World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:12, No:01, 2018

Gradient Boosted Trees on Spark Platform for Supervised Learning in Health Care Big Data

Authors : Gayathri Nagarajan, L. D. Dhinesh Babu

Abstract : Health care is one of the prominent industries that generate voluminous data thereby finding the need of machine learning techniques with big data solutions for efficient processing and prediction. Missing data, incomplete data, real time streaming data, sensitive data, privacy, heterogeneity are few of the common challenges to be addressed for efficient processing and mining of health care data. In comparison with other applications, accuracy and fast processing are of higher importance for health care applications as they are related to the human life directly. Though there are many machine learning techniques and big data solutions used for efficient processing and prediction in health care data, different techniques and different frameworks are proved to be effective for different applications largely depending on the characteristics of the datasets. In this paper, we present a framework that uses ensemble machine learning technique gradient boosted trees for data classification in health care big data. The framework is built on Spark platform which is fast in comparison with other traditional frameworks. Unlike other works that focus on a single technique, our work presents a comparison of six different machine learning techniques along with gradient boosted trees on datasets of different characteristics. Five benchmark health care datasets are considered for experimentation, and the results of different machine learning techniques are discussed in comparison with gradient boosted trees. The metric chosen for comparison is misclassification error rate and the run time of the algorithms. The goal of this paper is to i) Compare the performance of gradient boosted trees with other machine learning techniques in Spark platform specifically for health care big data and ii) Discuss the results from the experiments conducted on datasets of different characteristics thereby drawing inference and conclusion. The experimental results show that the accuracy is largely dependent on the characteristics of the datasets for other machine learning techniques whereas gradient boosting trees yields reasonably stable results in terms of accuracy without largely depending on the dataset characteristics.

Keywords : big data analytics, ensemble machine learning, gradient boosted trees, Spark platform
Conference Title : ICMLC 2018 : International Conference on Machine Learning and Cybernetics
Conference Location : Singapore, Singapore
Conference Dates : January 08-09, 2018