Available online at www.sciencedirect.com
ScienceDirect
European Journal of Operational Research 182 (2007) 612–625
EUROPEAN JOURNAL OF OPERATIONAL RESEARCH
www.elsevier.com/locate/ejor

Discrete Optimization

Integer programming models for the q-mode problem

Girish Kulkarni [a], Yahya Fathi [b,*]

[a] Fedex Express, 3680 Hacks Cross Road, Memphis, TN 38125, United States
[b] Department of Industrial Engineering, North Carolina State University, 430 Daniels Hall, Box 7906, Raleigh, NC 27695-7906, United States

Received 7 September 2005; accepted 2 August 2006
Available online 13 November 2006

Abstract

The q-mode problem is a combinatorial optimization problem that requires partitioning of objects into clusters. We discuss theoretical properties of an existing mixed integer programming (MIP) model for this problem and offer alternative models and enhancements. Through a comprehensive experiment we investigate computational properties of these MIP models. This experiment reveals that, in practice, the MIP approach is more effective for instances containing strong natural clusters and it is not as effective for instances containing weak natural clusters. The experiment also reveals that one of the MIP models that we propose is more effective than the other models for solving larger instances of the problem.

Keywords: Integer programming; Linear programming; Cluster analysis; Data mining

1. Introduction

The problem of forming clusters among a given collection of objects with attribute data is addressed in a number of different articles in the open literature. Its applications include the problem of pattern discovery in computational biology [12], design of cluster-based document retrieval systems [7], creating clusters among a given collection of transactional data in market analysis [4], data mining [6], and the model configuration problem in the context of switching cabinet manufacturing [9], among others. Of course each application could have certain distinguishing characteristics that are unique to that application. Such characteristics typically lead to an appropriate definition for the notion of *similarity* (or *dissimilarity*) between two objects and a corresponding evaluation criterion which could also be unique to that application. For instance, in the context of clustering transactional data, Guha et al. [4] propose the concept of *links* in the algorithm ROCK, while the algorithm CACTUS [2] creates *summaries* from the values and each record is assigned to a cluster based on these summaries. For a general discussion of various dissimilarity measures and their ramifications, see Gordon [3].

In this article we focus on the notion of defining the dissimilarity (i.e., distance) between two vectors as *the number of positions in which the two vectors are not identical* (i.e., $D(U^1, U^2)$ defined in the next section), and we refer to the resulting cluster

* Corresponding author. Tel.: +1 919 515 6417; fax: +1 919 515 5281.
E-mail addresses: girishmk@yahoo.com (G. Kulkarni), fathi@ncsu.edu (Y. Fathi).

0377-2217/$ - see front matter
doi:10.1016/j.ejor.2006.08.039

problem as the q-mode problem. This problem has been previously discussed in [6,9]. The q-mode problem requires partitioning a given collection of objects into clusters and hence it can be considered a special case of the set partitioning problem. We propose several integer programming models for the q-mode problem and discuss their properties. For a comprehensive discussion of mathematical programming models in cluster analysis, see [5].

In Section 2, we formally define the q-mode problem and introduce notation that we shall use in the rest of the paper. We also include a previously developed mixed integer programming (MIP) model for the q-mode problem. In Section 3, we show that a relaxation of this MIP model can be used to obtain optimal solutions for the q-mode problem. In Section 4, we introduce modifications to the existing MIP model and propose a new MIP model for the q-mode problem, with the aim of improving the overall computational requirements for solving the problem. In Section 5, we discuss the results of a computational experiment, and empirically evaluate the proposed approaches. Concluding remarks are in Section 6.

2. Background

2.1. Definition of the q-mode problem and notation

In the context of the q-mode problem our objective is to partition a given collection of objects into q mutually exclusive and collectively exhaustive groups so as to minimize the total *distance* from the objects to the "mode" of the cluster to which each object is assigned. We refer to each object in this context as a *record*, and each record is represented by a vector of size n. Each element (position) of this vector corresponds to an attribute of the object, and we assume that all attributes are categorical in nature, i.e., each attribute can take one of m different values (or categories) for that attribute. Without loss of generality we assume that all attributes have the same number m of possible values or categories, but this assumption can be easily removed with minor adjustments. We refer to the space of all such vectors of size n as CV^{nm}.

For a given collection (group) of records Φ in CV^{nm}, and for each value of $i = 1$ to m and $j = 1$ to n, let $F(i,j) =$ Number of records in the collection Φ that have value i in position j, and let $F^{\max}(j) = \max_i F(i,j)$ and $i^*(j) = \arg\max_i F(i,j)$. Clearly $i^*(j)$ represents the category that is most frequently observed in position j among all members of the collection Φ. If more than one category tie at achieving this maximum value at position j we break the tie arbitrarily and select any one of these categories as $i^*(j)$. We now define *mode* of Φ as a vector of size n where its jth element is $i^*(j)$, for all $j = 1$ to n. We denote this vector by $\mathrm{mod}(\Phi)$ and its jth element by $\mathrm{mod}_j(\Phi) = i^*(j)$ for all j. Note that given a collection of p vectors $\Phi \subseteq CV^{nm}$ its mode vector can be determined efficiently and the corresponding computational requirement is $o(np)$ [9].

Given two vectors U^1 and U^2 of the same size we define the distance between these vectors $D(U^1, U^2)$ as the number of positions at which the two vectors are not identical; i.e., letting $d_j(U^1, U^2) = 1$ for all j such that $u_j^1 \neq u_j^2$, and $d_j(U^1, U^2) = 0$ otherwise, it follows that $D(U^1, U^2) = \sum_{j=1}^{n} d_j(U^1, U^2)$. In these expressions the notation u_j is used to represent the jth element of the vector U.

For a given collection of p vectors in CV^{nm} (i.e., vectors of categorical data as defined above), say $\Phi = \{U^1, \ldots, U^p\}$, it can be shown that the total distance between these vectors and their mode is smaller than or equal to the total distance between these vectors and any other vector of the same size in CV^{nm} [9]. In other words

$$\sum_{k=1}^{p} D[U^k, \mathrm{mod}(\Phi)] = \min_{V \in CV^{nm}} \left\{ \sum_{k=1}^{p} D[U^k, V) \right\}.$$

Clearly $\sum_{k=1}^{p} D[U^k, \mathrm{mod}(\Phi)]$ is a characteristic value of the collection Φ; we denote this value by $MD(\Phi)$ and refer to it as the *total distance of the collection Φ from its mode*. In context, this value is comparable to the total distance of a collection of vectors in $\Re^n$ from their geometric center or their median [10].

We can interpret the distance between a vector $U^k \in \Phi$ and its mode, i.e., $D[U^k, \mathrm{mod}(\Phi]$, as the total number of positions where a replacement of the value (change of category) is required to make this vector identical to the mode vector. An interpretation for $MD(\Phi)$ depends on the specific application at hand, but this value serves as a metric to measure the similarity of objects in the given collection of vectors. Obviously a smaller value for $MD(\Phi)$ implies that the members of the collection Φ are more similar to each other, and a larger value implies otherwise.

We are now prepared to give a formal definition of the q-mode problem.

The q-mode problem: Given a collection of p vectors in CV^{nm} and a positive integer q, partition these