Available online at www.sciencedirect.com
ScienceDirect
European Journal of Operational Research 183 (2007) 1067–1085
EUROPEAN JOURNAL OF OPERATIONAL RESEARCH
www.elsevier.com/locate/ejor

Optimization based DC programming and DCA for hierarchical clustering ☆

Le Thi Hoai An [a], Le Hoai Minh [a], Pham Dinh Tao [b,*]

[a] *Laboratory of Theoretical and Applied Computer Science (LITA), UFR MIM, University of Paul Verlaine, Metz, Ile du Saulcy, 57045 Metz, France*

[b] *Laboratory of Modeling, Optimization and Operations Research, LMI, National Institute for Applied Sciences, Rouen BP 08, Place Emile Blondel F 76131 Mont Saint Aignan Cedex, France*

Received 15 October 2004; accepted 15 July 2005
Available online 12 June 2006

Abstract

One of the most promising approaches for clustering is based on methods of mathematical programming. In this paper we propose new optimization methods based on DC (Difference of Convex functions) programming for hierarchical clustering. A bilevel hierarchical clustering model is considered with different optimization formulations. They are all nonconvex, nonsmooth optimization problems for which we investigate attractive DC optimization Algorithms called DCA. Numerical results on some artificial and real-world databases are reported. The results demonstrate that the proposed algorithms are more efficient than related existing methods.

Keywords: Clustering; Multilevel hierarchical clustering; K-means algorithm; Nonsmooth nonconvex programs; DC programming; DCA

1. Introduction

Multilevel hierarchical clustering consists of grouping data objects into a hierarchy of clusters. It has a long history (see e.g., [2,5,15]) and has many important applications in various domains, since many kinds of data, including observational data collected in the human and biological sciences, have a hierarchical, nested, or clustered structure. Hierarchical clustering algorithms are useful to determine hierarchical multicast trees for Gird computing using in e-Science, e-Medicine or e-Commerce, Multimedia conferencing, Large-scale dissemination of timely information, etc.

A hierarchical clustering of a set of objects can be described as a tree, in which the leaves are precisely the objects to be clustered. A hierarchical clustering scheme produces a sequence of clusterings in which each

☆ This work is partially supported by the Conseil Régional Lorraine via the project "Optimisation et Aide à la Décision dans les Systèmes d'Information".

* Corresponding author.
E-mail addresses: lethi@univ-metz.fr (L.T. Hoai An), lehoai@univ-metz.fr (L.H. Minh), pham@insa-rouen.fr (P.D. Tao).

0377-2217/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejor.2005.07.028

clustering is nested into the next clustering in the sequence. Standard existing methods for Multilevel hierarchical clustering are often based upon nonhierarchical clustering algorithms coupled with several iterative control strategies to repeatedly modify an initial clustering (reordering, and reclustering) in search for a better one.

To our knowledge, while mathematical programming is widely used for nonhierarchical clustering problems there exist a few optimization models and techniques for multilevel hierarchical clustering ones. Except the work in [14] we have not found other approaches using mathematical programming model for multilevel hierarchical clustering.

In this paper we investigate an efficient optimization approach for a model of this class, that is bilevel hierarchical clustering. The problem can be stated as follows. Given a set $\mathscr{A}$ of p objects $\mathscr{A} := \{a_j \in \mathbb{R}^n : j = 1, \ldots, p\}$, a measured distance, and an integer k. We are to choose $k+1$ members in $\mathscr{A}$, one as the total centre (the root of the tree) and others as centres of k disjoint clusters, and assign other members of $\mathscr{A}$ to their closest centre. The total centre is defined as the closest object to all centres (in the sense that the sum of distances between it and all centres is the smallest).

Let us mention a practical application widely studied in the networking community that is the network topology identification based on end-to-end measurements. Consider a communication network, in which a sender node transmits information packets to a set of receiver nodes. The receivers are, in this case, the usual "objects" to be clustered. Assume that the routes from the sender to the receivers are fixed. The physical network topology is essentially a graph, where each node corresponds to a physical device (e.g., router, switch, terminal, etc.) and the links correspond to the connections between them. Knowledge of the network topology is essential for tasks like monitoring and provisioning a network.

Our approach is based on DC (Difference of Convex functions) programming – which deals with DC programs, i.e., the minimization of a DC function over a convex set – and DC optimization Algorithm called DCA. They were introduced by Pham Dinh Tao in their preliminary form in 1986 and have been extensively developed since 1993 by Le Thi Hoai An and Pham Dinh Tao to become now classic and more and more popular (see e.g., [8–13,16,17] and references therein). DCA has been successfully applied to many large-scale (smooth or nonsmooth) nonconvex programs in various domains of applied sciences, in particular in data analysis and data mining [1,6,12,19,20], for which it provided often a global solution and proved to be more robust and efficient than standard methods.

In this work different mathematical programs have been proposed for the considered bilevel hierarchical clustering problem. They are all nonconvex, nonsmooth optimization problems that can be reformulated as DC programs in a suitable matrix space. More precisely these appropriate formulations lead us to minimizations of differences of simple convex quadratic functions and nonsmooth convex functions, for which the resulting DCA schemes are all explicit, and very inexpensive. Numerical results on some artificial and real-world databases demonstrate that the proposed algorithms are more efficient than some existing optimization based clustering algorithms.

The paper is organized as follows. Section 2 presents different optimization models for the problem. Section 3 provides an introduction of DC programming and DCA. Section 4 deals with a special realization of DCA to the underlying bilevel hierarchical clustering problem. Computational results are reported in the last section.

2. Optimization formulations

In [14] the authors have proposed two nonsmooth, nonconvex optimization models for this problem in the context of determining a multicast group. They considered the set $\mathscr{A}$ as the set of p nodes in the plane, and the measured distance is the Euclidean distance. In the first model they have considered the artificial centres of clusters, denoted x_i, $i = 1, \ldots, k$, as variables. The total centre is then defined according to the centres x_i via the formula

$$x^* = \frac{1}{k}\sum_{i=1}^{k} x_i. \tag{1}$$