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*

March 5, 2017

This paper provides a summary about Apache Sparks Machine learning library (MLlib) and its functionality.

Keywords: MLlib, Apache Spark, machine learning

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

This review document is provided for you to achieve your best. We have listed a number of obvious opportunities for improvement. When improving it, please keep this copy untouched and instead focus on improving report.tex. The review does not include all possible improvement suggestions and if you sea comment you may want to check if this comment applies elsewhere in the document.

Abstract: remove This paper, its clear that this is a a paper so you do not have to mention it. Furthermore, its actually a technology review. Abstract has grammar errors.

1. INTRODUCTION

Apache Spark

Citation

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. Spark has many advantages when compared to other technologies such as Hadoop and Storm. Hadoop

Citation

is also a big data processing technology

Grammar

which is proved to be a solution for processing large data sets. But

Grammar

it is not efficient in cases involving machine learning or streaming data. In these cases, Hadoop requires other tools such as Mahout

Citation

or Storm

Citation

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.[**?**]

A citation should be place when first introduced. Also, please place a comma after But l. 56

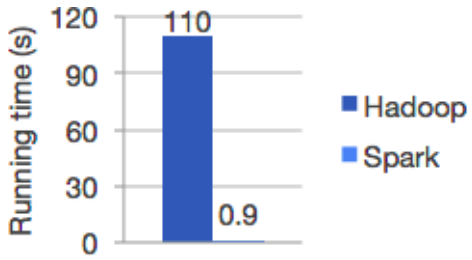


Fig. 1. Hadoop vs Spark

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 (Kraska et al., 2013)

Citation

. 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.

please provide the proper citation in latex for (Kraska et al., 2013)

3. COMPONENTS OF MLLIB

MLib provides various linear models, Naive bayes

Citation

and decision trees

Citation

for classification and regression problems. With the help of these models, problems such as alternating least squares(ALS), k-means problem

Citation

, PCA (principal component analysis)

Citation

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.[?]

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

Citation

, descision

Spelling

trees and naive Bayes. Multiclass Classification also has decision trees and naive Bayes as its supported algorithms. Regression has linear least squares, Lasso and decision trees.

please provide the proper citation for machine learning algorithms and check spelling in line 111

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

Citation

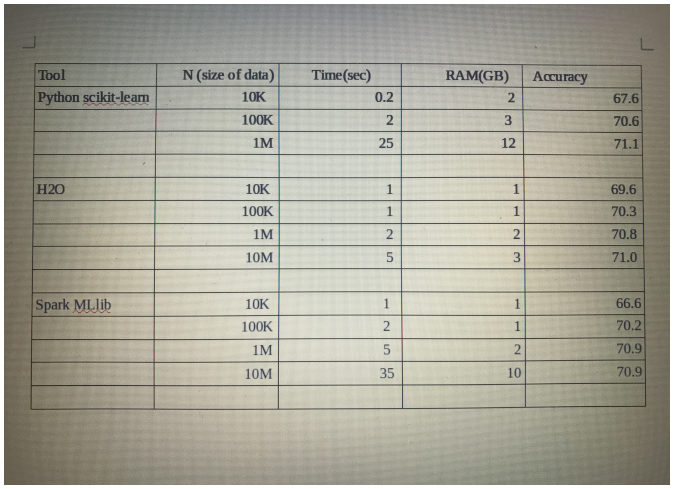
, xgboost

Citation

, python scikit-learn

Citation

are few other alternatives to MLlib. Scalability, speed and performance are measured for these tools and are shown in the table below.



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

please provide proper citations

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).[?]

The graph for the results is shown below. H2O is memory efficient and faster than MLlib. But

Grammar

MLlib is the better choice of the two as it has variety of functionalities.

please provide comma after But in line 144

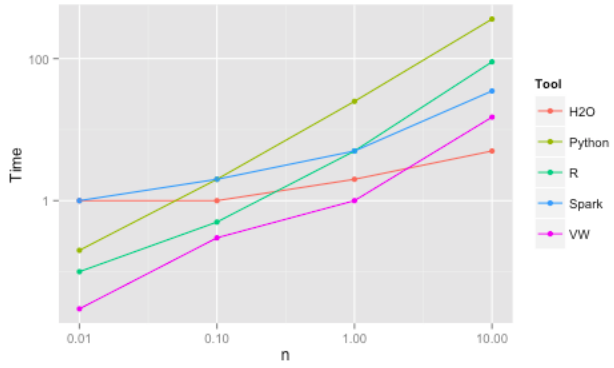


Fig. 3. Graph analysis

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.[?]

5.2. Predict Telco Churn with Apache Spark MLlib

Churn prediction

Citation

, 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.[**?**]

please provide proper citation

6. USEFUL RESOURCES

[**?**] 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 one of the best libraries to perform machine learning as a part of big data analysis. 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.

please consider revising MLIb is one of the best libraries to perform machine learning – this is the opinion of the author

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.