Information Sciences 320 (2015) 12–25

Contents lists available at ScienceDirect

Information Sciences

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

Toward high performance solution retrieval in multiobjective clustering

Alvaro Garcia-Piquer [a], Andreu Sancho-Asensio [b,*], Albert Fornells [c], Elisabet Golobardes [b], Guiomar Corral [d], Francesc Teixidó-Navarro [c]

[a] Institut de Ciències de l'Espai (IEEC – CSIC), Campus UAB, Facultat de Ciències, Torre C5 – parell – 2a planta, E-08193 Bellaterra, Spain
[b] Research Group in Electronic and Telecommunications Systems and Data Analysis, Ramon Llull University, Quatre Camins 2, 08022 Barcelona, Spain
[c] Research Group in Tourism, Hospitality and Mobilities, School of Tourism and Hospitality Management – Sant Ignasi, Ramon Llull University, Marqués de Mulhacén 40-42, 08034 Barcelona, Spain
[d] Research Group in Internet Technologies and Storage, Ramon Llull University, Quatre Camins 2, 08022 Barcelona, Spain

ARTICLE INFO

Article history:
Received 26 September 2014
Received in revised form 4 February 2015
Accepted 18 April 2015
Available online 23 April 2015

Keywords:
Soft-computing
Genetic algorithms
Multiobjective optimization
Clustering
Pareto set filtering

ABSTRACT

The massive generation of unlabeled data of current industrial applications has attracted the interest of data mining practitioners. Thus, retrieving novel and useful information from these volumes of data while decreasing the costs of manipulating such amounts of information is a major issue. Multiobjective clustering algorithms are able to recognize patterns considering several objective function which is crucial in real-world situations. However, they dearth from a retrieval system for obtaining the most suitable solution, and due to the fact that the size of Pareto set can be unpractical for human experts, autonomous retrieval methods are fostered. This paper presents an automatic retrieval system for handling Pareto-based multiobjective clustering problems based on the shape of the Pareto set and the quality of the clusters. The proposed method is integrated in CAOS, a scalable and flexible framework, to test its performance. Our approach is compared to classic retrieval methods that only consider individual strategies by using a wide set of artificial and real-world datasets. This filtering approach is evaluated under large data volumes demonstrating its competence in clustering problems. Experiments support that the proposal overcomes the accuracy and significantly reduces the computational time of the solution retrieval achieved by the individual strategies.

1. Introduction

Clustering [39,15,32] is a trending data mining technique used in real-world situations to partition a data set into several groups according to some criteria and therefore identifying novel and potentially useful patterns from data. Conventional clustering algorithms are focused on obtaining groups by optimizing a single fitness function. In contrast, it can be difficult to obtain good data partitions in some real-world problems using a single objective function, and it is necessary to define several of them to obtain more accurate clusters [35]. These objective measures can be summarized in a single fitness function if they are disjoint. However, when the defined objectives conflict with each other it is necessary to define a fitness

* Corresponding author.
E-mail addresses: agarcia@ice.csic.es (A. Garcia-Piquer), andreus@salleurl.edu (A. Sancho-Asensio), albert.fornells@tsi.url.edu (A. Fornells), elisabet@salleurl.edu (E. Golobardes), guiomar@salleurl.edu (G. Corral), francesc.teixido@tsi.url.edu (F. Teixidó-Navarro).

http://dx.doi.org/10.1016/j.ins.2015.04.041
0020-0255/© 2015 Elsevier Inc. All rights reserved.

function for each objective in order to find a solution which would give acceptable values for all of them [7]. A widely used technique to competently carry out this is multiobjective clustering (MC) [30], which uses the concept of Pareto Optimum with a posteriori approach [8] for simultaneously optimizing a set of mutually confronted objectives in order to promote the definition of clusters. This technique returns a collection that contains a number of Pareto optimal solutions (the so called Pareto set), none of which can be further improved on any objective without degrading another one [12].

