Apache Spark's Machine Learning Library (MLlib)

ANVESH NAYAN LINGAMPALLI[1]

[1] *School of Informatics and Computing, Bloomington, IN 47408, U.S.A.*
[*] *Corresponding authors: anveling@umail.iu.edu*

April 12, 2017

MLlib is a machine learning library that runs on top of Apache Spark. With Spark and MLlib, jobs that reference a number of predefined machine learning algorithms are used to build applications.

Keywords: MLlib, Apache Spark, machine learning

https://github.com/Anveling/sp17-i524/paper1/S17-IR-2016/report.pdf

1. INTRODUCTION

Apache Spark[1] is a open source processing engine with consists of elegant APIs, for performing efficient data analytics. It provides a framework to process big data which are diverse in nature. MLlib(Machine Learning Library) is Apache Spark's scalable machine learning library[2]. Spark has many advantages when compared to other technologies such as Hadoop and Storm. Hadoop[3] is a big data processing technology, which is proved to be a solution for processing large data sets. In cases involving machine learning or streaming data, Hadoop is not efficient. It requires other tools such as Mahout[4] or Storm[5] to process the data. This is the most important advantage that the Apache Spark has on Hadoop. Spark is faster in run times than Hadoop MapReduce[6].

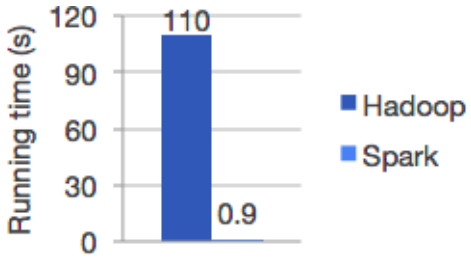

Fig. 1. Hadoop vs Spark

Apache Spark in addition to Map and Reduce functions, also supports SQL queries and machine learning. It has many libraries in Big Data analytics and Machine Learning domains. MLlib is one of the top level libraries that Spark offers. MLlib (Machine Learning Library) is Apache Spark's scalable machine learning library with APIs in Java, Python, R and Scala. It has the algorithms and tools for performing various tasks on the data such as, clustering, classification, regression and dimensionality reduction. The main goal of this library is to make machine learning easy.

2. HISTORY AND DEVELOPMENT

Development of MLlib began in 2012 as a part of MLBase project[7] (Kraska et al., 2013)[8]. It is an open source since September 2013. It has since then, been integrated into the Spark as an in-built package. The original version of MLlib was developed in UC Berkeley and provided a limited set of machine learning methods. Since it is an open source community, MLlib developed and now has additional functionality.

3. COMPONENTS OF MLLIB

MLib provides various linear models, Naive bayes[9] and decision trees[10] for classification and regression problems. With the help of these models, problems such as alternating least squares(ALS), k-means problem[11], PCA (principal component analysis)[12] for clustering have been successfully implemented. Text mining, predictive analysis of data are certain areas where MLlib is being used as an efficient tool.

MLlib has a package named spark.ml, which provides APIs for the functionality of the pipelines. This package enables users to swap the existing algorithms with their own algorithms[13].

MLlib supports various methods for binary classification, multiclass classification, and regression analysis. Each type of problem has its own supported algorithms. Binary Classification has Linear SVMs, Logistic regression[14], decision trees and naive Bayes. Multiclass Classification also has decision trees and naive Bayes as its supported algorithms. Regression has linear least squares[15], Lasso[16] and decision trees.

4. PERFORMANCE ANALYSIS BETWEEN MLLIB AND ITS ALTERNATIVES

Hadoop Mahout is one of the alternative choice for a machine learning library. Mahout uses Hadoop as underlying framework

whereas in the case of MLlib, it is Spark. In terms of features, support and performance MLlib performs better. In 2014, Mahout announced it would not accept Hadoop MapReduce and completely switched to Spark.

H2O[17], xgboost[18], python scikit-learn[19] are few other alternatives to MLlib. Scalability, speed and performance are measured for these tools and are shown in the table.

Tool	N (size of data)	Time(sec)	RAM(GB)	Accuracy
Python scikit-learn	10K	0.2	2	67.6
	100K	2	3	70.6
	1M	25	12	71.1
H2O	10K	1	1	69.6
	100K	1	1	70.3
	1M	2	2	70.8
	10M	5	3	71.0
Spark MLlib	10K	1	1	66.6
	100K	2	1	70.2
	1M	5	2	70.9
	10M	35	10	70.9

Fig. 2. Analysis of performance

For each tool and each size N, observations of the training tie, memory usage, and accuracy are presented. These tests have been carried out on a Amazon EC2 instance (32 cores, 60GB RAM)[20].

The graph for the results is shown below. H2O is memory efficient and faster than MLlib. But, MLlib is the better choice of the two as it has variety of functionalities.

