CS 6240: Parallel Data Processing in MapReduce

Project Presentation

ROHIT PATNAIK | SANKAR GIREESAN NAIR

Foreground-Background Prediction

Approaches Used:

- Ensemble of Random Forest, Logistic Regression and Gradient Boosted Trees Models
- Ensemble of 3 Random Forest Models
- Ensemble of 5 Random Forest Models

Random Forest Parameters

- numTrees – 12

  As the number of trees increased, the accuracy increased

  Runtime increased roughly linearly

- maxDepth – 10

  As the depth of the trees increased, accuracy was expected to increase.

  Accuracy reduced due to the possibility of overfitting

Experiments on Models

- Ensemble model with Gradient Boosted Trees, Logistic Regression, and Random Forest. It was done without sampling with replacement with each model getting 1/3rd of the Training data on AWS. Running this ensemble with small amount of data gave good accuracy, however, we couldn't train the GBT model on AWS with 1/3rd data because it took around 12-13 hours. When we ran the prediction code on AWS for the models we received for Logistic Regression and Random Forest individually, the overall accuracy for 1 was 66% and 78% respectively.
- Ensemble model with 3 Random Forest models. It was done without sampling with replacement with each model getting 1/3rd of Training data on AWS. The parameters set for the Random Forest model was numTrees=10, maxDepth=5, maxBins = 100. Predicting for a test file on the ensemble gave us a very low accuracy of ~60%.
- Ensemble model with 5 Random Forest models. It was done with sampling with replacement which increased overall Training Data by nearly ~1.5 times the original data. Each model received around 20% of the sampling data. The parameters set for the Random Forest model was numTrees=10, maxDepth=5, maxBins = 100. Predicting for a test file on the ensemble gave us an accuracy of 74.8%.

Accuracy Obtained

Models	Amount of Data for each model	Number of Trees	Max Depth	Max Bins	Accuracy*
3 Random Forest	33% of Training Data (without replacement)	10	5	100	75.04%
5 Random Forest	20% of Training Data (without replacement)	30	20	3000	78.88%
5 Random Forest	50% of Training Data (with replacement)	10	10	100	74.8%

*All the models gave good accuracy of 99.9 % for the validation data. However, the accuracy mentioned is for all the foreground values. It will be like **recall(tp/(tp+fn)).**

Scalability

Models	Number of Machines	Running Time(in mins)	Speed Up
Random Forest (3 RF)	11	104	
Random Forest (3 RF)	21	59	1.76
Random Forest (5 RF)	11	381	
Random Forest (5 RF)	21	240	1.59

Thank You! Questions?