Applied Soft Computing 69 (2018) 816–817

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Editorial

Special issue on evolutionary multi-objective optimization and applications in big data

An increasing number of real-life problems can be effectively represented and addressed in terms of multi-objective optimization, where several conflicting objective functions are optimized at once with the aim of finding a set of Pareto optimal solutions. For this reason, the topic has attracted a relevant amount of research efforts in the last twenty years, with important results and applications in many fields, including engineering, economics, logistics, bioinformatics, medicine, and others. According to the literature, optimization metaheuristics, and in particular evolutionary algorithms, are considered among the most effective methods for solving multi-objective optimization problems by providing a number of Pareto-optimal solutions in a single run.

Nowadays, some of the most relevant applications of multi-objective optimization algorithms belong to the field of big data analytics. Indeed, with the ever-increasing production of data from various heterogeneous sources, processing large and varied data sets to uncover hidden patterns, previously unknown correlations, people preferences, and other information that may be relevant for tackling real-life problems, became one of the most intriguing challenge. Big-data analytics processes may involve, in various phases and for different aspects, the solution of optimization problems. However, such optimizations have to be carried out by accounting for the specific research challenges related to big data, which may originate from some well-known characteristics, such as large volume, variety of different sources, dynamicity, and fast real-time generation of input data. All such aspects pose new research challenges, essentially related to the need of achieving a scalable data-intensive processing, as well as to design specific and more efficient optimization strategies. Therefore, it is of great interest to investigate the role and the effectiveness of multi-objective optimization metaheuristics, including evolutionary algorithms, for addressing the optimization and learning problems involving big data analytics.

In the above context, the overall aim of this special issue was that of providing to researchers and practitioners, in both academy and industry, an important opportunity to express and discuss their views on current challenges, trends, state-of-the art solutions and to illustrate new advancements and applications in the field of multi-objective optimization and big data.

The special issue received more than twenty submissions. The manuscripts were rigorously peer reviewed, according to the journal standards, by recognized researchers in the specific topics. At the end of the process, we selected seven high-quality articles concerning relevant applications in big-data analytics. We grouped the articles into two sets based on the adopted metaheuristic approach, namely Evolutionary Multiobjective Optimization (EMO group) and Swarm Intelligence (SI group).

In particular, the EMO group contains four articles concerning the use of evolutionary algorithms to address typical problems related to big data, as well as the design and development of frameworks to facilitate the usage of evolutionary algorithms for coping with such problems. The articles are briefly summarized below.

First, in the article "Data Analysis Framework of Sequential Clustering and Classification Using Non-Dominated Sorting Genetic Algorithm", Chao-Lung Yang and Thi Phuong Quyen Nguyen propose the NSGAII-SCC framework, which can be considered as a preliminary data analysis tool when the labels for classification are not available or the objective of the data analysis itself is only roughly defined. By combining feature selection, clustering, and classification, NSGAII-SCC can: (*i*) reveal the hidden patterns on one set of data; (*ii*) select the data features of another set, which are correlated with the discovered patterns; (*iii*) train a model to predict those patterns. More in detail, in order to balance the quality of clustering and classification, NSGAII-SCC formulates the problems as a multi-objective optimization. The latter is tackled using the well-known non-dominated Sorting Genetic Algorithm II, together with a specifically designed chromosome encoding. By using two public datasets, the proposed approach is extensively investigated. According to the results, NSGAII-SCC can outperform other methods and is effective in guiding the data analysis task, revealing hidden patterns and exploring the data features correlated with the discovered patterns.

The article "Multiobjective Characteristic-based Framework for Very-Large Multiple Sequence Alignment", authored by Álvaro Rubio-Largo, Leonardo Vanneschi, Mauro Castelli, and Miguel A. Vega-Rodriguez, proposes a new approach for addressing the Multiple Sequence Alignment problem. The work is based on the observation that the vast majority of existing algorithms requires the setting of specific flags in order to modify certain alignment parameters and run in an optimal way. Therefore, the framework exploits an intelligent mechanism for setting the best configuration of parameters. In particular, depending on the biological characteristics of the input dataset, the devised strategy runs the aligner with the best parameters found for another dataset that has similar biological characteristics, in such a way to improve the accuracy and conservation of the obtained alignment. In the experimental part of the article, the proposed framework is trained using three well-known multi-objective evolutionary algorithms: NSGA-II, IBEA, and MOEA/D. Moreover, the article includes a comprehensive comparative study between the characteristic-based version of

https://doi.org/10.1016/j.asoc.2018.06.016

three well-known aligners (Kalign, MAFFT, and MUSCLE) and several aligners proposed in the literature. The results show the significant advantages of the proposed framework when dealing with well-known benchmarks, such as PREFAB v4.0 and SABmark v1.65 and very-large benchmarks with thousands of unaligned sequences (HomFam).

Furthermore, the article "JMetalSP: a Framework for Dynamic Multi-Objective Big Data Optimization", authored by Cristobal Barba-González, José García-Nieto, Antonio J. Nebro, José A. Cordero, Juan J. Durillo, Ismael Navas-Delgado and José F. Aldana-Montes, presents jMetalSP, which combines the multi-objective optimization features of the jMetal framework with the streaming facilities of the Apache Spark cluster computing system. jMetal stands for Metaheuristic Algorithms in Java, and it is an object-oriented Java-based framework for multi-objective optimization with metaheuristics. The article adapts and extends jMetal in order to solve dynamic multi-objective optimization problems that may easily arise when dealing with big data. Besides describing the proposed architecture, the authors show how JMetalSP can be used to solve a dynamic bi-objective instance of the Traveling Salesman Problem (TSP) based on New York City's real-time traffic data. Moreover, an experimental study assesses the performance of the jMetalSP application in a Hadoop cluster composed of 100 nodes and shows that jMetalSP is able to adapt to changing traffic conditions when solving the dynamic TSP test problem.

