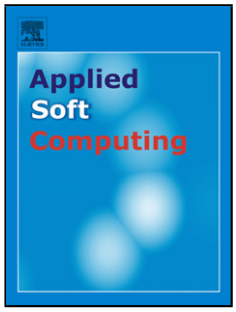

Title: A Tabu search based clustering algorithm and its parallel implementation on Spark

Authors: Yinhao Lu, Buyang Cao, Cesar Rego, Fred Glover

PII: S1568-4946(17)30701-9
DOI: https://doi.org/10.1016/j.asoc.2017.11.038
Reference: ASOC 4582

To appear in: *Applied Soft Computing*

Received date: 15-5-2017
Revised date: 21-11-2017
Accepted date: 24-11-2017

Please cite this article as: Yinhao Lu, Buyang Cao, Cesar Rego, Fred Glover, A Tabu search based clustering algorithm and its parallel implementation on Spark, Applied Soft Computing Journal https://doi.org/10.1016/j.asoc.2017.11.038

A Tabu Search based clustering algorithm and its parallel implementation on Spark

Yinhao Lu, Buyang Cao#
School of Software Engineering
Tongji University, 4800 Cao'An Road, Shanghai, China 201804

Cesar Rego
School of Business Administration
University of Mississippi, University, MS 38677, USA.

Fred Glover
ECEE, College of Engineering & Applied Science
University of Colorado, Boulder, CO 80309, USA

Highlights

- Our Tabu Search based clustering algorithm produces more accurate and stable solutions compared to the widely-applied Spark MLlib K-means algorithm.
- In addition to these advantages, the parallelized Tabu Search based clustering algorithms achieves an acceleration rate similar to that of the K-means algorithm in Spark MLlib.
- To our knowledge, this is the first Tabu Search based clustering algorithm that has been implemented on the Spark platform.

Abstract The well-known *K*-means clustering algorithm has been employed widely in different application domains ranging from data analytics to logistics applications. However, the *K*-means algorithm can be affected by factors such as the initial choice of centroids and can readily become trapped in a local optimum. In this paper, we propose an improved *K*-means clustering algorithm that is augmented by a Tabu Search strategy, and which is better adapted to meet the needs of big data applications. Our design focuses on enhancements to take advantage of parallel processing based on the Spark framework. Computational experiments demonstrate the superiority of our parallel Tabu Search based clustering algorithm over a widely used version of the *K*-means approach embodied in the parallel Spark MLlib system, comparing the algorithms in terms of scalability, accuracy, and effectiveness.

Keywords: Clustering, K-means, Tabu Search, parallel computing, Spark

1. INTRODUCTION

The purpose of a clustering process is to group a set of (abstract or physical) objects into multiple classes, so that the objects in each class (cluster) are similar according to certain rules or criteria. A clustering algorithm in general seeks to build the clusters by the two interrelated

Corresponding author