac{x-\mu}{\sigma}$, where x is the data value, μ is the mean, and σ is the attribute standard deviation. Categorical attributes were encoded by using non-binary encoded values prior to standardization.

1.3 Stratified K-fold Cross Validation

The data was separated into several stratified train and validation sets for k-fold cross validation, where k = 5. The full data was stored as a pandas dataframe but was subsequently broken into smaller dictionaries for the randomized sampling of the data and each dictionary was appended to a list that was iterated through to obtain the five train and validation sets.

2 Algorithms

This section discusses the machine learning algorithms throughout this experiment.

2.1 Iterative Dichotomiser 3 (ID3)

The ID3 algorithm uses a tree-like structure to partition data into subsets. The subsets of data are separated using the entropy algorithm $E = -\sum_{i=1}^{c} p_i log_2 (p_i)$, where p is the probability of a value being in the subset. The lower the entropy, the more information can be gained from the data of an attributes. The information gain, calculated as $I = E_{parent} - E_{feature}$ is the weighted entropy for each of the feature sets. The highest information gain is used to partition the data into unique values found in the feature subset. The result of partitioning the data using ID3 is a decision tree with bins of decision values based on input data.

2.2 Classification and Regression Trees (CART)

CART is like ID3 and can be used for both classification and regression tasks. For classification, the Gini-index is substituted in lieu of the entropy calculation. There are several possible methods to decide on dataset splits, but the method used here is using the mean-squared-error (MSE). MSE is calculated as $MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2$, where n is the sample number, Y is the truth label and $\bar{Y}$ is the predicted label which is usually a type of mean. The result of partitioning data in using the CART method is a decision tree that is like ID3. The difference between the two algorithms is not in the structure of the tree, but rather the method used to build the tree. CART is a more flexible approach and generally easier to implement.

2.3 Early-stopping (ES)

ES is a type of tree pruning, and it specifically belongs to the class of techniques called pre-pruning. Pre-pruning is used to reduce the size of the tree, thereby reducing inference time while potentially lowering the test accuracy as a tradeoff. In ES a method is used to stop tree growth by train-cycle-termination. This can be done by stopping the training prematurely and testing the validation or holdout set to compare the accuracies of pre-pruned vs not pre-pruned methods.

2.4 Reduced Error Pruning (REP)

REP type of tree pruning, and it specifically belongs to the class of techniques called post-pruning. Post-pruning is used to reduce the size of the tree, thereby reducing inference time while potentially lowering the test accuracy as a tradeoff. REM takes the trained tree as input and begins pruning the lower most decision nodes that are least used. Once the tree is pruned, the holdout or validation set is used to calculate the difference in accuracy. In some cases, the accuracy does not change, but the tree is smaller. In yet other cases, the accuracy fluctuates in either the positive or negative but with a significantly smaller tree.

3 Results

This section discusses the results of the implementation of the ID3 algorithm with REP and without post-pruning, on the breast-cancer, car, and house-votes classification datasets. The results of the implementation of CART with and without ES on abalone, forestfires, and machine regression datasets

