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

Project 3 Appendix due 20091029

Decision trees have different predictive capabilities that depend on many factors. The differences in our trees seemed large when comparing the printout of each next to each other, but the difference in accuracy between one and the next was not very large. Clearly the training data used had a strong affect on the structure of the resulting decision tree, but it did not seem to have as strong an effect on the predictive accuracy of the tree as a whole.

Our very simple fattest-value algorithm performed better than we expected. It seems very close to the highest information gain algorithm, if not better in some cases. Computationally, the fattest-value algorithm is extremely fast to compute the decision tree, although it tends to produce much larger trees. Since creation of the tree is more intensive than classification through the tree, our algorithm might be preferable to the highest information gain function in cases where a data set needs to be analyzed quickly, and accuracy is perhaps not as important.

The two data sets that we used seemed to indicate different predictive accuracy for the algorithms. This was not established as statistically significant at this small number of runs, but it leaves open the possibility that the highest information gain algorithm is more accurate for voting data, while the fattest value algorithm could be more accurate for the cancer malignancy data. This implies that there are other hidden factors that we have not uncovered in this research paper that make one algorithm more accurate than another based on the type of data being classified. We do not know what attribute of the data is more susceptible to accurate classification under one algorithm as opposed to the other, but it must be something intrinsic to the data itself; such as, perhaps a hidden variable.