Anthony Gray
EN.605.649.81
Intro to Machine Learning
Project 4

Abstract

The ID3 algorithm was used to classify examples from three datasets using the gain ratio of the dataset's features. It performed well in two cases, and very poorly in the third. Performance was related to the number of classes in the dataset, with a lower class count correlated with classification accuracy. The evaluation of the implementation is undermined by the presence of several bugs, and the absence of a reduced-error pruning function, in part due to those same bugs.

Problem Statement and Algorithms

The ID3 algorithm (1) was applied to three datasets. ID3 grows a decision tree by recursively selecting the feature with the highest information gain ratio, splitting the dataset by that feature, and repeating the process until the resulting node contains only one class. Categorical data is split by category, while continuous data was split by value that maximized information gain. All possible splits of the feature were considered, excluding the ones between equal values or instances that shared a class. The gain ratio for each split was then calculated, and the data split around the associated value.
The gain ratio is computed using Shannon's Entropy, subtracting the entropy of a feature from that of the dataset, then dividing by the intrinsic value of the dataset. The trees were evaluated using classification error and five-fold validation.
Reduced-error pruning increases the accuracy of a decision tree and reduces overfitting by replacing nodes with leaf nodes containing the most common value reachable by that node. Nodes are replaced until the error of the tree on a validation set begins to diverge from the error of the tree on a test set.
Decision trees were built for three datasets Abalone (2), Car Evaluation (3) and Image Segmentation (4). It is anticipated that the Car Evaluation set will be the easiest to classify by decision tree, as it has both the lowest number of attributes and the lowest number of classes and consists entirely of categorical data. Image Segmentation, while it has the most attributes, also has a relatively low number of classes. However, all of it's attributes are continuous, so relatively deep trees are expected to be generated. The Abalone dataset is likely to be the most challenging to classify with a decision tree

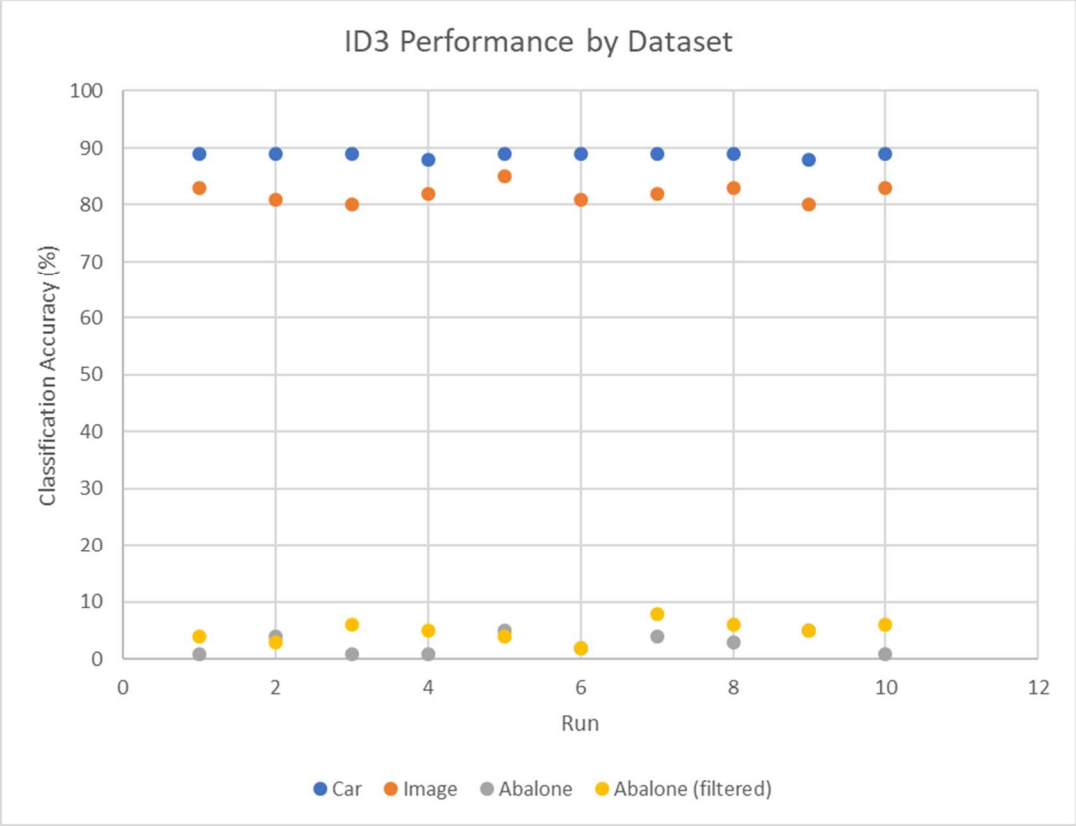
Procedure/Tuning

Dataset	**Instances**	**Attributes**	**Classes**
Abalone	4177	8	29
Car Evaluation	1728	6	4
Image Segmentation	210	19	7

