Investigation of Mahout Machine Learning Sub-Project

Hossein KardanMoghaddam *, Amir Rajaei**
* Faculty Member of Birjand University of Technology
Birjand, Iran,
h.kardanmoghaddam@birjandut.ac.ir
** Faculty Member of Computer Engineering, Velayat University, Iranshahr, Iran
a.rajaei@velayat.ac.ir

Abstract: *Machine learning is a currently available method to improve services used by man. Mahout project is an environment to develop machine learning applications and algorithms in a distributive manner. This sub-project constitutes various libraries and algorithms for data mining including clustering algorithms (e.g. K-Means) and classification algorithms (e.g. Naïve Bayes). This sub-project is in the process of completion; however, a plethora of algorithms have been ever implemented for it. Although Mahout is mostly known as one of the Hadoop sub-projects, it fails to necessarily mean that it is Hadoop-dependent. Mahout could be used without Hadoop and on a single node and even non-Hadoop cluster. In the present study, the attempt is made to investigate the structures and functions of Mahout machine learning.*

Keywords: Mahout, Big Data, Data Mining, Hadoop.

1. Introduction

One of the objectives of artificial intelligence is to imitate man behaviour, to this end, a machine requires learning capabilities. One of the broad and widely used branches of artificial intelligence is machine learning dealing with setting and exploring the algorithms and procedures based on which computers and systems acquire learning ability. In the most general sense, machine learning aims at enabling computers to gradually achieve better efficiency in performing the given task by increasing data.

Machine learning simply refers to the way of writing a computer program that could learn from experiences and improve its performance as well. Learning may lead to change either in the program or data.

Algorithms used in machine learning suffer multiple limitations and they fail to be generalized over a large group of different patterns. These algorithms may react differently to a pattern having been never observed by any algorithm; while man has a broad knowledge of training and experience in order to act based on them. Moreover, these algorithms have remarkable ability to detect similar conditions as making a decision on new data. Machine learning practices may only generalize what has been ever observed; however, in the aforementioned condition, there are still a number of limitations.

The development trend of low-cost cloud environments, as well as rapid and powerful graphics cards, has had a considerable role in the field of machine learning. The aforementioned factors have led to the exploitative growth of frameworks required for machine learning.

Nowadays, there are a wide range of free and open-source software for training computers and performing online tasks, which highly facilitate machine learning.

One of the oldest machine learning libraries is Shogun developed in 1999. This library has developed based on C++ programming language; however, if fails to be merely limited to this language, it has the ability to work in other environments as well, including Java, Python, C #, Matlab. Shogun is focused on kernel machines like support vector machines to remove regression problems and perform classification task. Shogun also offers a complete implementation of Hidden Markov model [1].

Machine learning tool of Scikit-learn is designed based on Python language packages and it could help statistical and computational tasks in machine learning as well. This machine learning kit works with BSD license; moreover, it is fully free and open-source [2].

Accord.NET Framework is a Net-based signal processing and machine learning framework. This machine learning framework combined with audio and image processing libraries and could cover a wide range of specialized audio and image processing algorithms; additionally, this framework is completely written in C# [3]. Weka (Waikato Environment for Knowledge Analysis) is a popular machine learning software written in Java and includes a set of data mining and machine learning algorithms. This tool was developed at the University of Waikato, New Zealand. Weka was released under the GNU General Public License. It is used for big data analysis. Weka serves the same as a workbench and constitutes a set of visualization tools and algorithms to analyze data and forecast models by graphical interfaces. The original version of Weka written in non-Java was a primary design like a tool for analyzing agriculture-related data. However, its new versions (Weka 3) were completely written in Java [4][5].

Mahout sub-project is an environment for developing distributed machine learning algorithms and applications. This sub-project composes various algorithms and libraries for data mining. Data mining is simply defined as data exploration to extract the valuable data in the