Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 113 (2017) 559–564

Procedia Computer Science

www.elsevier.com/locate/procedia

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

Evaluation of classification algorithms for banking customer's behavior under Apache Spark Data Processing System

Wael Etaiwi*, Mariam Biltawi and Ghazi Naymat

Princess Sumaya University for Technology, Amman. Jordan

Abstract

Many different classification algorithms could be used in order to analyze, classify or predict data. These algorithms differ in their performance and results. Therefore, in order to select the best approach, a comparison studies required to present the most appropriate approach to be used in a certain domain. This paper presents a comparative study between two classification techniques namely, Naïve Bayes (NB) and the Support Vector Machine (SVM), of the Machine Learning Library (MLlib) under the Apache Spark Data processing System. The comparison is conducted after applying the two classifiers on a dataset consisting of customer's personal and behavioral information in Santander Bank in Spain. The dataset contains: a training set of more than 13 million records and a testing set of about 1 million records. To properly apply these two classifiers on the dataset, a preprocessing step was performed to clean and prepare data to be used. Experimental results show that Naïve Bayes overcomes Support Vector Machine in term of precision, recall and F-measure.

Peer-review under responsibility of the Conference Program Chairs.

Keywords: Naïve Bayes, Spark, Machine Learning, Support Vector Machine, Big Data.

1. Introduction

The active generation and analysis of large volumes of structured and unstructured data caused the big data problem, this problem is due to three main characteristics: volume, velocity and variety of the data, which are referred to as the 3Vs, in turn these characteristics lead to system challenges in implementing machine learning framework[1]. Thus a

* Corresponding author. Tel.: +962795744288.
E-mail address: w.etaiwi@psut.edu.jo

1877-0509
Peer-review under responsibility of the Conference Program Chairs.
10.1016/j.procs.2017.08.280

powerful machine learning tools, strategies and environment are needed to properly analyze the large volumes of data. The term volume of data refers to the large amount of data collected from many different sources, such as: sensors, databases, multimedia or websites. The speed at which data is collected and analyzed is the key factor when dealing with real time systems such as: sensors and Radio-Frequency Identification (RFID) systems, and this is what the term Velocity means. While Variety of data concerns about different formats of data such as video, audio, email messages, text files, etc. Because big data problems concern about collecting data from its different sources, processing, analyzing and extracting knowledge from it; many frameworks have been proposed in order to deal with such problems, these frameworks are available, but they required to be tested and evaluated in order to select the most suitable framework that can solve a specific big data problem quickly and precisely, knowing that the traditional machine learning algorithms are not applicable on such kind of data[2]. Apache Spark is an open source programming frameworks for data processing originated in the University of California, Berkeley[3]. It has the capability to analyze, manage, process and solve big data problems using an expressive development APIs to allow data workers to develop and execute their works. The Apache Spark operates data processing tasks on many distributed data processing machines, which requires a file management system to collaborate data on those machines such as Hadoop Distributed File System (HDFS), and distribute storage system such as Spark standalone and Hadoop YARN. Because Apache Spark completes data analysis in-memory, it is fast and near real-time framework, in comparison to other big data processing modules, such as MapReduce. Apache Spark can be as 10 times faster for batch processing[4]. Apache Spark architecture consists of three main components: Driver Program, which has the main function to be distributed and executed on other machines. Cluster Manager, which manages cluster resources, and the Worker Node, which is a machine that executes application code. Machine Learning Library (MLlib) is one of the Apache Spark components that consists of common machine learning algorithms and utilities. This paper focuses on two MLlib classification algorithms used for prediction; Naïve Bayes (NB) and support vector machine (SVM). NB is a linear classifier based on the Bayes theorem, it creates simple and well performed models, and it assumes that the features in the dataset are mutually independent, thus the term naïve came along[5]. While, SVM is a learning algorithm that performs classification by finding the hyper plain that maximizes margin between two classes, and the nearest points to the hyper plain are the support vectors that determine the maximum margin[6]. Knowing that Spark MLlib is a new library established in 2014, with little number of published research papers providing evaluation and comparison studies, the goal of this paper is to evaluate and compare two main machine learning algorithms of the MLlib under Apache Spark through predicting bank customer's behaviours. A preprocessing step required to prepare dataset to be analysed. Experimental results were conducted by applying two prediction algorithms on a dataset consisting of customer's personal information and their behaviour in Santander Bank†. The remaining of this paper is structured as follows; section 2 presents the related work. Methodology is presented in section 3, experimental results and evaluation are discussed in section 4, and finally the conclusion is presented in section 5.

2. Related Work

In general two main machine learning tools were considered the best as noted by Richter et al.[7], they presented a multidimensional comparison of four main open source machine learning tools that are used in big data; Mahout, MLlib, H2O, and SAMOA in terms of algorithm availability, scalability, and speed. Although the choice of using one of the tools depends on the goal of the application, the authors conducted that MLlib and H2O are the best tools in terms of algorithm availability, task pipelining and data manipulation. Another research by Landset et al.[8] also claimed that MLlib and H2O are the best machine learning tools in terms of speed, usability, algorithms covered, and scalability to different sizes of datasets. Several research papers were published in the domain of big data, these papers showed that the Spark MLlib is either compared with other machine learning tools or was treated as a part of an architecture/software. A research paper that is an example of comparing the MLlib with other tools is the one presented by Kholod et al.[9]. They proposed a Cloud for Distributed Data Analysis (CDDA) based on the actor model. CDDA is compared with Spark MLlib and Azure ML in terms of performance. Both CDDA and Spark MLlib were tested on a high performance hardware and systems. Experiments were conducted on datasets from Azure ML and results showed

† Santander Banks: Retail banking company, https://www.santanderbank.com/us/personal