3.1 ID3 Results

Test Set	Not Pruned Accuracy
1	0.85
2	0.77
3	0.7
4	0.82
5	0.66
Mean	0.76

Table 2: Breast-cance-wisconsin validation test statistic prior to pruning

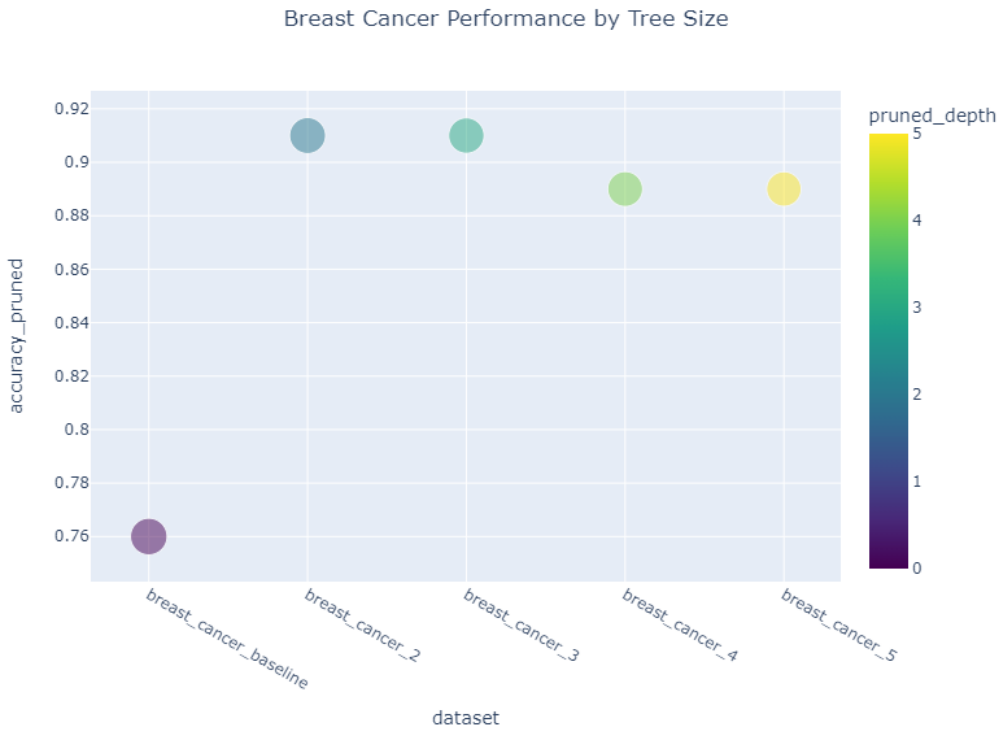

Figure 1: Breast-cancer-baseline is the baseline accuracy prior to pruning. It can be seen that pruning had a significant impact on the test accuracy.

Test Set	Not Pruned Accuracy
1	0.51
2	0.56
3	0.55
4	0.52
5	0.62
Mean	0.552

Table 3: Car validation test statistics prior to pruning

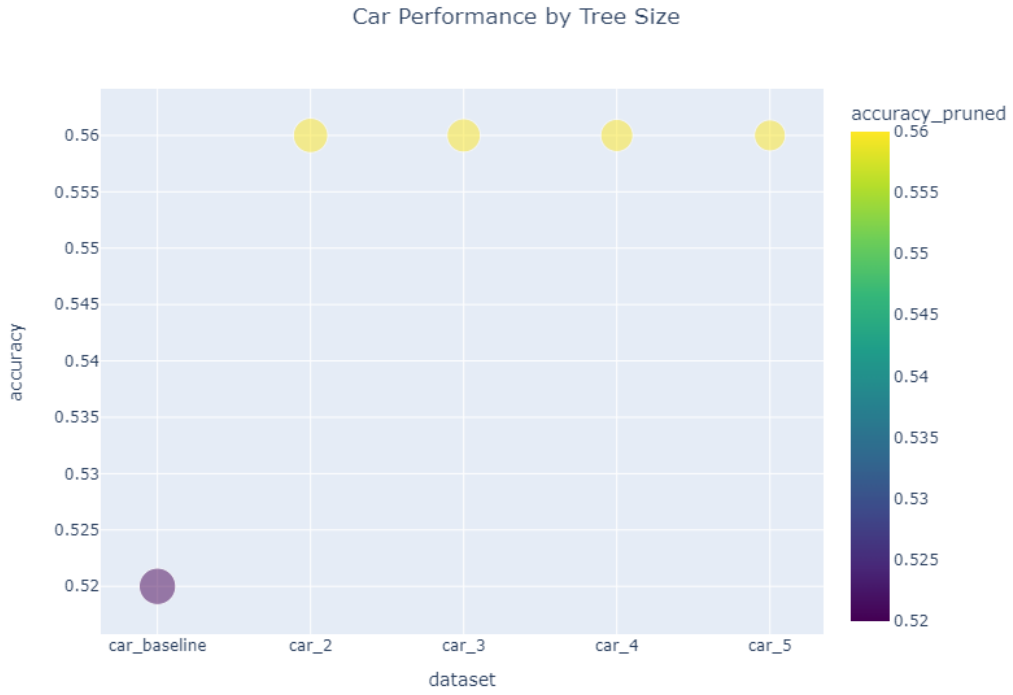

Figure 2: Car tests low at about 52 percent accuracy given four classes. It is definitely better than guessing, but pruning did seem to help by about four percent. It is possible that the tree was over-pruned.

