Computers & Operations Research 43 (2014) 280–285

Contents lists available at ScienceDirect

Computers & Operations Research

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

A mixed integer linear model for clustering with variable selection

Stefano Benati [a], Sergio García [b,*]

[a] Dipartimento di Sociologia e Ricerca Sociale, Università degli Studi di Trento, Trento, Italy
[b] Kent Business School, University of Kent, Chatham (Kent), United Kingdom

ARTICLE INFO

Available online 19 October 2013

Keywords:
Clustering
p-median
Variable selection
Radius formulation

ABSTRACT

This paper introduces an extension of the *p*-median problem in which the distance function between units is calculated as the distance sum on the q most important variables out of a set of size m. This model has applications in cluster analysis (for example, in sociological surveys), where analysts have a large list of variables available for inclusion, but only a subset of them (true variables) is appropriate for uncovering the cluster structure. Therefore, researchers must carefully separate the true variables from the other before computing data partitions. Here we show that this problem can be formulated as a mixed integer non-linear optimization model where clustering and variable selection are done simultaneously. Then we provide two different linearizations and compare their performance with the default method of clustering with all the variables (which is a *p*-median model) on a set of artificially generated binary data, showing that the model based on a radius formulation performs the best.

1. Introduction

This paper studies the following clustering problem: suppose that we are given a set $\mathcal{U}=\{u_i\}_{i=1}^{n}$ of statistical units (for example, persons answering a survey) that are measured with a set of quantitative or qualitative features $\mathcal{F}=\{f_k\}_{k=1}^{m}$ (for example, questions of the survey). This information is collected in a data matrix $V=[v_{ik}]$, where v_{ik} is the value that feature f_k takes for unit u_i. Our goal is to find a subset of variables $Q\subseteq\mathcal{F}$ of fixed size q and to cluster the n statistical units into p clusters such that the resulting clustering of $\mathcal{U}$ is the most accurate when only the information of the variables of Q is used to decide the clusters.

This problem emerges in Statistics, where it is recognized that not all the variables are equally important in uncovering cluster structure because including them deteriorates the effectiveness of the clustering procedures [16,4] up to the point that misclassification may become a serious problem. Variables that do not define cluster structure are called "masking variables" [6] in order to differentiate them from the other "true variables". Two methods considering variable selection that can be found in the clustering literature are variable weighting and model based clustering. Variable weighting [5] consists on determining weights for each variable so that the distance function between units is not affected by the masking variables. In model-based clustering [14,19] the units are assumed to have been generated with certain probabilistic models and clustering is done by learning the associated parameters and probabilities. Other methods can be found in [22]. It must be noted that all these methods are applied to continuous variables and are computationally demanding.

In this paper we propose a combinatorial model for clustering that selects simultaneously the best set $Q\subseteq\mathcal{F}$ of variables, the best set of medians $P\subseteq\mathcal{U}$ and the optimal data partition when the criterion used is the minimization of the total distance inside the clusters between the median of the cluster and the units that belong to the cluster. This model can be seen as an extension of the classical *p*-median problem where the computation of the distances depends on the selected variables Q. It is formulated as follows:

$$\min_{\substack{P\subseteq\mathcal{U},\ |P|=p,\\ Q\subseteq\mathcal{F},\ |Q|=q}} \sum_{u_i\in\mathcal{U}} \min\left\{d_{ij}^{Q}/u_j\in P\right\},$$

where d_{ij}^{Q} is the distance between units i and j restricted to the features of Q. Particularly, we use the Manhattan distance or distance ℓ_1, that is one of the most commonly used to cluster ordinal or qualitative data, as in the applications that motivated this research (see [1,3]). The Euclidean distance is also quite popular in clustering (see, for example, [18]).

Almost all the previous papers in the literature on clustering with variable selection have dealt exclusively with continuous variables and, therefore, cannot be directly compared to the research of this paper, which is focused on binary data. The only exception is [4], where a heuristic method is developed for

* Corresponding author. Tel.: +44 131 650 5038.
E-mail addresses: Stefano.Benati@unitn.it (S. Benati), S.Garcia-Quiles@kent.ac.uk, sergio.garcia-quiles@ed.ac.uk (S. García).

