Christopher Morgan

CS478: Brother Christophe

Project 1: Decision Tree

**Correctly implement the ID3**

* + **An option for choosing the splitting criterion: information gain or accuracy.**

The command line argument –UA (Use Accuracy) will flip the algorithm from its default maximize Information Gain to maximizing Accuracy.

* + **Some mechanism to handle continuous-valued attributes.**

My algorithm will simply divide a continuous-valued attribute in two. The attribute is split along the mean of its data values.

* + **Some mechanism to handle unknown (or missing) attribute values.**

My algorithm uses a preprocessing technique to fill in any missing attributes. The preprocessing strategy replaces any missing attribute values with the majority value for that attribute. This seems reasonable given that majority for that attribute is the best guess one can make about a given attribute value.

* + **You should be able to get ~68% predictive accuracy on lenses data with cross-validation.**

|  |  |
| --- | --- |
| Dataset name: ../data/lenses.arff  Number of instances (rows): 24  Number of attributes (cols): 5  Learning algorithm: DecisionTree  Evaluation method: cross  Rep: 0, Fold: 0, Accuracy: 0.5  Rep: 0, Fold: 1, Accuracy: 0  Rep: 0, Fold: 2, Accuracy: 1  Rep: 0, Fold: 3, Accuracy: 1  Rep: 0, Fold: 4, Accuracy: 1  Rep: 0, Fold: 5, Accuracy: 0.5  Rep: 0, Fold: 6, Accuracy: 1  Rep: 0, Fold: 7, Accuracy: 0.5  Rep: 0, Fold: 8, Accuracy: 1  Rep: 0, Fold: 9, Accuracy: 0.5  Mean predictive accuracy: **0.7** | **Using Accuracy:**  Dataset name: ../data/lenses.arff  Number of instances (rows): 24  Number of attributes (cols): 5  Learning algorithm: DecisionTree  Evaluation method: cross  Rep: 0, Fold: 0, Accuracy: 1  Rep: 0, Fold: 1, Accuracy: 0.5  Rep: 0, Fold: 2, Accuracy: 1  Rep: 0, Fold: 3, Accuracy: 0.5  Rep: 0, Fold: 4, Accuracy: 0  Rep: 0, Fold: 5, Accuracy: 0.5  Rep: 0, Fold: 6, Accuracy: 1  Rep: 0, Fold: 7, Accuracy: 1  Rep: 0, Fold: 8, Accuracy: 0.5  Rep: 0, Fold: 9, Accuracy: 1  Mean predictive accuracy: **0.7** |

**Use your ID3 algorithm on the**[**Iris**](http://dml.cs.byu.edu/~cgc/docs/mldm_tools/Assignments/Datasets/iris.arff)**problem.**

|  |  |
| --- | --- |
| **Information Gain** | **Accuracy** |
| Node: Property: petallength  Node: Property: petalwidth  Node: Class: 0  Node: Property: sepalwidth  Node: Class: 1  Node: Class: 0  Node: Property: petalwidth  Node: Property: sepalwidth  Node: Property: sepallength  Node: Class: 1  Node: Class: 1  Node: Property: sepallength  Node: Class: 1  Node: Class: 1  Node: Property: sepalwidth  Node: Class: 2  Node: Property: sepallength  Node: Class: 2  Node: Class: 2 | Node: Property: petalwidth  Node: Property: petallength  Node: Class: 0  Node: Property: sepallength  Node: Property: sepalwidth  Node: Class: 1  Node: Class: 0  Node: Class: 1  Node: Property: petallength  Node: Property: sepallength  Node: Property: sepalwidth  Node: Class: 1  Node: Class: 1  Node: Property: sepalwidth  Node: Class: 1  Node: Class: 1  Node: Property: sepallength  Node: Property: sepalwidth  Node: Class: 2  Node: Class: 2  Node: Class: 2 |

* + **Compare this tree with the one obtained with information gain as the splitting criterion.**

The main order of the Entropy algorithm was to choose petalLength, petalWidth, and then sepalWidth in that order. However, when switching to the Accuracy algorithm it favored petalWidth, petalLenght, then sepalLength in that order. Changing algorithms to accuracy therefore cause an increase in weight for petalWidth, cause it to be selected first over petalLength. In addition to the order of attributes chosen the switch in algorithms also created a slightly larger tree when accuracy was used. The Entropy tree consisted of 19 nodes where the Accuracy tree was made up of 21.

