Adnan Fazeli
Casey Rosenthal
October 29st, 2009
COS 472
Artificial Intelligence

Project 3 due 20091029

Introduction

Decision trees can be used to classify data based on item attributes. We investigated three methods of creating a decision tree. The decision trees are split on nodes that are selected as having the highest information gain, in alphabetical order, or representative of the most items. We call the last algorithm the "Fattest Value" algorithm.

Methodology

The information gain algorithm calculates the entropy of nodes in a potential split and uses that to infer that a node split which ends up closest to having the class on one side and everything else on the other would make for the shortest, possibly the most efficient decision tree. The alphabetical algorithm iterates through all of the values available for splitting, and selects the one which is smallest in ASCII. The fattest value algorithm iterates through all of the values available for splitting, and selects the one which represents the most items in the data set.

Information gain is a proven algorithm. The alphabetical algorithm is trivial, and almost random in that the node chosen for splitting has presumably nothing to do with the classification of items. This algorithm is similar to an experimental control, because we do not expect it to perform well. It is easy to calculate, and will provide a data point for comparison. The fattest value algorithm was chosen because we do not expect to find another algorithm that performs better than the information gain algorithm, but we thought we might be able to construct a decision tree faster [computationally] and shorter. By splitting on the node that represents the most items, we thought we might get fat trunks to our decision tree quickly, and then end up with many leaves.

In order to create our decision tree, we needed test data. We take each data set, and randomly pop a data item until 20% of the data has been separated. This 20% is used as training data, and is consumed in the construction of the decision tree. The other 80% of the data is then processed, and each item

classification prediction is compared with the actual classification, which is used to estimate the accuracy of the decision tree.

Because the training data is chosen at random, each decision tree is created from different data. This gives us different decision trees, each of which performs with a different level of accuracy given the remaining test data. We use the accuracy of the decision trees to compare the different algorithms, and a statistical function to tell us whether the differences in the accuracy of the different methods is significant.

Data

Our data was selected and formatted to test our classification system. We chose two datasets from the UCI Machine Learning Repository. The first is a set of Congressional voting records including the House Representative's party affiliation and vote recorded for 16 key votes. 435 representatives are included in the dataset, which we modified slightly to be read by our program. The second is data collected on breast cancer. This data contains 9 attributes of the tumor or patient as well as whether or not the tumor is classified as malignant. 699 examples were included in this dataset, which was also modified so that it could be easily consumed by our program.

Results

House Voting Data	Total Prediction Accuracy					
Metric	Data 1	Data 2	Data 3	Data 4	Data 5	Average
Information Gain	92.24	92.82	91.95	93.68	93.1	92.76
Alphabetical	71.26	85.06	84.77	82.47	75.57	79.83
Fattest Value	92.53	92.53	88.22	94.83	80.46	89.71

Table 1

Table 1 contains the prediction accuracy for items that were properly chosen as either in-class or not-in-class for the three decision tree methods trained and tested using the Congressional voting data.

House Voting Data	Class (Republican) Prediction Accuracy					
Metric	Data 1	Data 2	Data 3	Data 4	Data 5	Average
Information Gain	84.67	94.82	94.91	93.64	96.19	92.85
Alphabetical	58.29	84.98	87.14	83.62	61.83	75.17
Fattest Value	88.08	95.59	84.06	91.55	83.93	88.64

Table 2

Table 2 contains the prediction accuracy for items that were properly classified as Republicans for the three decision tree methods trained and tested using the Congressional voting data.

Cancer Data	Total Prediction Accuracy					
Metric	Data 1	Data 2	Data 3	Data 4	Data 5	Average
Information Gain	91.43	93.93	93.57	91.25	93.93	92.82
Alphabetical	91.25	95.18	93.04	93.39	94.64	93.5
Fattest Value	94.29	92.5	92.86	93.75	91.79	93.04