A problem of information retrieval in very large datasets is addressed by the article "Topic Relevance and Diversity in Information Retrieval from Large Datasets: a Multi-Objective Evolutionary Algorithm Approach", authored by Rocio L. Cecchini, Carlos M. Martín Lorenzetti, Ana G. Maguitman and Ignacio Ponzoni. In particular, the authors exploit multi-objective evolutionary algorithms to tackle topical search problems (i.e., searching for material that is relevant to a given topic). The adopted approach consists of evolving a population of queries towards successively better individuals, for a given topic. However, in a straightforward evolutionary approach, previous studies have revealed that the population tends to converge to a solution set composed of few queries leading to the exploration of a very limited region of the search space. The authors successfully mitigate such an issue through some effective strategies to favour diversity in evolutionary topical search, based on novel fitness functions, different parameterization for the crossover and mutation rates, and the use of multiple populations to promote diversity preservation. The article proves the effectivity and efficiency of the proposed strategies through a computational study based on a dataset including more than 350,000 labelled web pages.

The SI group contains the remaining three articles summarized below. In them, swarm intelligence was adopted for coping with some typical issues related to big data, namely clustering, dataset balancing, and need for fast processing in case of real-time data analysis.

In particular, in the article "Parallel Swarm Intelligence Strategies for Large-scale Clustering based on MapReduce with Application to Epigenetics of Aging", the authors Zakaria Benmounah, Souham Meshoul, Mohamed Batouche and Pietro Liò address the task of clustering, one of the most relevant techniques for data analysis and knowledge discovery. Clustering big data is a challenging issue for which conventional algorithms as well as optimization metaheuristics exhibit limited efficiency. To tackle such an issue, the authors developed a decentralized distributed big data clustering solution using three swarm intelligence algorithms according to the MapReduce approach. The latter enables an effective cooperation between the involved swarm algorithms to achieve largely scalable data partitioning through a migration strategy. The article investigates the proposed approach trough Amazon Elastic MapReduce service on 192 computer nodes and using 30 gigabytes of data. Compared with some state-of-the-art big data clustering results, the proposed framework shows a significant improvement, both in terms of computing time and quality of the achieved solution. Moreover, the authors discuss an interesting application to epigenetics data clustering, in order to study the epigenetics impact on aging.

When dealing with big data, another typical issue originates from imbalanced datasets, in which the amounts of labelled patterns of the different classes are very different. Such a problem, considering the case of binary data classification, is the object of "A Suite of Swarm Dynamic Multi-Objective Algorithms for Rebalancing Extremely Imbalanced Datasets", by Jinyan Li, Simon Fong, Raymond Wong, Sabah Mohammed, Jinan Fiaidhi, Yunsick Sung. Rather than just trying to match the data quantities of the two classes, the proposed algorithms focus on guaranteeing the credibility of the classification model and reaching the greatest possible accuracy by dynamically rebalancing the training dataset with multi-objective swarm intelligence multi-objective optimisation. The proposed approach, first find a set of solutions that satisfy the kappa criterion, then it searches for the solution in the set that offers the highest accuracy. The search is based on multi-objective optimisation in order to find a solution that satisfies several criteria at the same time. Moreover, in order to handle properly streaming data, the optimization operates incrementally. The article includes several experimental results and comparisons that prove the effectivity of the approach when imbalanced datasets are used as training datasets for inducing a classifier.

A new Single- and Multi-objective version of the Firefly Algorithm (FA), particularly efficient in dealing with big data, is presented by Hui Wang, Wenjun Wang, Laizhong Cui, Hui Sun, Jia Zhao, Yun Wang and Yu Xue, in "A Hybrid Multi-Objective Firefly Algorithm for Big Data Optimization". Unlike existing multi-objective FAs, the proposed approach employs simultaneously three techniques: (*i*) a crossover strategy from Differential Evolution (*ii*) parameter adaptation and (*iii*) the non-dominated sorting method with density estimation strategy used in NSGAII to generate Pareto fronts. In the article, the new algorithm is tested on the benchmarks proposed for the Big Data Optimization Competition at IEEE CEC 2015, which includes different optimization problems concerning the processing of electroencephalographic (EEG) signals. According to the experimental results, the proposed approach proved very promising.

Overall, we think that all the above articles can capture some of the most up-to-date research trends and challenges in the field of multi-objective optimization applied to big data.

As guest editors, we would like to thank the Editor-in-Chief of Applied Soft Computing for giving us the opportunity of working on this Special Issue. We would also like to thank the journal Managing Editor and the Elsevier team for all the operational support, the authors for their submissions and, last but not least, all the reviewers, who have provided their valuable feedback on each article in a timely and professional manner. We hope that this collection of research articles will suggest to the readers interesting directions for new development and useful real-life applications.

Gianluigi Folino*
ICAR-CNR, Rende, Italy

Giuseppe A. Trunfio
DADU, University of Sassari, Alghero, Italy

* Corresponding author.
E-mail addresses: gianluigi.folino@icar.cnr.it (G. Folino), trunfio@uniss.it (G.A. Trunfio).