* + **Evaluate predictive accuracy using 10-fold cross-validation for information gain and accuracy.**

|  |  |
| --- | --- |
| Dataset name: ../data/iris.arff  Number of instances (rows): 150  Number of attributes (cols): 5  Learning algorithm: DecisionTree  Evaluation method: cross  Rep: 0, Fold: 0, Accuracy: 1  Rep: 0, Fold: 1, Accuracy: 1  Rep: 0, Fold: 2, Accuracy: 1  Rep: 0, Fold: 3, Accuracy: 1  Rep: 0, Fold: 4, Accuracy: 0.933333  Rep: 0, Fold: 5, Accuracy: 0.866667  Rep: 0, Fold: 6, Accuracy: 1  Rep: 0, Fold: 7, Accuracy: 0.933333  Rep: 0, Fold: 8, Accuracy: 0.933333  Rep: 0, Fold: 9, Accuracy: 0.933333  Mean predictive accuracy: **0.96** | **Using Accuracy:**  Dataset name: ../data/iris.arff  Number of instances (rows): 150  Number of attributes (cols): 5  Learning algorithm: DecisionTree  Evaluation method: cross  Rep: 0, Fold: 0, Accuracy: 1  Rep: 0, Fold: 1, Accuracy: 0.866667  Rep: 0, Fold: 2, Accuracy: 1  Rep: 0, Fold: 3, Accuracy: 0.933333  Rep: 0, Fold: 4, Accuracy: 0.866667  Rep: 0, Fold: 5, Accuracy: 1  Rep: 0, Fold: 6, Accuracy: 1  Rep: 0, Fold: 7, Accuracy: 0.933333  Rep: 0, Fold: 8, Accuracy: 1  Rep: 0, Fold: 9, Accuracy: 1  Mean predictive accuracy: **0.96** |

* + **Compare the results.**

On average the two different branching criteria appear to perform about the same, with a mean accuracy of .96 over 10 folds on the Iris data set. Using Information Gain the algorithm produced 5 perfect run where as the Accuracy driven algorithm produced 6. In the end they appear equally effective for training on the Iris data set.

**Repeat the experiment with the**[**Voting**](http://dml.cs.byu.edu/~cgc/docs/mldm_tools/Assignments/Datasets/voting.arff)**problem.**

|  |  |
| --- | --- |
| **Information Gain** | **Accuracy** |
| Node: Property: 'physician-fee-freeze'  Node: Property: 'synfuels-corporation-cutback'  Node: Property: 'crime'  Node: Property: 'anti-satellite-test-ban'  …  Node: Property: 'anti-satellite-test-ban'  Node: Class: 0  Node: Property: 'el-salvador-aid'  …    Node: Class: 0  Node: Property: 'synfuels-corporation-cutback'  Node: Property: 'duty-free-exports'  N: P: 'adoption-of-the-budget-resolution'  …  Node: Property: 'immigration'  …  Node: P: 'adoption-of-the-budget-resolution'  Node: Property: 'el-salvador-aid'  Node: Class: 0  Node: Property: 'immigration'  Node: P: 'superfund-right-to-sue'  Node: Class: 0  N: P: 'anti-satellite-test-ban'  …  Node: Class: 1  Node: Property: 'anti-satellite-test-ban'  Node: Class: 0  Node: Class: 1 | Node: Property: 'physician-fee-freeze'  Node: Property: 'handicapped-infants'  Node: Property: 'religious-groups-in-schools'  Node: Class: 0  Node: P: 'water-project-cost-sharing'  …  Node: Property: 'water-project-cost-sharing'  Node: Property: 'anti-satellite-test-ban'  …  Node: Class: 0  Node: Property: 'handicapped-infants'  Node: Property: 'water-project-cost-sharing'  N: P: 'adoption-of-the-budget-resolution'  N: P: 'religious-groups-in-schools'  …  Node: Property: 'el-salvador-aid'  Node: Class: 1  N: P: 'religious-groups-in-schools'  …  Node: Property: 'el-salvador-aid'  …  Node: P: 'synfuels-corporation-cutback'  Node: Class: 1  Node: Property: 'mx-missile'  Node: P: 'water-project-cost-sharing'  …  Node: P: 'water-project-cost-sharing'  … |