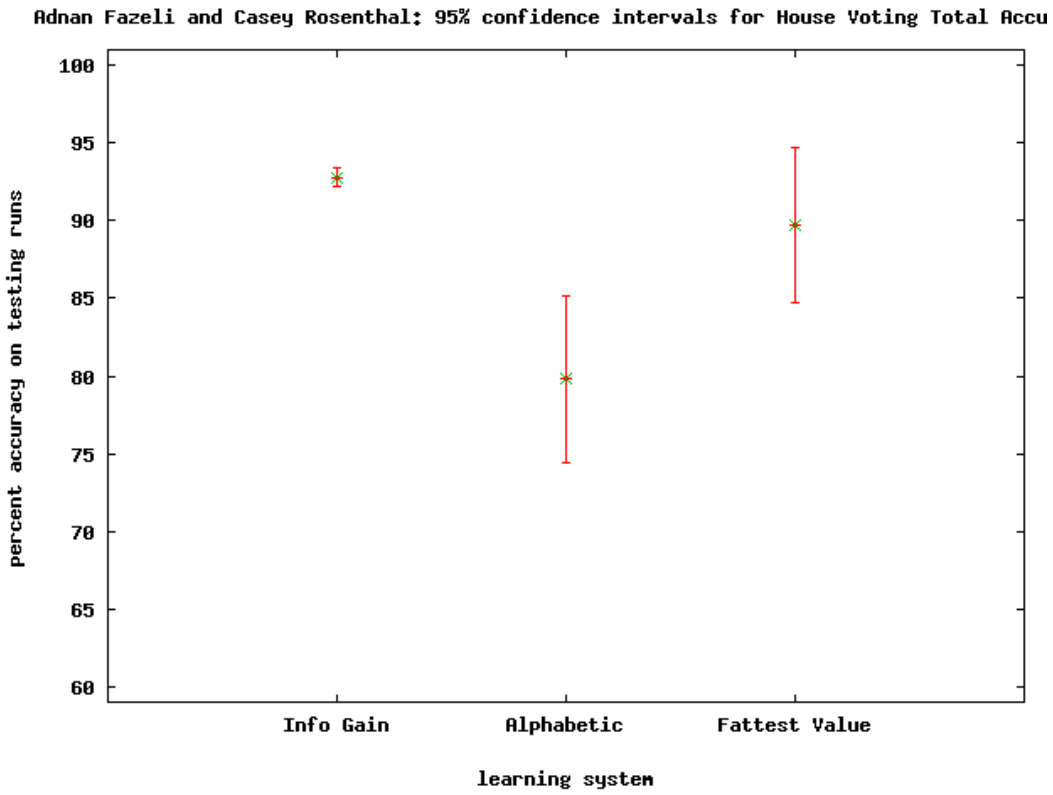
Table 3

Table 3 contains the prediction accuracy for items that were properly chosen as either in-class or not-in-class for the three decision tree methods trained and tested using the breast cancer malignancy data.

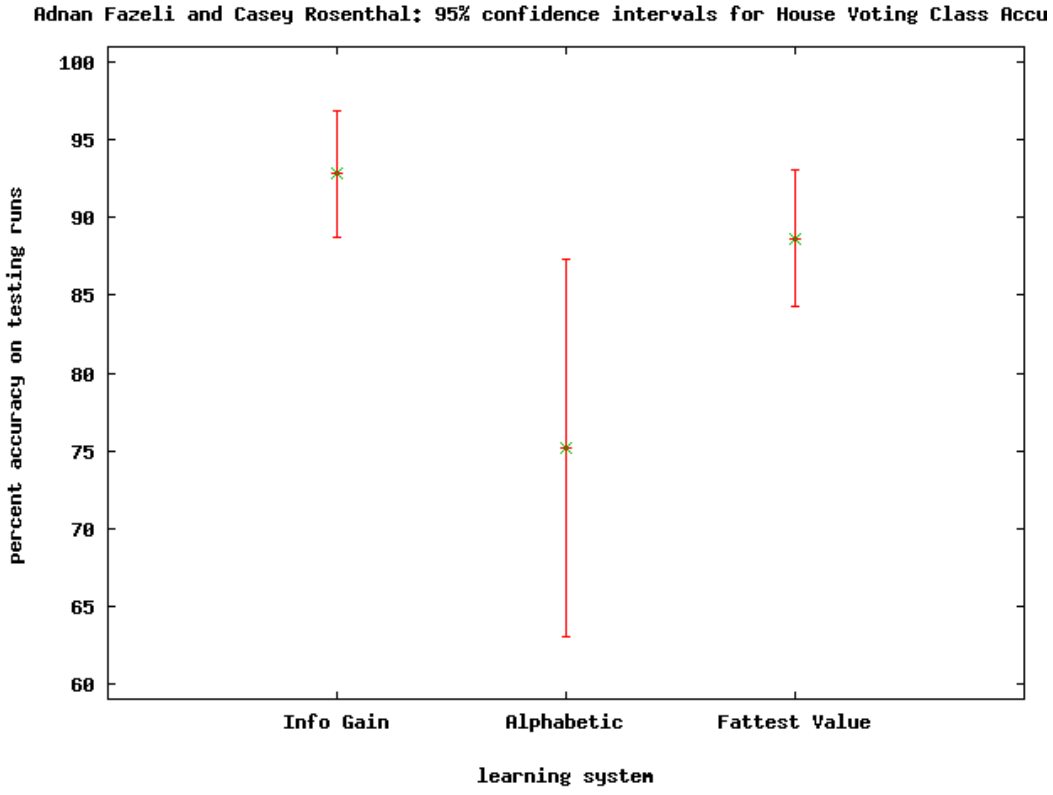
Cancer Data	Class (Malignant) Prediction Accuracy					
Metric	Data 1	Data 2	Data 3	Data 4	Data 5	Average
Information Gain	84.47	93.6	97.95	84.54	89.32	89.98
Alphabetical	90.36	90.82	92.54	87.62	87.94	89.86
Fattest Value	88.24	93.68	96.35	90.52	97.15	93.19