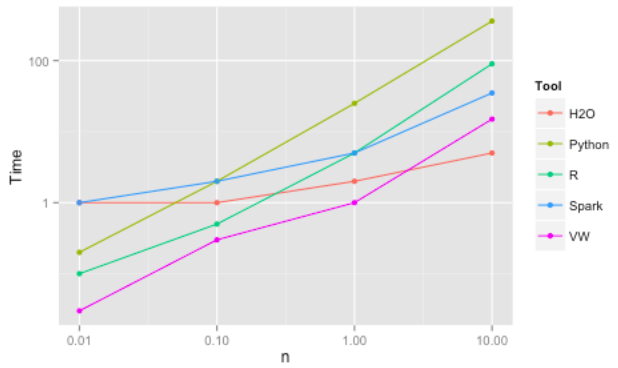

Fig. 3. Graph analysis

5. USE CASES

Apache Spark Machine Learning Library is used in wide range of applications in research and industry. Here two such applications are described briefly.

5.1. Movie Recommendation with MLlib

In this mini course project MLlib library is used to make personalized movie recommendations.[21]

5.2. Predict Telco Churn with Apache Spark MLlib

Churn prediction[22], is one of the most common applications of machine learning in the telecommunications industry, as well as many other subscriptions-based industries.MLlib is used here to fit a machine-learning model that can predict which customers of a telecommunications company are likely to stop using their service.[23]

6. USEFUL RESOURCES

[24] also has some good step by step tutorials on how to use Machine learning library to work on big data anlytics involving machine learning learning studio.

7. CONCLUSION

In conclusion, MLlib is a library used to perform machine learning as a part of big data analysis. It is designed for simplicity, scalability, and easy integration with other tools It is still in active development phase, and there have been many improvements over the previous versions over time. MLlib provides developers with a wide range of tools to make machine learning easy and scalable.

8. ACKNOWLEDGEMENTS

This work was done as part of the course "I524: Big Data and Open Source Software Projects" at Indiana University during Spring 2017.

REFERENCES

[1] "Apache spark," webpage. [Online]. Available: http://spark.apache.org/
[2] "Apache spark's mllib," webpage. [Online]. Available: http://spark.apache.org/mllib/
[3] "Apache hadoop," webpage. [Online]. Available: http://hadoop.apache.org/
[4] "Apache mahout," webpage. [Online]. Available: http://mahout.apache.org/
[5] "Storm," webpage. [Online]. Available: http://storm.apache.org
[6] "Hadoop mapreduce," webpage. [Online]. Available: https://en.wikipedia.org/wiki/MapReduce
[7] T. Kraska, A. Talwalkar, J. Duchi, R. Griffith, , M. J. Franklin, and M. Jordan, "Mlbase: A distributed machine-learning system," in *6th Biennial Conference on Innovative Data Systems Research(CIDR'13)*, Asilomer, CA, USA, 2013. [Online]. Available: http://cidrdb.org/cidr2013/Papers/CIDR13_Paper118.pdf
[8] "Mlbase project paper," webpage. [Online]. Available: http://cidrdb.org/cidr2013/Papers/CIDR13_Paper118.pdf
[9] "Naive bayes," webpage. [Online]. Available: https://en.wikipedia.org/wiki/Naive_Bayes_classifier
[10] "Decision trees," webpage. [Online]. Available: https://en.wikipedia.org/wiki/Decision_tree
[11] "K-means problem," webpage. [Online]. Available: https://en.wikipedia.org/wiki/K-means_clustering
[12] "Principal component analysis," webpage. [Online]. Available: https://en.wikipedia.org/wiki/Principal_component_analysis
[13] X. Meng, J. K. Bradley, B. Yavuz, E. R. Sparks, S. Venkataraman, D. Liu, J. Freeman, D. B. Tsai, M. Amde, S. Owen, D. Xin, R. Xin, M. J. Franklin, R. Zadeh, M. Zaharia, and A. Talwalkar, "Mllib: Machine learning in apache spark," *CoRR*, vol. abs/1505.06807, 2015. [Online]. Available: http://arxiv.org/abs/1505.06807
[14] "Logistic regression," webpage. [Online]. Available: http://www.statisticssolutions.com/what-is-logistic-regression/
[15] "Linear least squares," webpage. [Online]. Available: https://en.wikipedia.org/wiki/Linear_least_squares_(mathematics)

[16] "Lasso statistics," webpage. [Online]. Available: https://en.wikipedia.org/wiki/Lasso_(statistics)
[17] "H2o," webpage. [Online]. Available: https://en.wikipedia.org/wiki/H2O_(software)
[18] "xgboost," webpage. [Online]. Available: https://xgboost.readthedocs.io/
[19] "python scikit-learn," webpage. [Online]. Available: http://scikit-learn.org/stable/
[20] "Analysis of various machine learning packages," webpage. [Online]. Available: https://github.com/szilard/benchm-ml
[21] "Movie recommender using mllib," webpage. [Online]. Available: http://ampcamp.berkeley.edu/big-data-mini-course/movie-recommendation-with-mllib.html
[22] "Churn prediction," webpage. [Online]. Available: https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-azure-ml-customer-churn-scenario
[23] "Prediction of telco churn," webpage. [Online]. Available: https://blog.cloudera.com/blog/2016/02/how-to-predict-telco-churn-with-apache-spark-mllib/
[24] "Guide for mllib," webpage. [Online]. Available: https://spark.apache.org/docs/latest/ml-guide.html