**Decision Tree Learning Algorithms**

**Results**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Accuracy** | **Precision** | **Recall** | **F-Measure** | **Run-Time** |
| **ID3** | 0.8103925 | 0.5974326 | 0.6050442 | 0.60121435 | 93 secs |
| **Random Forest**  **(with 10 trees)** | 0.8108839 | 0.596381 | 0.6170047 | 0.60651755 | 3 secs |
| **Random Forest**  **(with 50 trees)** | 0.82722193 | 0.6373305 | 0.62324494 | 0.63020897 | 15 secs |
| **Random Forest**  **(with 100 trees)** | 0.8296173 | 0.6395107 | 0.63884556 | 0.6391779 | 33 secs |
| **Pruned Tree**  **(1 node pruned)** | 0.82396656 | 0.6394422 | 0.58424336 | 0.61059785 | 426 secs |
| **Pruned Tree**  **(2 nodes pruned)** | 0.83090717 | 0.66231066 | 0.5798232 | 0.618328 | 809 secs |
| **Pruned Tree**  **(3 nodes pruned)** | 0.8332412 | 0.68608093 | 0.5421217 | 0.6056645 | 1041 secs |
|  |  |  |  |  |  |

**Interpretation of results**

* As can be clearly seen, ID3 by itself does not give a very high accuracy when tested on validation data due to overfitting.
* This accuracy is increased in the case of both, random-forests and reduced error pruning.
* Further, the more trees we consider in random-forests, the more the accuracy of the result is increased. This can be seen when number of trees change from 10 to 50 to 100.
* In the case of reduced error pruning the accuracy keeps increasing as we prune more nodes. But, to get the final pruned tree (after having pruned all possible nodes), a lot of time is consumed.
* Hence, in reduced error pruning we have mentioned only 3 cases for comparison, when 1, 2 and 3 nodes have been pruned and the algorithm is stopped. As we can clearly see, even for pruning just 3 nodes, a LOT of time is consumed.

**Applications ID3**

* ID3 is used in computer crime forensics.
* ID3 is ideal for situations where the dataset is not large enough to make use of tree pruning.
* When less data is available, we cannot create a separate validation set. If this is done the training examples used for ID3 will decrease significantly.
* Some applications are finding the cause of equipment malfunctions, medical diagnosis, etc.

**Applications of Random Forest**

* ID3 tree tends to overfit the training data giving us incorrect predictions.
* In such cases, random forest is a good solution.
* Random forest does not need a large dataset as training data and additionally gives the result as the mode of many trees generated randomly on the same dataset.
* Random forests are used in radar, LiDAR , thermal remote sensing imagery and other applications in remote sensing.

**Applications of Reduced-Error Pruning**

* Reduced error is one of the ideal solutions when a large dataset of training examples is available.
* About one third of the training examples are randomly selected to get reserved as the validation set and the other two-thirds is used to train and form the tree using ID3.
* This is an effective method to deal with overfitting in ID3, as the final tree in ID3 is pruned till the point where further pruning will decrease its accuracy over the validation set.
* It is one of the leading techniques used in artificial neural networks.
* One direct application is “traffic incident detection” from previously available dataset.

**Team Members**

1. Siddhant Shenoy (2015A7PS0057H)

2. Rohith Pulipaka (2015AAPS0153H)

3. Kriti Goyal (2015A7PS0100H)