Image Prediction

Neha Pradhan
Rakesh Buchan Krishna Murthy

Random Forest

- Random Forests are a combination of tree predictors where each tree depends on the values of a random vector sampled independently with the same distribution for all trees in the forest.
- The basic principle is that a group of “weak learners” can come together to form a “strong learner”.
- Random Forests attempts to mitigate the problems of high variance and high bias by averaging to find a natural balance between the two extremes.

Model Implementation

Parameters are tuned as follows:

- numClasses = 2 (1 or 0 – being the possible labels to predict)
- numTrees = 40 (the number of trees used to make a decision, majority predicts the label)
- featureSubsetStrategy = "auto" (is set to “all”, if numTrees = 1 else will be set to “sqrt”)
- impurity = "gini" (allows split at node if information gain > 0.0)
- maxDepth = 15 (the depth of each tree, more the depth of the tree more are the feature considered during splitting)
- maxBins = 60 (allows to categorize continuous features and sets the conditions for splitting at each node)

Analysis

- Feature extraction strategy
  - We assumed that the pixel present at the center column (1543) of each record was the main factor contributing to the label. So we filtered center column and it's nearest neighbors up to two levels.
  - Overfitting the model was the concern.

- Working on Sampled Data set
  - We reduced the training data set to 11GB in order to train the model faster with high values for numTrees and maxDepth.
  - Inadequate data to train the model was the concern.

Experiments

- Tests performed with sampled data set

numTrees	maxDepth	maxBins	Accuracy
20	10	40	0.976931
30	10	40	0.980552
40	10	40	0.981847
15	7	40	0.964912
15	8	40	0.970588
15	11	40	0.953704
15	15	40	0.955752
15	20	40	0.955682

Experiments

- Tests performed with feature extractions

numTrees	maxDepth	maxBins	Worker machines	filtered Data set	Execution Time(in min)	Accuracy
80	15	250	10	First level neighbors	53	0.997294408
80	15	250	10	First & Second level neighbors	63	0.997421152
75	15	250	20	First & Second level neighbors	33	0.997382311

Experiments

- Tests performed with all features

numTrees	maxDepth	maxBins	Worker machines	Execution Time (in min)	Accuracy
40	15	40	20	59	0.99750599
45	15	40	20	62	0.99749065
50	15	40	20	69	0.99749576
40	15	60	20	53	0.99754278
40	15	80	20	62	0.99757856
40	15	100	20	67	0.99758469
40	15	140	20	80	0.99759695

Results

numTrees	maxDepth	maxBins	Worker machines	Execution Time (in min)	Accuracy
40	15	140	20	80	0.99759695

- Running Times:

Job	Workers	Running Time (in mins)
Training	20	80
Prediction	20	9

Thank You!!