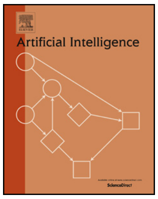
Artificial Intelligence ••• (••••) •••–•••

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint



Constrained clustering by constraint programming

Thi-Bich-Hanh Dao*, Khanh-Chuong Duong, Christel Vrain

Univ. Orléans, INSA Centre Val de Loire, LIFO, EA 4022, F-45067, Orléans, France

ARTICLE INFO

Article history:
Received in revised form 12 May 2015
Accepted 23 May 2015
Available online xxxx

Keywords:
Constrained clustering
Bi-criterion clustering
Constraint programming
Modeling
Global optimization constraint
Filtering algorithm

ABSTRACT

Constrained Clustering allows to make the clustering task more accurate by integrating user constraints, which can be instance-level or cluster-level constraints. Few works consider the integration of different kinds of constraints, they are usually based on declarative frameworks and they are often exact methods, which either enumerate all the solutions satisfying the user constraints, or find a global optimum when an optimization criterion is specified. In a previous work, we have proposed a model for Constrained Clustering based on a Constraint Programming framework. It is declarative, allowing a user to integrate user constraints and to choose an optimization criterion among several ones. In this article we present a new and substantially improved model for Constrained Clustering, still based on a Constraint Programming framework. It differs from our earlier model in the way partitions are represented by means of variables and constraints. It is also more flexible since the number of clusters does not need to be set beforehand; only a lower and an upper bound on the number of clusters have to be provided. In order to make the model-based approach more efficient, we propose new global optimization constraints with dedicated filtering algorithms. We show that such a framework can easily be embedded in a more general process and we illustrate this on the problem of finding the optimal Pareto front of a bi-criterion constrained clustering task. We compare our approach with existing exact approaches, based either on a branch-and-bound approach or on graph coloring on twelve datasets. Experiments show that the model outperforms exact approaches in most cases.

1. Introduction

Constrained Clustering has received much attention this last decade. It allows to make the clustering task more accurate by integrating user constraints. Several kinds of constraints can be considered. First, constraints may be used to limit the size or the diameter of clusters; second, they can enforce expert knowledge instances that must be or cannot be in the same cluster (must-link or cannot-link constraints). Much work has focused on instance-based constraints and has adapted classical clustering methods to handle must-link or cannot-link constraints. A small number of earlier studies have considered the integration of different kinds of constraints. These studies are based on declarative frameworks and offer exact methods that either enumerate all the solutions satisfying the user constraints, or find a global optimum when an optimization criterion is given. For instance, in [1] a SAT based framework for constrained clustering has been proposed, integrating many kinds of user constraints but limited to clustering tasks into two clusters. A framework for conceptual clustering based on Integer Linear Programming has also been proposed in [2]. In [3], we have presented a model based on Constraint Programming for

* Corresponding author.
E-mail addresses: thi-bich-hanh.dao@univ-orleans.fr (T.-B.-H. Dao), khanh-chuong.duong@univ-orleans.fr (K.-C. Duong), christel.vrain@univ-orleans.fr (C. Vrain).

http://dx.doi.org/10.1016/j.artint.2015.05.006
0004-3702/© 2015 Elsevier B.V. All rights reserved.

constrained clustering. This model allows to choose one among different optimization criteria and to integrate various kinds of user constraints. As far as we know, the approach we propose is the only one able to handle different optimization criteria and all popular constraints, for any number of clusters. It is based on Constraint Programming (CP): in such a paradigm, a constraint optimization problem or a constraint satisfaction problem is modeled by defining variables with their domains and by expressing constraints on these variables. Solving a CP problem relies on two operations: constraint propagation that reduces the domain of the variables by removing inconsistent values and branching that divides the problem in subproblems, by taking an unassigned variable and by splitting its domain into several parts. It is important to notice that modeling a task in Constraint Programming implies several choices, which have a high impact on the efficiency of the approach: the choice of the variables and the choice of the constraints for the model, the development of filtering algorithms dedicated to the task and the use of adapted search strategies for solving the model. A point in favor of CP is that the requirement of getting an exact solution can be relaxed by using metaheuristics or local search methods. For the time being, we have fully investigated exact methods, to push the efficiency of the framework as far as possible. Approximate search strategies could be integrated in the future.

In this paper, we propose a new model for Constrained Clustering, still based on Constraint Programming, but significantly improved compared to the previous model [3]. In the previous model, two sets of variables were introduced, namely a variable for each cluster identifying a cluster by one of its points and a variable for each point expressing its assignment to a cluster. The number of clusters had to be set beforehand. The new model we present here contains only a variable for each point, giving the index of the cluster the point belongs to. As a result, the constraints enforcing the solution to be a partition and breaking symmetries are entirely different. The new model is lighter in terms of the number of variables. It also enables to remove the restriction on the number of clusters; only bounds on the number of clusters are required. Moreover, in order to make this model efficient, we have developed dedicated global constraints for three optimization criteria: minimizing the maximal diameter, maximizing the split between clusters, and minimizing the within-cluster sum of dissimilarities.

The approach we propose may be easily embedded in a general process for the task of Constrained Clustering. Considering Data Mining as an iterative and interactive process composed of the classical steps of task formulation, data preparation, application of a tool, thus requiring to set parameters, and validation of the results, a user can specify the task at hand including or not some constraints and decide to change the settings according to the results. He/she may decide to change the constraints, removing or relaxing some constraints, adding or hardening other constraints. The modularity and declarativity of our model allow this easily. In this paper, we illustrate the integration of our model in a more complex process by considering a bi-criterion clustering problem, namely finding the Pareto front when minimizing the maximal diameter and maximizing the minimal split. To achieve this, our framework is integrated in an algorithm, which alternatively calls our model to minimize the maximal diameter and then to maximize the split between clusters with adapted constraints.

Our contributions are as follows.

- We propose a new model based on Constraint Programming, allowing to find an optimal solution for clustering under constraints, given an optimization criterion. This new model improves substantially the previous one, it is more modular (each criterion is implemented by a global constraint) and it is much more efficient.
- We show that such a framework can easily be embedded in a more general process and we illustrate this on the problem of finding the optimal Pareto front of a bi-criterion constrained clustering task. As far as we know, this is the first approach to handle bi-criterion clustering in presence of user-constraints.
- We propose new global optimization constraints with dedicated filtering algorithms, thus allowing to make the model more efficient.
- We compare this model with existing exact approaches, based either on a branch-and-bound approach [4] or on graph coloring [5] on twelve datasets. Experiments show that the model we propose is generally more efficient. Moreover we compare the two models based on CP that we have developed and we show that the different changes (search strategy and development of global constraints) allow to improve the model.

The paper is organized as follows. Section 2 is dedicated to preliminaries on Constrained Clustering and Constraint Programming. Related work is presented in Section 3. Section 4 is devoted to the presentation of both CP models, the first one presented in [3] and the new one. The filtering algorithms for the optimization criteria are presented in Section 5. We show in Section 6 how our framework can be easily integrated for solving a bi-criterion constrained clustering task. Experiments are presented in Section 7, showing the performance and the flexibility of our approach.

2. Preliminaries

2.1. Constrained clustering

Cluster analysis is a Data Mining task that aims at partitioning a given set of objects into homogeneous and/or well-separated subsets, called classes or clusters. It is often formulated as the search for a partition such that the objects inside the same cluster are similar, while being different from the objects belonging to other clusters. These requirements are usually expressed by an optimization criterion and the clustering task is usually defined as finding a partition of objects