* + **Compare this tree with the one obtained with information gain as the splitting criterion.**
  + **Evaluate predictive accuracy using 10-fold cross-validation for information gain and accuracy.**

|  |  |
| --- | --- |
| Dataset name: ../data/voting.arff  Number of instances (rows): 435  Number of attributes (cols): 17  Learning algorithm: DecisionTree  Evaluation method: cross  Rep: 0, Fold: 0, Accuracy: 0.883721  Rep: 0, Fold: 1, Accuracy: 0.906977  Rep: 0, Fold: 2, Accuracy: 0.976744  Rep: 0, Fold: 3, Accuracy: 0.883721  Rep: 0, Fold: 4, Accuracy: 0.930233  Rep: 0, Fold: 5, Accuracy: 0.906977  Rep: 0, Fold: 6, Accuracy: 0.953488  Rep: 0, Fold: 7, Accuracy: 0.953488  Rep: 0, Fold: 8, Accuracy: 0.930233  Rep: 0, Fold: 9, Accuracy: 0.953488  Mean predictive accuracy: **0.927907** | **Using Accuracy:**  Dataset name: ../data/voting.arff  Number of instances (rows): 435  Number of attributes (cols): 17  Learning algorithm: DecisionTree  Evaluation method: cross  Rep: 0, Fold: 0, Accuracy: 0.930233  Rep: 0, Fold: 1, Accuracy: 0.906977  Rep: 0, Fold: 2, Accuracy: 0.976744  Rep: 0, Fold: 3, Accuracy: 0.906977  Rep: 0, Fold: 4, Accuracy: 0.930233  Rep: 0, Fold: 5, Accuracy: 0.930233  Rep: 0, Fold: 6, Accuracy: 0.930233  Rep: 0, Fold: 7, Accuracy: 0.860465  Rep: 0, Fold: 8, Accuracy: 0.883721  Rep: 0, Fold: 9, Accuracy: 0.930233  Mean predictive accuracy: **0.918605** |

* + **Compare the results.**
  + **Describe and justify the method you used to handle missing values.**

For each missing attributes I calculate the majority value for that attribute from the original data set. I then replace the missing value with the majority for that attribute. This seems reasonable given that the majority value, for a missing attribute, is a simple best guess for what it might be given our current data set.

**Extend your algorithm so that, when accuracy is the splitting criterion, it may use up to 2 conditions in the tests at each node (e.g., attrX = Vx and attrY = Vy). You may choose to make that an user-specified option.**

* + **Induce a decision tree using the entire dataset with this extended algorithm for both the**[**Iris**](http://dml.cs.byu.edu/~cgc/docs/mldm_tools/Assignments/Datasets/iris.arff)**problem and the**[**Voting**](http://dml.cs.byu.edu/~cgc/docs/mldm_tools/Assignments/Datasets/voting.arff)**problem. Give a visual representation of the trees.**

|  |
| --- |
| **Iris** |
| Node: Property: sepallength & sepalwidth  Node: Property: petallength  Node: Property: petalwidth  Node: Class: 0  Node: Class: 1  Node: Property: petalwidth  Node: Class: 1  Node: Class: 2  Node: Class: 0  Node: Property: petallength & petalwidth  Node: Class: 1  Node: Class: 2  Node: Class: 2  Node: Class: 2  Node: Property: petallength & petalwidth  Node: Class: 1  Node: Class: 2  Node: Class: 2  Node: Class: 2 |
| **Voting** |
| Node: Property: 'physician-fee-freeze' & 'superfund-right-to-sue'  Node: Property: 'handicapped-infants'  Node: Property: 'water-project-cost-sharing' & 'aid-to-nicaraguan-contras'  ~Class Nodes  Node: Property: 'el-salvador-aid' & 'religious-groups-in-schools'  Node: Property: 'adoption-of-the-budget-resolution' & 'crime'  ~Class Nodes  Node: Class: 0  Node: Class: 0  Node: Class: 0  Node: Property: 'handicapped-infants' & 'crime'  Node: Property: 'education-spending'  Node: Class: 0  Node: Class: 1  Node: Property: 'water-project-cost-sharing'  Node: Class: 0  Node: Property: 'adoption-of-the-budget-resolution' & 'mx-missile'  Node: Class: 0  Node: Class: 0  Node: Class: 0  Node: Property: 'el-salvador-aid' & 'synfuels-corporation-cutback'  …  Node: Class: 0  Node: Class: 0  Node: Property: 'water-project-cost-sharing' & 'adoption-of-the-budget-resolution'  ~Class Nodes  Node: Property: 'adoption-of-the-budget-resolution' & 'synfuels-corporation-cutback'  Node: Property: 'water-project-cost-sharing' & 'duty-free-exports'  ~Class Nodes  Node: Property: 'anti-satellite-test-ban' & 'education-spending'  Node: Property: 'handicapped-infants'  …  Node: Property: 'handicapped-infants'  …  Node: Class: 0  Node: Class: 1  Node: Property: 'anti-satellite-test-ban' & 'immigration'  ~Class Nodes  Node: Property: 'water-project-cost-sharing'  Node: Class: 1  Node: Class: 0 |

* + **Compare them with those obtained above.**
  + **Explain why it may be necessary to thus extend the decision tree learning algorithm when using accuracy as the splitting criterion (and why the extension is of little value when information gain is the splitting criterion).**