Distributed nearest neighbor classification for large-scale multi-label data on spark

Jorge Gonzalez-Lopez, Sebastián Ventura, Alberto Cano

PII: S0167-739X(17)32775-9
DOI: https://doi.org/10.1016/j.future.2018.04.094
Reference: FUTURE 4172

To appear in: *Future Generation Computer Systems*

Received date : 1 December 2017
Revised date : 18 April 2018
Accepted date : 28 April 2018

Please cite this article as: J. Gonzalez-Lopez, S. Ventura, A. Cano, Distributed nearest neighbor classification for large-scale multi-label data on spark, *Future Generation Computer Systems* (2018), https://doi.org/10.1016/j.future.2018.04.094

Distributed Nearest Neighbor Classification for Large-Scale Multi-label Data on Spark

Jorge Gonzalez-Lopez[a], Sebastián Ventura[b,c,d], Alberto Cano[a]

[a] Department of Computer Science, Virginia Commonwealth University, USA
[b] Department of Computer Science and Numerical Analysis, University of Cordoba, Spain
[c] Computing and Information Technology, King Abdulaziz University, Saudi Arabia
[d] Maimonides Biomedical Research Institute of Cordoba, Spain

Abstract

Modern data is characterized by its ever-increasing volume and complexity, particularly when data instances belong to many categories simultaneously. This learning paradigm is known as *multi-label classification* and one of its most renowned methods is the multi-label k nearest neighbor (Ml-knn). The traditional implementations of this method are not feasible for large-scale multi-label data due to its complexity and memory restrictions. We propose a distributed Ml-knn implementation based on the MapReduce programming model, implemented on Apache Spark. We compare three strategies for distributed nearest neighbor search: 1) iteratively broadcasting instances, 2) using a distributed tree-based index structure, and 3) building hash tables to group instances. The experimental study evaluates the trade-off between the quality of the predictions and runtimes on 22 benchmark datasets, and compares the scalability using different sizes of data. The results indicate that the tree-based index strategy outperforms the other approaches, having a speedup of up to 266x for the largest dataset, while achieving an accuracy equivalent to the exact methods. This strategy enables Ml-knn to scale efficiently with respect to the size of the problem.

Keywords: Apache Spark, MapReduce, Distributed Computing, Big Data, Multi-label classification, Nearest Neighbors

Email addresses: gonzalezlopej@vcu.edu (Jorge Gonzalez-Lopez), sventura@uco.es (Sebastián Ventura), acano@vcu.edu (Alberto Cano)