Table 4

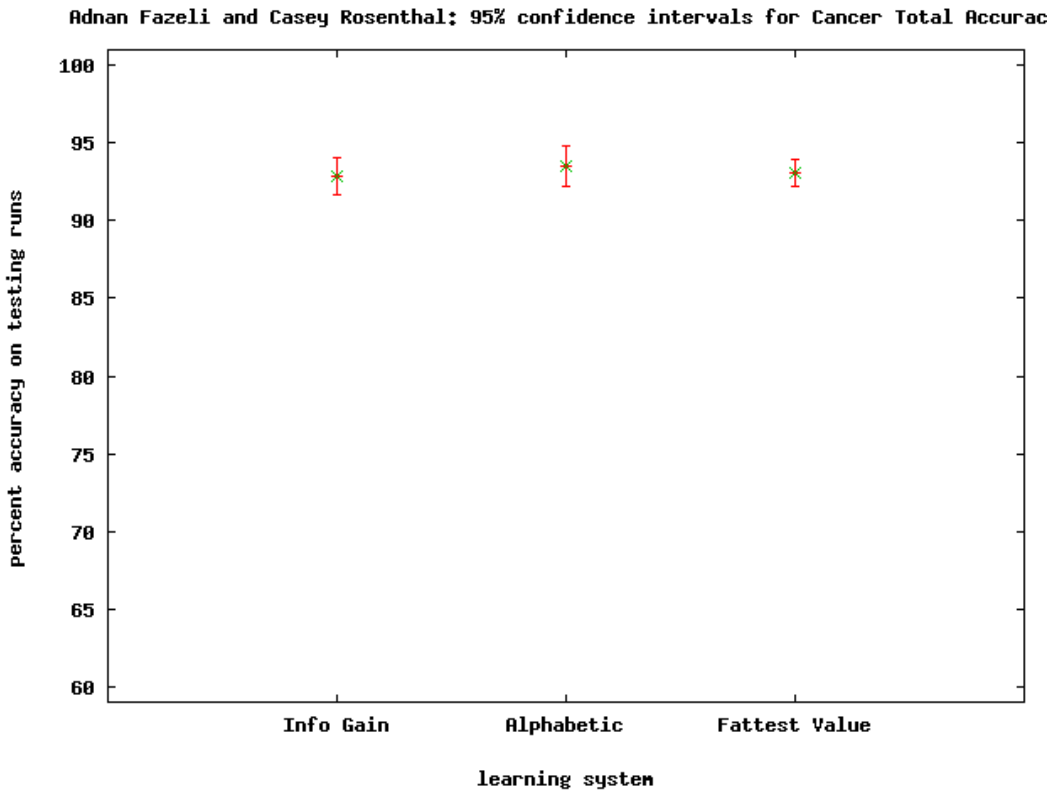
Table 4 contains the prediction accuracy for items that were properly classified as malignant for the three decision tree methods trained and tested using the breast cancer malignancy data.



