**NLP Project Report**

CMPT 310

Simon Fraser University

Fall 2018

Kevin Park, Michael Wong, Alan Ou

**Decision Tree Learning Analysis**

**Methods and Results: Effects of the Confidence Factor**

We tested the J48 classifier with confidence factors ranging from 0.001 to 0.99, and all other options kept at default; this includes keeping the cross-validation at 10 folds and the minimum number of instances per leaf at 2.

Our results, as shown in the diagram on the left, show that as confidence factor increases, the number of leaves and the size of the tree increases as well; however, the percentage of correctly identified instances increases to a certain point where it then begins to decrease. The performance measure of the program, the percentage of correctly identified instances, starts low at low confidence factors, reaches a maximum at confidence factor 0.25, and starts decreasing as confidence factor continues increasing.

**Methods and Results: Effect of Modifying the Minimum Number of Instances per Leaf**

For testing the effect that changing the minimum number of instances per leaf had on the percentage of correctly identified instances, we chose minNumObj to range from 1 to 10, and again, all other options kept at default.

Our results, as shown in the diagram on the right, show that as the minimum number of instances per leaf increases, the size of the tree decreases, as expected; however, like changing the confidence factor, the percentage of correctly identified instances increases to a certain point, reaches a peak, then begins to decrease. The percentage of correctly identified instances again, starts low when minNumObj is 1, reaches a maximum when minNumObj is 2 and starts decreasing as minNumObj continues increasing.

**Methods and Results: Other J48 Classifier Options**

Further testing of different J48 classifier options was conducted. While modifying batchSize, binarySplit, collapseTree, doNotMakeSplitPointActualValue, saveInstanceData, useLaplace, and useMDLcorrection did not change the performance measure, other options did. Options causing change are shown below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Option | Option Modified to | Percentage of Correctly Identified Instances | Number of Leaves | Size of Tree |
| reducedErrorPruning | True | 84.93% | 75 | 149 |
| subtreeRaising | False | 85.93% | 100 | 199 |
| unpruned | True | 84.80% | 204 | 407 |

**Results: Optimal Options for J48 Classifier**

Our results show that the J48 classifier performs optimally at default options. By leaving the confidence factor at 0.25 and the minimum number of instances per leaf (minNumObj) at 2, an optimal 86.43% correctly identified instances is observed. Any modifications decrease the performance of the program, as shown in the above charts.

**Insights**

In our model, the root node is the word “share", however, it is not the most informative feature. While “share” did correctly classify all 170 of the 3000 news articles as fake, it did not classify as many articles as “said”. “Said” correctly identified 312 news articles as real while incorrectly classifying 6 articles as real when they were fake.

While some conjunctions surprisingly make sense as in this case “share, nomination, sen, email,” many do not make sense; for example, the conjunction “share, s, I, US, GOP” does not make any sense.

**Common Incorrectly Identified Cases:** Two common cases that the tree gets wrong are:

1. Commonly used words such as “Said,” which the program correctly used to identify 213 articles as fake, and incorrectly classifying 38; and “Monday” which was correctly used to identify 57 articles as fake, and incorrectly classifying 17.

**Reason:** One possible reason for commonly used words accounting for many of the incorrect classifications may be because these words are simply too common. These are words that we use every day in settings where we are lying and also in settings where we are telling the truth.

1. Single and double letter features such as “s”, “t”, and “re” were each used to correctly identify 242, 32 and 47 articles, and incorrectly classifying 14, 10 and 4 articles respectively.

**Reason:** Single and double letter features were also common cases the tree got wrong because single and double letter features are common components in many words. For example, to give possession to a person, we add a ‘s to a person’s name; similarly, we add ‘t to certain contractions.