Test Set	Not Prune Accuracy
1.00	0.84
2.00	0.84
3.00	0.78
4.00	0.66
5.00	0.81
Mean	0.79

Table 4: House-votes-84 validation test statistics prior to pruning

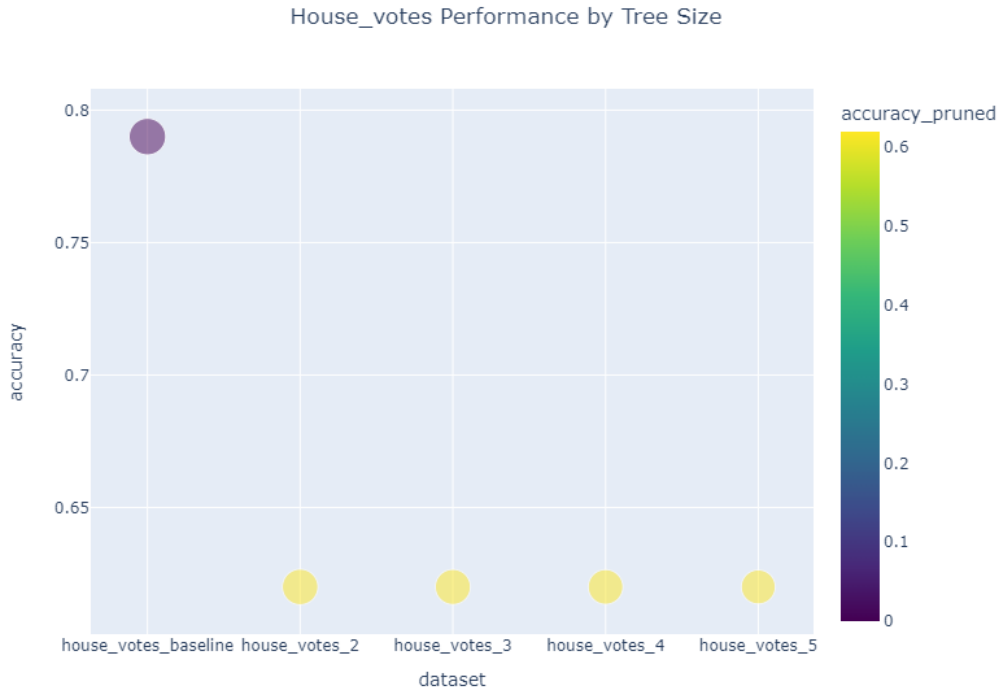

Figure 3: The house-votes-84 dataset is tricky. Even for a small pruning of just two nodes, the impact is obvious. It is possible that all the last nodes are being used to make decisions for the holdout dataset.

3.2 CART Results

Test Set	Not Pruned MSE
1	0.74
2	0.78
3	0.8
4	0.67
5	0.87
Mean	0.772

Table 5: Abalone validation test statistics prior to early stopping

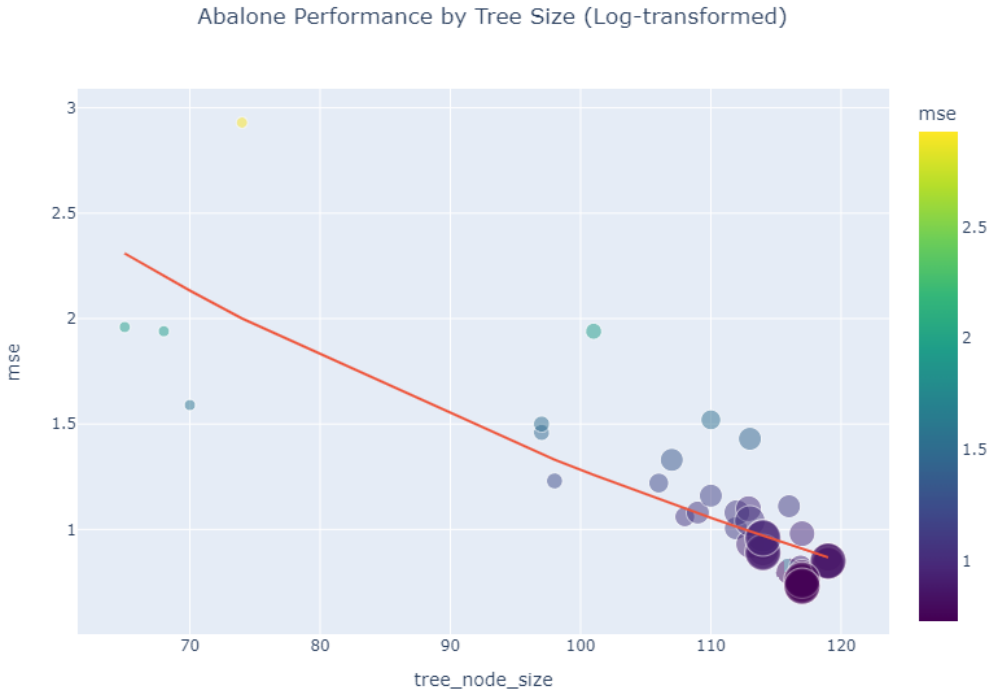

Figure 4: Logistic regression was used to fit a line in the general direction of test mse score. The tree_node vs mse plot shown here suggests that the mse is highest in the 70 to 90 tree-node-sizes, and as the tree increases in size, the test mse approaches its lowest value at around about 115 nodes.

