World Academy of Science, Engineering and Technology
International Journal of Mathematical and Computational Sciences
Vol:14, No:12, 2020

MapReduce Logistic Regression Algorithms with RHadoop

Authors : Byung Ho Jung, Dong Hoon Lim

Abstract : Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. Logistic regression is used extensively in numerous disciplines, including the medical and social science fields. In this paper, we address the problem of estimating parameters in the logistic regression based on MapReduce framework with RHadoop that integrates R and Hadoop environment applicable to large scale data. There exist three learning algorithms for logistic regression, namely Gradient descent method, Cost minimization method and Newton-Rhapson's method. The Newton-Rhapson's method does not require a learning rate, while gradient descent and cost minimization methods need to manually pick a learning rate. The experimental results demonstrated that our learning algorithms using RHadoop can scale well and efficiently process large data sets on commodity hardware. We also compared the performance of our Newton-Rhapson's method with gradient descent and cost minimization methods. The results showed that our newton's method appeared to be the most robust to all data tested.

Keywords : big data, logistic regression, MapReduce, RHadoop
Conference Title : ICSRD 2020 : International Conference on Scientific Research and Development
Conference Location : Chicago, United States
Conference Dates : December 12-13, 2020

Open Science Index, Mathematical and Computational Sciences Vol:14, No:12, 2020 publications.waset.org/abstracts/41569.pdf