Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 113 (2017) 539–544

Procedia Computer Science

www.elsevier.com/locate/procedia

The 4th International Symposium on Emerging Information, Communication and Networks (EICN 2017)

An Empirical Evaluation of Intelligent Machine Learning Algorithms under Big Data Processing Systems

Dima Suleiman[a,b,*], Malek Al-Zewairi[a], Ghazi Naymat[a]

[a]*Computer Science Department, King Hussein Faculty of Computing Sciences, Princess Sumaya University for Technology, Amman 11941 Jordan P.o.Box 1438 Al-Jubaiha, Amman, Jordan*

[b]*Teacher at the University of Jordan*

Abstract

The rapid increase in the magnitude of data produced by industries that need to be processed using Machine Learning algorithms to generate business intelligence has created a dilemma for data scientists. This is due to the fact that traditional machine learning platforms such as Weka and R are not designed to handle data with such Volume, Velocity and Variety. Several machine learning algorithms and associated toolkits have been built specifically to work with big data; however, their performance is yet to be evaluated to allow researchers to get the most of these platforms. In this paper, the authors intend to provide an empirical evaluation of two emerging machine learning platforms under big data processing systems namely, H2O and Sparkling Water, by performing an experimental comparison between the two platforms in terms of performance over several generalization error metrics and model training time using the Santander Bank Dataset. Up to the authors' knowledge, this is the first time such a study is conducted. The evaluation results showed that the H2O platform has significantly outperformed the Sparkling Water platform in terms of model training time almost by fifty percent, while achieving convergent results.

Peer-review under responsibility of the Conference Program Chairs.

Keywords: Big Data; H2O; Sparkling Water; Prediction; Spark; Santander Bank Dataset;

* Corresponding author. Tel.: +962-6-5359949; fax: +962-6-5347295.
E-mail address: dimah_1999@yahoo.com

1877-0509
Peer-review under responsibility of the Conference Program Chairs.
10.1016/j.procs.2017.08.270

1. Introduction

It has been predicted that the amount of data that will be available in 2020 will be ten times more than what was available in 2013, and it might reach forty-four zettabytes according to the International Data Corporation's annual Digital Universe study[1]. This proliferation of data has led to the blossom of Big Data science. In general, Big Data concept refers to the data that is huge, complex and/or has several formats to be processed by a single computing machine. For a problem to be contemplated as a big data problem, it must have at least one of the three features of big data, that is, Volume, Velocity and Variety, which often referred as the 3Vs. The first V (i.e. volume) means that the data come in huge size; thus, it cannot be processed using simple tools or commodity computers. The second V is the Velocity and it refers to the speed in which the data is collected. Often, big data comes with high velocity. Finally, the third V, which refers to Variety, and it means that data comes in different format, i.e. structured data, such as data in tabular format, semi-structured data, such as an XML and unstructured data, such as multimedia files[2].

The rapid increase in the magnitude of data produced by many industries (e.g. IoT sensors, user's click stream, etc.) that need to be processed by machine learning algorithms creates a dilemma for researchers since traditional machine learning tools are not designed to handle this amount of data. The main purpose of machine learning is to use the knowledge from the past in order to learn how to make educated guesses in the future. In general, machine learning workflow consists of building a model, then making the necessary tuning after making evaluations to achieve the appropriate results, finally using the model to make predictions[3,4]. Moreover, Deep Learning algorithms are closely related to Artificial Intelligence. It aims to analyze and learn complex problems in order to make decisions similar to what the human brain can do[5].

Although, big data combined with machine learning have opened the door for unique research opportunities in several areas such as healthcare, users' behavior analysis and threat intelligence; it has been proven that traditional machine learning toolkits such as Weka and R cannot handle the large proliferation of data that came with big data. Therefore, a new generation of data processing systems has emerged to handle big data, such as Hadoop, Spark, H2O, Sparkling Water and Steam.

Hadoop (Apache Hadoop) is an open source implementation of MapReduce processing engine designed to distribute the processing of large datasets using clusters of commodity computers[6]. On the other hand, MapReduce is a programming model, which takes large task and divides it into subtasks. However, the reason for this division is to produce the results faster by enabling the subtasks to be done in parallel[7].

Similar to Hadoop, Spark also supports iterative computations in cluster environment. However, it features in-memory computations, making it process data much faster than its competitor technologies. In Spark, the main abstraction is Resilient Distributed Datasets, which used to store data in memory[2].

H2O is an open source platform that provides libraries for machine learning, parallel processing engines, scalable and fast deep learning, math, and data analytics, in addition to providing tools to facilitate the processing of data and building evaluations. On the other hand, Sparkling Water takes advantages of H2O and Spark by making combination between them. In addition, Sparkling water can provide fast, ideal and scalable machine learning platform of H2O to developers in order to use them in their applications[8].

One of hot research area is the personalized product recommendations where the individual user shopping behavior, habits and activities are tracked, recorded and analyzed to provide service providers with a better understanding of their users' preferences. However, it is a challenge to create the prefect prediction model available online. Therefore, this allows service providers to deliver customized, targeted ads based on the users liking, which improves the overall user experience and increases the likelihood of purchasing extra products. These problems mandate the intervention of intelligent machine learning algorithms such as deep learning accompanied with big data processing systems.

Several machine learning algorithms and associated toolkits were built specifically to work with big data problems such as the personalized product recommendations problem. Nonetheless, their performance is yet to be evaluated. In this paper, the authors provide an empirical evaluation study of two machine learning platforms under big data processing systems (namely; H2O and Sparkling Water). Both platforms are evaluated against a publicly available