Decision Trees(ID3) and Reduced Error Pruning

Amrith Deepak

Abstract:

The abalone data to predict the age of abalone from physical measurements, image segmentation data to classify pixels in an image, and car evaluation data to predict car acceptability based on price, comfort, and technical specifications. were analyzed using the ID3 Decision tree algorithm with and without reduced error pruning. The performance with pruning was consistently better than without pruning. The performance for the car evaluation data was highest, then segmentation next highest, followed by the abalone data. The car evaluation and abalone data has very small differences in accuracy between the pruned and non-pruned version but the image segmentation data had a larger difference.

Problem Statement:

This project is examining how well ID3 works for classification and examining how effective pruning is. These algorithms are being evaluated on how they perform on different datasets to classify data. 8 attributes are being used to classify age of abalones, 19 attributes are being used to classify pixels of an image, and 6 attributes are being used to predict whether the condition of a car is acceptable.

Hypothesis:

For classifying data, decision trees should give high performance. One problem with decision trees though is that they are prone to overfitting. They are extremely sensitive to small perturbations in the data. This might reduce performance, but overall high performance (>75% accuracy) is predicted. Decision trees also have problems with out of sample space prediction, which should not be an issue with these data sets.

Pruning should improve accuracy. When pruning, the decision tree is simplified only when it improves performance, so the performance in the datasets is expected to stay the same or go higher.

Algorithms:

ID3 first calculates the entropy of the data, which is the measure of impurity, disorder, or uncertainty. Then the information gain is calculated for each of the features and the feature with highest information gain is chosen as the feature to branch on. For each value of the feature, a decision tree is built for the feature being equal to that value. This process is done until all features have been examined in which case the most frequently occurring class is chosen or until all values are of the same class. The accuracy is the proportion of correctly classified points.

For reduced error pruning, first ID3 algorithm is used to create a decision tree. Then for each level of the tree, one branch of the tree will be removed and then the performance will be calculated. If removing the branch does not decrease accuracy, it will be removed. This will be repeated for all branches until the performance worsens.

Experimental Approach:

For each of the three datasets, the data was split into training and testing data using 5-fold cross validation. The data was chosen randomly. The algorithms described above were followed.

For the abalone dataset and the segmentation dataset, each of the variables were discretized before applying ID3. The variables were changed to 0 or 1 based on whether they were above or below the mean. That way there are branches that can be created. For the abalone dataset, the sex variable was not changed, since it already had 3 categories. The variable predicted is age, and has been grouped into 3 categories: 0-9 rings(1.5-10.5 years old), 10-19 rings(11.5-20.5 years old), and 20+ rings(>21.5 years old).

For the car evaluation data, all the variables had categories and the variable being predicted is car acceptability. The data was kept unchanged.

For the classification, accuracy rate is measured.

Results:

Dataset	**ID3 Performance With Pruning**	**ID3 Performance Without Pruning**
Abalone	0.5549	0.5185
Segmentation	0.7905	0.7143
Car Evaluation	0.9432	0.9299

The ID3 performed best for the car evaluation data for both with and without pruning. The acuracy was 0.9432 and 0.9299 respectively. For the abalone dataset, the performance was worst and the accuracy was 0.5549 and 0.5185 for pruned and not-pruned. For the segmentation data, the accuracy was 0.7905 and 0.7143 for pruned and not pruned.

The difference in performance for the abalone and car evaluation data was negligible, but pruning gave a considerable performance increase(almost 9% more accuracy) for the segmentation data.

Discussion:

The ID3 performed best for the car evaluation data for both with and without pruning. For the other 2 datasets, the features were classified into categories, where as in the car evaluation data, the original dataset was used. The accuracy was around 95% for the car data. One of the major problems of decision trees is overfitting. This occurred with the abalone data, especially since it had fewer features. The segmentation data had more features and the reduced error pruning further aided in lowering the overfitting. The difference in performance for the abalone and car evaluation data was negligible, but pruning gave a considerable performance increase(almost 9% more accuracy) for the segmentation data.

One of the major advantages of decision trees is that it is comprehensive. It considers of all possible outcomes of a decision and traces each path to a conclusion. Decision trees also assign specific values to problems, decision path, and outcome. This reduces uncertainty and ambiguity. Despite being prone to overfitting, these advantages of decision trees led to the decision tree performing above 50% in all datasets.

Summary:

The performance for the car evaluation data was highest, then segmentation next highest, followed by the abalone data. The car evaluation and abalone data has very small differences in accuracy between the pruned and non-pruned version but the image segmentation data had a larger difference. In all datasets, the pruned version performance was higher than the non-pruned version.

The code trained using 5-fold cross validation where each time 10% of the data was used for testing and 90% for training.

Works Cited

Bhardwaj, Rupali, and Sonia Vatta. “Implementation of ID3 Algorithm.” *Implementation of ID3 Algorithm*, International Journal of Advanced Research in Computer Science and Software Engineering, June 2013.

Elomaa, T., and M. Kaariainen. “An Analysis of Reduced Error Pruning.” *Journal of Artificial Intelligence Research*, vol. 15, 2001, pp. 163–187., doi:10.1613/jair.816.

Datasets:

https://archive.ics.uci.edu/ml/datasets/Image+Segmentation

https://archive.ics.uci.edu/ml/datasets/Abalone

https://archive.ics.uci.edu/ml/datasets/Car+Evaluation