Applied Soft Computing 54 (2017) 256–266

Contents lists available at ScienceDirect

Applied Soft Computing

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

A comparison of two hybrid methods for constrained clustering problems

Rudinei Martins de Oliveira[a,*], Antonio Augusto Chaves[a], Luiz Antonio Nogueira Lorena[b]

[a] Federal University of São Paulo - UNIFESP, Brazil
[b] National Institute for Space Research - INPE, Brazil

ARTICLE INFO

Article history:
Received 5 May 2016
Received in revised form 11 January 2017
Accepted 12 January 2017
Available online 20 January 2017

Keywords:
Column generation
BRKGA
Clustering

ABSTRACT

This paper proposes two hybrid heuristics to solve the constrained clustering problem. This problem consists of partitioning a set of objects into clusters with similar members that satisfy must-link and cannot-link constraints. A must-link constraint indicates that two selected objects must be in the same cluster, and cannot-link constraint means that two selected objects must be in distinct clusters. The two proposed hybrid methods are biased random key genetic algorithm (BRKGA) with local search (LS) heuristic and column generation (CG) with path-relinking (PR) and local search (LS) heuristic. Computational experiments considering instances available in the literature are presented to demonstrate the efficacy of the proposed methods to solve the constrained clustering problem. Moreover, the results of the CG and BRKGA are compared with the CCCG, CP and CPRBBA method.

1. Introduction

Clustering problems involve analysis of large volumes of data and can have many applications, such as identifying patients with similar clinical conditions, group of public health problems, or hereditary health problems; identifying customers' buying patterns, verifying the users' behavior on a given web page; and clustering actions with the same price fluctuations [2].

Clustering searches for patterns in the data often by trying to extract some kind of information that makes sense. The tendency is to work with large scale data and improve the accuracy of the obtained results. Data may be plants, proteins, cancer cells, image segmentation, pattern recognition, GPS, internet pages accessed, etc.

The clustering problem is defined as the process of partitioning a set of objects into groups such that members of a group are similar to each other [5]. This problem is known to be NP-hard, and finding optimal feasible solutions can be very complicated [21]. Therefore, to improve the separation of objects into clusters, constraints have been added in the clustering process. The constrained clustering problem seeks to separate a set of data into similar clusters that satisfy certain constraints. Each constraint is related with two objects and is known a priori.

According to Davidson and Ravi [10], in general, there are at least two models of constraints. The first model employs a learning algorithm that uses the results of the clustering algorithm to classify new objects. The second model uses the original clusters to create the constraints, and the inference is made from them; the constraints are used to guide the algorithm. In the second model, constraints can be must-link and cannot-link. Must-link indicates that if an object is associated with a group, the other must also be. Cannot-link indicates that if an object is associated with a group, the other cannot be. The second model of constraints is the focus of this paper.

To group the objects, the clusters must be created. Some authors create clusters by choosing medians through the solution formulation of a p-median problem [27]. Others create clusters by choosing centroid using variations of the k-Means method [32].

Objects are clustered considering the similarity with the other group members. For each object in a set of data, it is necessary to identify which group is the most similar. This is very important, because depending on the procedure used to separate the data, different partitions may be obtained. A procedure is to create an array with the distance between all objects. In this paper, we use the Euclidean distance to verify the similarity between objects. The created groups are analyzed by a cost function, which seeks

* Corresponding author.
E-mail addresses: rudmart@gmail.com (R.M.d. Oliveira), antonio.chaves@unifesp.br (A.A. Chaves), luiz.lorena@inpe.br (L.A.N. Lorena).

http://dx.doi.org/10.1016/j.asoc.2017.01.023
1568-4946/© 2017 Elsevier B.V. All rights reserved.

to minimize the sum of distances of elements of a cluster to the centroid of that cluster.

The difficult problem is to identify data that are similar and develop methods to group them. Due to the large number of solutions, the methods may not classify the data efficiently. This classification can be improved when it imposes must-link and cannot-link constraints to formulate the problem [32].

In this context, the aim of this paper is to solve the constrained clustering problem. The two proposed heuristic methods are a biased random key genetic algorithm (BRKGA), with a local search (LS), and a column generation (CG), which combines path-relinking (PR) and local search (LS). The BRKGA was based on Festa [14] and CG was based on Senne et al. [30] and Oliveira et al. [27]. Both methods were improved and applied to a new problem, which is known as the constrained clustering problem.

The BRKGA represents solutions on vectors with random-keys. This vector is decoded into a solution for the specific problem. This feature makes the method independent of the problem. Thus, the only component that needs to be implemented is a decoder function. In the BRKGA, we also apply a local search heuristic to speed up the convergence of the method.

The column generation method is used to generate the clusters to solve the formulation of *p*-median. Intermediary solutions compose a path-relinking that begins from an initial solution and a solution guide, which searches for a better constrained cluster that is subsequently improved by a local search. The local search accelerates the convergence of the method by exploring new regions in the search space.

