6240-V PARALLEL PROCESSING WITH MAPREDUCE

FLIGHT DELAY PREDICTION

AUTHORS: Sujith Narayan, Karthik Chandranna

INTRODUCTION:

In this assignment we have been introduced to the classifier algorithms such as Random Forest in a MapReduce environment. This assignment deals with predicting the flight delays. We use the train dataset to construct a model using the Random forest algorithm provided by Weka and test is against a given dataset. We then validate the predicted results given by the model against a validation dataset. We use the comparison results to construct a Confusion Matrix.

IMPLEMENTATION DETAILS:

OVERALL DESIGN:

Our design involves 2 MapReduce jobs run subsequently. The first MapReduce job reads the training dataset and generates a Random Forest model which is stored in the HDFS. The second MapReduce job reads the test dataset and predicts the results using the generated model.

MODEL DETAILS:

We use the Random Forest classifier provided by Weka to build our model. The model requires a set of factors that determine the prediction. The factors that we have considered as suitable for our model are-

Year, Quarter, Month, DayofMonth, DayofWeek, Date, FlightNumber, Origin, Destination, ScheduledDepartureTime, ScheduledArrivalTime, ScheduledElapsedTime, Carrier, Distance and a set of synthesized factors which are DepartsFromChicago, IsHoliday, IsPopularSrcDest, IsBusyDay.

All these factors are used to predict if the flight is delayed or not.

FIRST MR JOB:

Mapper

This mapper class reads the values from the train dataset and creates an AirlineDetails object. We also synthesize a number of factors that are required for the model building. We then sanitize the object to remove bad data. Now we create a Writable object called AirlineCompositeKey which has only the required fields and synthesized factors. The mapper emits data, with the month as the Key and the AirlineCompositeKey as the value.

Reducer

The reducer builds the attributes required for the model generation. It then creates instances for each object in the Iterable list of AirlineCompositeKey. All the instances are then added to the training dataset. We then set the class index of the dataset to the IsDelayed attribute. A Random Forest object is then created with the number of trees set to 5. We then feed the training dataset to build the Random Forest. We store the built model on the Hadoop FileSystem.

SECOND MR JOB:

Mapper

This mapper class reads the values from the test dataset and creates an AirlineDetails object. We also synthesize a number of factors that are required for the prediction. Now we create a Writable object called AirlineCompositeKey which has only the required fields and synthesized factors. The mapper emits data, with the month as the Key and the AirlineCompositeKey as the value.

Reducer

The reducer builds the attributes required for prediction. It then creates instances for each object in the Iterable list of AirlineCompositeKey. All the instances are then added to the test dataset. We set the class index of the dataset to the IsDelayed attribute. Now we load the model from the filesystem. We classify each instance to predict the value of IsDelayed. The reducer emits the combination of <FL_NUM>_<FL_DATE>_<CRS_DEP_TIME> as the key and a Boolean value.

VALIDATION:

The obtained results are validated against the validation dataset and the confusion matrix is drawn

RESULTS:

N = 4671724	**Predicted**: TRUE	**Predicted**: FALSE
Actual: TRUE	1193824	1127649
Actual: FALSE	1048188	1302063

CONFUSION MATRIX

Based on the above seen number in the Confusion Matrix, the calculated percentages are –

Correct Predictions: 53.43%
Incorrect Predictions: 46.57%

NOTE:

- Program was executed on the AWS cloud using cluster of 10 m3:XLarge machines.
- The execution time was 6 minutes on the AWS cloud.