There are different strategies for multiobjective optimization such as Simulated Annealing [47] and Ant Colony Optimization [37], but Multiobjective Evolutionary Algorithms (MOEAs) [7] have become one of the most capable strategies to solve this kind of problems [17,51] since they (1) work with a collection of solutions with different trade-offs among objectives, which are improved until a Pareto set with optimal trade-offs is obtained; (2) can be easily adapted to the type of data of the studied domain, due to the flexible knowledge representation used; and (3) are able to optimize different objectives without assuming any underlying structure of the objective functions. However, the performance of MOEAs can be compromised in large databases due to their high computational and memory usage requirements [19]. Moreover, one of the key challenges in Pareto-based MOEAs is the retrieval of the most suitable solution from the final Pareto set. This solution is typically identified by an expert in the domain. Nonetheless this process results in a subjective criterion and in a non trivial and tedious task if there are several solutions in the Pareto set. Thus, automatic methods are strongly required in order to help experts and simplify the identification of the most suitable solution, which can be beneficial in challenging domains such as health, smart networks or education. These are areas in which large volumes of data are generated.

In MC algorithms there are mainly two approaches to retrieve the most suitable solution from the Pareto set: (1) consider the shape of the Pareto set [43] or (2) consider the features related to the morphological properties of clusters [30]. The first method tries to identify the knee of the Pareto set to retrieve the solution with the best trade-off between objectives, but it does not take into account the resulting quality of clusters. The term quality is defined as how useful the solution is for the expert in the domain. Furthermore, quality is directly related to the shape, size and compactness of the clusters and the separation between them, characteristics which can be evaluated using clustering validation indexes [25,26,40]. The second method retrieves the best solution according to clustering validation indexes but its objective values could be unbalanced and the solution may only properly optimize a single objective.

The purpose of this paper is to propose a scalable retrieval filtering method that contemplates both the shape of the Pareto set and the quality of the clusters. The goal is to retrieve explanatory solutions with an acceptable trade-off between objectives in MC based on MOEAs. The proposed retrieval method is based on the observation that solutions with acceptable balance between objectives are placed around the knee of the Pareto front. The aim is to filter clustering solutions with less objective trade-off in order to retrieve the best solution from the remaining ones according to a clustering validation index. Thus, extra computations to evaluate non-interesting solutions are avoided. To test our approach we use the *Clustering Algorithm based on multiObjective Strategies* (CAOS) [10,22], a MC algorithm based on PESA-II [9]. CAOS uses a representation that does not depend on the number of instances of the data set, subsequently it is memory scalable [21]. Moreover, it scales the computational time of the clustering process by dividing the original data set to several subsets that are alternatively used in each generation of the MOEA process, thus it uses less data in each evolutionary cycle. This is performed in this way to avoid biasing the population by using only a single sample, while achieving low penalization in accuracy [2]. More specifically, the approach acts iteratively through the evolutionary cycle, being an automatic, adaptive system, thence fostering objectivity in the filtering parameters.

We compare the proposed method with the retrieval strategies based on (1) the shape of the Pareto set and (2) the morphological properties of clusters. All approaches are compared along a wide set of synthetic data sets [30] and real-world ones from the UCI [18] and KEEL [1] repositories. Furthermore, we carry out another set of experiments in data sets with large amounts of data in order to test the scalability capabilities of the method. Results show that accuracy and retrieval time are improved with this new proposal with a negligible additional cost to the evolutionary cycle. For a comparison between CAOS and other clustering methods, the reader is referred to [21].

The contributions of this paper are the following:

- It explores a filtering method that greatly increases the efficiency in retrieving solutions in two-objective clustering MOEAs.
- It integrates the proposed method in a scalable and flexible clustering framework.
- It tests the filtering method in a massive amount of data sets, including large ones.
- It shows a high performance in solution retrieval in both moderate and large data sets.
- It encourages practitioners to exploit the presented filtering technique to address the problem of retrieving the most suitable solutions from Pareto-based MOEAs.

The remainder of the paper is organized as follows. Section 2 briefly summarizes the related work on retrieving solutions in MC based on MOEAs. Section 3 introduces CAOS and describes the required modifications in order to adapt it to (1) become memory scalable and (2) the new filtering method. Section 4 describes the proposed retrieval method. Section 5 describes the experimentation and discusses the results. Finally, Section 6 ends with conclusions and further work.