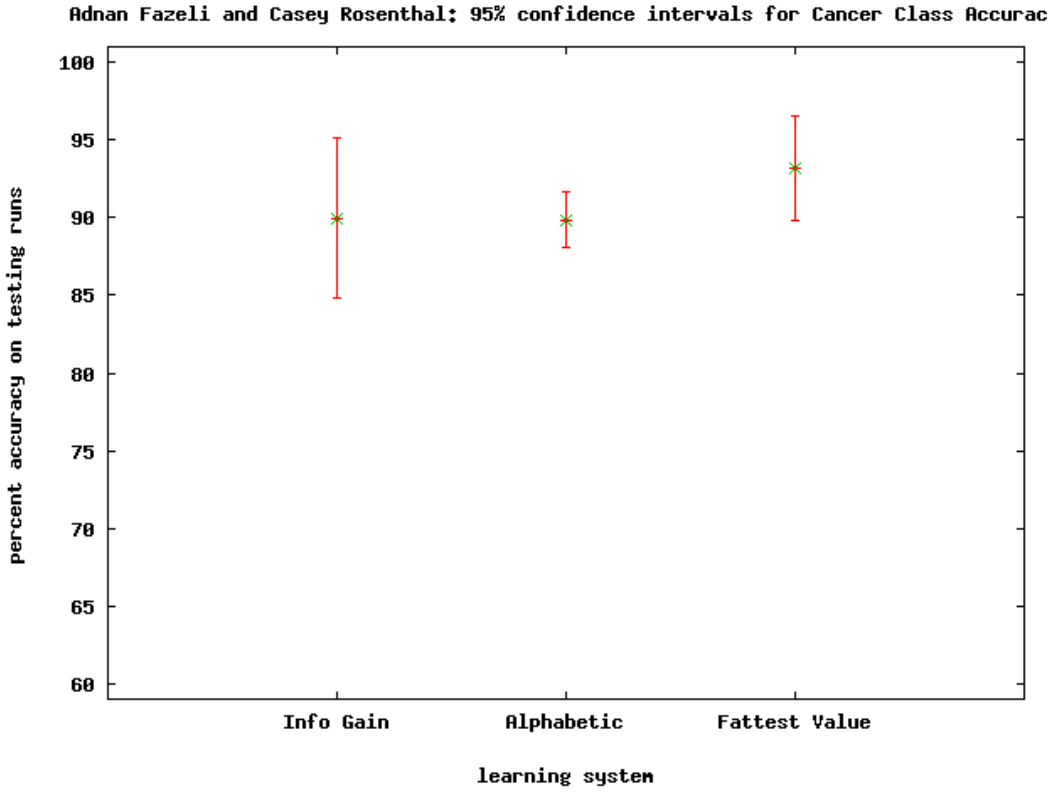
Graph 1



Graph 2



Graph 3



Graph 4

In addition to the tables and charts above, we computed whether the difference between any of the three methods was statistically significant. The following differences were significant.

Total predictive accuracy for Congressional voting decision tree:
Information Gain vs Alphabetical: significant to 99.5%
Alphabetical vs Fattest Value: significant to 97.5%

Classified predictive accuracy for Congressional voting decision tree:
Information Gain vs Alphabetical: significant to 97.5%
Alphabetical vs Fattest Value: significant to 95%

Total predictive accuracy for cancer malignancy decision tree:
none

Classified predictive accuracy for cancer malignancy decision tree:
Information Gain vs Fattest Value: significant to 90%
Alphabetical vs Fattest Value: significant to 90%

Conclusions

We can be pretty confident that both the information gain and the fattest value decision trees were more accurate than the baseline alphabetical algorithm when predicting the political affiliation of Representatives based on voting record. This is surprising, because it suggests that the information gain and fattest value algorithms were more or less equivalent in accuracy, although the information gain algorithm results in a significantly smaller tree.

We cannot conclude with great confidence that any of the algorithms performed better when classifying the cancer malignancy data; however, it appears from the charts that if we gathered more data, there is a possibility that the fattest value algorithm has a chance of out-performing the information gain algorithm in terms of accuracy.

Future Work

Since we are unable to say with confidence which algorithm is most accurate at this point, future research will investigate the same three algorithms with larger data sets for longer runs. This will hopefully help us establish a stronger statistical confidence. We will also run these same algorithms on different data sets to see if some perform better with different types of data, since our research here indicates that we may have difference in accuracy between the methods applied to the voting data as opposed to the methods applied to the cancer malignancy data.