To compare and validate the results, we compared the CG and BRKGA with are compared with the CCCG [2], CP [8] and CPRBBA method [18]. The results obtained in this paper demonstrate the effectiveness of the proposed methods as an alternative to solve the constrained clustering problem to find good feasible solutions.

The paper is organized as follows. Section 2 gives a brief literature review on the constrained clustering problem. Section 3 is an overview of the BRKGA method, and Section 4 discusses the column generation heuristic. Section 5 presents the data and the computational results. In Section 6, some conclusions are mentioned.

2. Literature review

The clustering problem has been extensively studied, and researchers use a wide variety of methods seeking to solve it. For example, Rand [28] proposes patterns that isolate aspects of performance of a method, such as return, sensitivity, and stability. These patterns depend on a similarity measure between two different clusters of the same set of data; the measure should essentially consider how each data point pair is assigned in each set.

Handl et al. [19] show the large amount of techniques available to validate results obtained for the problem, with the principal focus of the paper being the application of post-genomics data analysis. The authors use synthetic and real biological data to demonstrate the benefits and also validation risks.

Chang et al. [6] propose a separation algorithm based on genetic algorithms with gene rearrangement to k-Means clustering, which seeks to remove the degeneracy for the purpose of a more efficient search. A crossover operator that exploits a measure of similarity between chromosomes in a population is also presented.

Nascimento et al. [25] present a mathematical formulation and a greedy randomized adaptive search procedure to solve the clustering problem using biological data. The computational results were compared with the direct application of CPLEX (ILOG, 2009) solver, k-Means, k-Medians, and partitioning around medoids (PAM). The CRand index is used to compare the methods.

Festa [14] provide an overview of the main types of clustering and criteria for separation of data. In this paper, the BRKGA is also described and compared with the hybrid GRASP with path-relinking algorithms used to cluster biological data. The data used are biological data: fold protein classification, prediction of protein localization sites, cancer diagnosis, and Iris data. The Adjusted Rand Index was used to compare in the methods.

Oliveira et al. [27] examine hybrid heuristics to solve clustering problems. The hybrid heuristics are based on the application of a column generation technique for solving *p*-medians problems. Five heuristics are derived directly from the column generation algorithm to solve the clustering problem.

The papers cited above solve the clustering problem without constraints. The principal focus of these papers was to group similar objects into the same clusters. An extension on the studies of clusters has added constraints to the clustering problem, which is the purpose of this paper. The following papers solve the constrained clustering problem. Constraints are named must-link and cannot-link.

Constrained clustering has many applications and has been extensively studied in literature. The *k*-Means was widely used in the initial studies on the constrained clustering problem and is still the most explored in the literature. For example, Wagstaff and Cardie [32] modify and add must-link and cannot-link constraints in the COP-kMeans algorithm that is a variation of *k*-Means. This paper shows that the type of constraint that is most effective can vary between data sets.

Wagstaff et al. [33] use the COP-kMeans algorithm that has background knowledge in the form of instance-level constraints. They apply it to the real-world problem of automatically detecting road lanes from GPS data. The COP-kMeans algorithm outperformed the unconstrained k-Means algorithm, for all data analyzed.

Kleinberg [22] develop an axiomatic framework for clustering. As a result, they develop the basic impossibility theorem for a set of three simple properties: essentially scale-invariance, a richness, and a consistency condition. This paper shows that no clustering function satisfies all three properties.

Basu et al. [3] present a framework for pairwise constrained clustering (PCC). The pairwise are named must-link and cannot-link constraints. The objective function in the PCC framework aims to minimize the sum of the total distance between the points and their cluster centroids as well as the cost of violating the pairwise constraints with a cost of violating each constraint. They also propose a method for actively selecting informative pairwise constraints to improve clustering performance. The results show that the Active selection of pairwise constraints significantly outperforms random selection of constraints.

According to Davidson and Ravi [9], constrained clustering can be divided into two types: constraints that help the algorithm to learn and constraints that guide the solution. In this paper, the *k*-Means algorithm is modified so that the must-link and cannot-link constraints are identified and interpreted as a conjunction or disjunction, respectively. A must-link for two objects is considered if the distance between them is at most a threshold and cannot-link when the distance is at least another threshold. The authors explore the feasibility and complexity of the problem.

Davidson et al. [12] create two quantitative measures, informativeness and coherence, which can be used to identify useful constraint sets. The authors study why some constraint sets increase the clustering accuracy while others have no effect or even decrease the accuracy. Different constrained clustering algorithms on some clustering problems were examined. The algorithms are COP-KMeans, which performs hard constraint satisfaction; PC-KMeans, which performs soft constraint satisfaction and permits some constraints to be violated; M-KMeans which performs metric learning from constraints and does not require that the constraints