Classification using Spark MLlib

Team: Drashti Bhuta & Shantanu Kawlekar

bhuta.d@husky.neu.edu | kawlekar.s@husky.neu.edu

Why Spark?

Spark programs are easy to write.

Faster than Hadoop MapReduce

Support for scalable machine learning with Spark MLlib

Explore and learn

The challenge

- The dataset was skewed
- Selection of model
- Tuning the model
- Transformation and diversity
- Huge Dataset
- Running on AWS

Considered all the features for classification.

Research

We explored Linear Regression, Logistics Regression, Gradient Boosted Trees and Random Forest.

Explored parameters for tuning

Transformation, OverSampling and Feature Selection (Reduction)

Functions: sampleByKey | takeSample

Model Used

We have used Random Forest for training the image - 1,2,3 & 6

Methods we used in our experiment?

- RDDs
- LabelPoint and Vector
- Sampling
- Transformations (Rotation)

Experimentation

We have tried various combinations of configurations.

Results based on validation set Image4

No of trees	Max Depth	Number of Bins	Accuracy
200	5	32	0.9965
50	**10**	**32**	**0.9978**
20	5	32	0.9967
200	10	32	0.9976
30	5	32	0.9973
100	10	40	0.9973
50	**10**	**50**	**0.9979**

Aha!

Our discoveries

Results with validation set Image4

1. Increasing these parameters increases the running time.
2. Values for depth and numTrees should be in proper ratio to get better results
3. Number of bins increases the choices from which the model can select the best one at a split

Number of trees	**Max Depth**	**Number of Bins**	**Run time (mins)**	**No of Machines**
50	10	32	65	10
30	5	32	35	10
200	10	32	83	15
50	10	50	40	15
30	30	32	600	10

Takeaways

Spark helps to simplify the challenging and computationally intensive task of processing high volumes.

MLlib requires more documentation.

Speed is as equally important as achieving higher accuracy.

Limitations and tradeoffs

Conclusion

For this project, we have learnt that in order to achieve most accurate classification we need to consider the following:

- Decide if there is enough diversity and try to make the data more diverse
- Understanding the distribution and nature of data set to decide the model to choose
- The most important is selecting values for tuning parameters of the model to achieve a greater accuracy. Each parameter of the model affects the outcome significantly.

After experimenting with various configurations of random forest , the results help us conclude that model with 50 trees, 10 depth and 50 bins gives us the highest accuracy. We also learnt that ML and large scale distributed processing compliment each other and the scope of growth these fields have together .

What will we do next?

Our plan is to dig deeper into this field and do quality application oriented research work.