All of the datasets were obtained from UCI Machine Learning Repository (5). The class values in the Image dataset were moved from the row labels to the last column in the dataset. The abalone dataset had the Sex value changed from characters to integers, with I mapping to 0, F mapping to 1, and M mapping to 2. None of the datasets had missing values.
The algorithm was run without pruning on the Car and Image sets in their entirety ten times, recording the average five-fold validation accuracy each time. The abalone set was run ten times on random subsets of between 500 and 2,000 rows, as the first few runs on the whole dataset took a very long time and were prone to crashing due to bugs in the implementation. Additionally, runs on the Abalone set were made with the low-count classes removed. Classes 1, 2, 24, 25, 26, 27, 28 and 29 all had at most two examples in the entire dataset and were removed to see if doing so improved performance.

Unfortunately, the design and implementation of the ID3 tree was not performed with enough care. The construction of the tree contained several bugs, and so much time was spent bringing it to a working standard that pruning was not implemented. The poor design of the tree made implementing the pruning algorithm too difficult and time-consuming to complete, and a re-factor was not possible by the time this was realized.

Results



Overall, the unpruned ID3 tree was successful in classifying items from the Car and Image datasets. Almost every run resulted in a classification accuracy between 80 and 90 percent. With pruning, this accuracy percentage could potentially be improved into the 90s, increasing the reliability even further. On the other hand, the unpruned ID3 algorithm performed terribly on the Abalone dataset, almost certainly due to the absence of pruning. The Abalone training sets frequently produced very deep and complex trees, often with only one meaningful classification value within a cluster of leaves. With pruning, these leaf clusters could be condensed, probably resulting in a dramatic increase in accuracy. The filtered Abalone dataset did perform better than the unfiltered set, though the classification accuracy remained below 10 percent.

Discussion

The lack of pruning makes a discussion of this ID3 implementation difficult, as it is not possible to determine how much the classification error would be improved. Still, even without pruning, the accuracy on two of the datasets was already high. Pruning would presumably improve accuracy in these cases,

though probably not by very much. It was very clear that pruning would be necessary for this ID3 implementation to be of any use on the Abalone dataset. The trees generated by this implementation and the Abalone dataset tended to be very complex, with great depth and breadth, and many branches ended in leaf clusters containing few or no classification values at all. This is almost certainly a bug in the tree-building implementation, but a usable pruning function would have helped to mitigate it by removing the useless clusters or replacing them with a useful one.

Summary

As expected, the ID3 algorithm performed the best on the Car dataset, did almost as well on the Image dataset, and had the worst performance on the Abalone dataset. As bad as the performance was on the Abalone dataset, proper pruning would likely result in a significant improvement, due to the reasons mentioned above. Additionally, filtering the least represented classes dataset resulted in a small increase in performance, even without pruning. Furthermore, the classes in the Abalone set represent a number, specifically the age of the animal in question. In many of the misclassification cases, the prediction was within a one year of the correct answer. Using mean-squared error, rather than classification error would have likely resulted in significantly better performance due to these cases. Regardless, a thorough debugging and the reduced-error pruning function would still be necessary for this implementation to reach a reliable level of accuracy.

References

1) Quinlan, J. R. 1986. Induction of Decision Trees. Mach. Learn. 1, 1 (Mar. 1986), 81–106
2) Warwick J Nash, Tracy L Sellers, Simon R Talbot, Andrew J Cawthorn and Wes B Ford (1994) "The Population Biology of Abalone (Haliotis species) in Tasmania. I. Blacklip Abalone (H. rubra) from the North Coast and Islands of Bass Strait", Sea Fisheries Division, Technical Report No. 48 (ISSN 1034-3288)
3) M. Bohanec and V. Rajkovic: Knowledge acquisition and explanation for multi-attribute decision making. In 8th Intl Workshop on Expert Systems and their Applications, Avignon, France. pages 59-78, 1988.
4) Image Segmentation data, Source Information, Creators: Vision Group, University of Massachusetts, Donor: Vision Group (Carla Brodley, brodley@cs.umass.edu), Date: November, 1990
5) Dua, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science