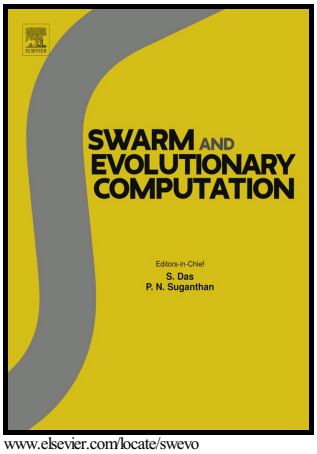

www.elsevier.com/locate/swevo

A Distributed Evolutionary Multivariate Discretizer for Big Data processing on Apache Spark

S. Ramírez-Gallego, S. García, J.M. Benítez, F. Herrera

PII: S2210-6502(16)30572-7
DOI: http://dx.doi.org/10.1016/j.swevo.2017.08.005
Reference: SWEVO303

To appear in: *Swarm and Evolutionary Computation*

Received date: 28 December 2016
Revised date: 19 April 2017
Accepted date: 19 August 2017

Cite this article as: S. Ramírez-Gallego, S. García, J.M. Benítez and F. Herrera, A Distributed Evolutionary Multivariate Discretizer for Big Data processing on Apache Spark, *Swarm and Evolutionary Computation,* http://dx.doi.org/10.1016/j.swevo.2017.08.005

A Distributed Evolutionary Multivariate Discretizer for Big Data processing on Apache Spark

S. Ramírez-Gallego[a,*], S. García[a], J.M. Benítez[a], F. Herrera[a,1]

[a] *Department of Computer Science and Artificial Intelligence, CITIC-UGR, University of Granada, 18071 Granada, Spain*

[b] *Faculty of Computing and Information Technology - North Jeddah, King Abdulaziz University, Saudi Arabia*

Abstract

Nowadays the phenomenon of Big Data is overwhelming our capacity to extract relevant knowledge through classical machine learning techniques. Discretization (as part of data reduction) is presented as a real solution to reduce this complexity. However, standard discretizers are not designed to perform well with such amounts of data. This paper proposes a distributed discretization algorithm for Big Data analytics based on evolutionary optimization. After comparing with a distributed discretizer based on the Minimum Description Length Principle, we have found that our solution yields more accurate and simpler solutions in reasonable time.

Keywords:
Discretizacion, Evolutionary Computation, Big Data, Data Mining, Apache Spark

1. Introduction

Among all Data Mining tasks, Data Preprocessing [1, 2] stands as one of the most important steps in the knowledge discovery process. As input data must be provided in a suitable structure and format for a subsequent high-quality mining process, Data Preprocessing becomes essential in most

*Corresponding author

Email addresses: sramirez@decsai.ugr.es (S. Ramírez-Gallego), salvagl@decsai.ugr.es (S. García), J.M.Benitez@decsai.ugr.es (J.M. Benítez), herrera@decsai.ugr.es (F. Herrera)