Test Set	Not Pruned MSE
1	0.34
2	0.67
3	0.44
4	1.63
5	3.97
Mean	1.41

Table 6: Abalone validation test statistics prior to early stopping

Figure 5: Logistic regression was used to fit a line in the general direction of test mse score. The tree_node_size vs mse seems to suggest that as tree node size increases, the mse approaches its lowest value at about 30 nodes.

Test Set	Not Pruned MSE
1	0.49
2	0.04
3	1.24
4	0.03
5	0.27
Mean	0.42

Table 7: Machine validation test statistics prior to early stopping

Figure 6: Logistic regression was used to fit a line in the general direction of test mse score. This example was interesing in that increasing the node size by just one node increases the mse by two-fold. Twenty two seems to be around the best node size for this dataset.

4 Conclusion

This was a challenging project for me since it required plenty of forethought prior to implementation, which meant using sudo-code and a combination of object-oriented with linear-programming. There were a few things, after implementation, something clicked, and I kicked myself for not implementing the code in a cleaner more precise manner. This cannot be made more apparent than the contrast in the implementation of the CART and ID3 algorithms. After ID3, I realized how linear it was and decided to focus on a more object-oriented approach, which is what is seen with the CART implementation. Once nearly done with CART, I realized that the only real difference was the split mechanism, and therefore using a more modular approach could have saved a time and effort. With that said, I learned a lot and can probably implement decision trees in several unique ways because of the lessons learned.

References

Marko Bohenec, B. Z. (1997, June). Car Evaluation Database. Avignon, France.

McCormick, C. (2014, Febuary 2014). *Kernal Regression.* Retrieved from mccormickml.com: https://mccormickml.com/2014/02/26/kernel-regression/

Paulo Cortez, A. M. (2007, December). Forest Fires. Guimaraes, Portugal.

Phillip Ein-Dor, J. F. (1987, October). Relative CPU Performance Data. Tel Aviv, Israel.

Schlimmer, J. (1984). 1984 United States Congressional Voting Records Database. Washington D.C., USA.

Waugh, S. (1995). Extending and benchmarking Cascade-Correlation. Hobart, Tasmania, Australia.

Wolberg, W. H. (1992, July 15). Wisconsin Breast Cancer Database. Madison, Wisconsin, USA.

Appendix

The following is a brief description of the dataset used, as well as source information.

- Abalone is a regression dataset used to predict the age of a gastropod species by counting the number rings on the inner shell of the sample. There are 4177 samples with eight attributes to include, sex, length, diameter, height, whole weight, shucked weight, viscera weight, shell weight, ring number (Waugh, 1995). There are no missing values in this dataset.

- Breast-cancer-wisconsin is a dataset that includes 699 samples with 10 attributes and is used to classify either benign or malignant cancer types (Wolberg, 1992). This dataset is missing 16 values that are denoted by a "?."

- Car is a classification dataset used to evaluate hierarchy induction tools. There are 1728 samples with six attributes and four classes including, unacc, acc, good, and v-good (Marko Bohenec, 1997). This dataset has no missing values.

- Forestfires is a regression dataset used to predict the burn area of a forest fire based on 517 samples, and >12. The attributes include special coordinates X and Y of a park map, month, data, indices from the Fire Warning Information system, temperature, relative humidity, wind speed, rain, and burn area (Paulo Cortez, 2007). This dataset has no missing values.

- House-votes-84 is a classification dataset used to predict whether a congressman from the house of representatives belongs to the republican or democrat parties based whether they voted on 16 key votes: handicapped-infants, water-project-cost-sharing, adoption-of-the-budget-resolution, physician-fee-freeze, El-Salvador-aid, religious-groups-in-school, anti-satellite-test-ban, aid-to-Nicaraguan-contras, mx-missile, immigration, synfuels-corporation-cutback, education spending, superfund-right-to-sue, crime, duty-free-exports, and export-administration-act-south-Africa (Schlimmer, 1984).

- Machine is a regression dataset that is used to predict relative CPU performance and includes 209 samples with 10 attributes. Attributes include, vendor name, model name, machine, cycle time, minimum main memory, maximum main memory, cache memory, minimum channels, maximum channels, published relative performance, and estimated relative performance (Phillip Ein-Dor, 1987). This dataset has no missing values.