

Accepted Manuscript

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>

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.



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)

Download English Version:

<https://daneshyari.com/en/article/6872902>

Download Persian Version:

<https://daneshyari.com/article/6872902>

[Daneshyari.com](https://daneshyari.com)