0305-0548/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cor.2013.10.005

clustering and selecting variables in the case of binary data. This method, based on a k-means approach [15], determines a promising initial subset of four features and then increases it one by one while the objective function of the k-means satisfies a certain condition. The model that we propose in this paper is different because it is an exact method and we changed from a k-means objective function to a p-median objective function. A p-median approach has a long tradition in the clustering literature [20,17,12,10,23] and, moreover, it is known that it has some advantages in terms of robustness and interpretation, for example, because the median representing the cluster is an element of the sample [13].

As will be shown in Section 2, the p-median model can be extended to consider the decision on what variables $Q \subseteq \mathcal{F}$ to select, but the natural formulation of this extension leads to a quadratic non-convex problem. Instead of developing new solution tools for this non-linear model, our approach is to study different mixed integer linearizations and to determine which one is the most efficient. The first formulation is a direct linearization of the initial quadratic model and the second is based on the so-called radius formulation of the p-median problem [7].

The rest of the paper is organized as follows. In Section 2 the non-linear model and the two linearizations are formulated. A computational study is carried out in Section 3 and, finally, some conclusions are given in Section 4.

2. Problem definition and formulations

Assume that we are given a sample $\mathcal{U} = \{u_i\}_{i=1}^n$ of statistical units. For every unit i, the set $\mathcal{F} = \{f_k\}_{k=1}^m$ of statistical variables (features) is measured. We assume that, as is common in opinion polling or attitude surveys, variables f_k are represented by qualitative or ordinal data. If the data are qualitative, they are represented by 0–1. If the data are ordinal with g occurrences, or they are represented by a Likert scale with a finite number g of tiers, then we will refer to g as the dimension of the scale.

Let v_{ik} be the record of variable k for unit i. The distance, or difference, between unit u_i and unit u_j with respect to the feature f_k is $d_{ijk} = |v_{ik} - v_{jk}|$ and the overall distance between u_i and u_j is the 1-norm:

$$d_{ij} = \sum_{k=1}^{m} d_{ijk} = \sum_{k=1}^{m} |v_{ik} - v_{jk}|.$$

Suppose now that only a subset $Q \subseteq \mathcal{F}$ of statistical variables are considered relevant for the analysis and that, as a consequence, the differences between the units are calculated using Q only. The distance formula is thus expressed using the incidence vector z of subset Q:

$$d_{ij} = \sum_{k=1}^{m} d_{ijk} z_k,$$

where $z_k = 1$ if $f_k \in Q$ and $z_k = 0$ otherwise.

The units are clustered using the p-median model and the min-sum criterion, so that its outcome consists of p clusters and its median is the most representative element (the cluster archetype). We define binary variables y_j, $j = 1, \ldots, n$, that take value one if unit j is a median (and zero otherwise), and binary allocation variables x_{ij}, $i, j = 1, \ldots, n$, that take value one if unit i is assigned to cluster j (and zero otherwise). Then, imposing that only a subset $Q \subseteq \mathcal{F}$ of variables is to be used, we introduce binary variables z_k, $k = 1, \ldots, m$, where $z_k = 1$ if $f_k \in Q$ (and $z_k = 0$ otherwise). The model obtained is the following:

$$(F_1)\begin{cases} \text{Min.} & \displaystyle\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\sum_{k=1}^{m} d_{ijk} z_k\right) x_{ij} & \\ \text{s.t.} & x_{ij} \le y_j, & i,j = 1, \ldots, n, \\ & \displaystyle\sum_{j=1}^{n} x_{ij} = 1, & i = 1, \ldots, n, \\ & \displaystyle\sum_{j=1}^{n} y_j = p, & \\ & \displaystyle\sum_{k=1}^{m} z_k = q, & \\ & x_{ij} \ge 0, & i,j = 1, \ldots, n, \\ & y_j \in \{0,1\}, & j = 1, \ldots, n, \\ & z_k \in \{0,1\}, & k = 1, \ldots, m. \end{cases}$$

This formulation is non-linear because of the quadratic terms in the objective function. Moreover, the objective function is non-convex because the distance matrix is not positive semidefinite (the terms d_{ijk} can be arranged in such a way that the matrix is composed of zeros on the main diagonal and has positive or zero terms elsewhere). Since the potential to solve to optimality large instances of a non-convex quadratic model is quite low, our goal is to linearize this formulation F_1.

Finally, it must be remarked that the classical p-median problem is a particular case of this problem that appears when all variables z_k take value one.

2.1. Direct linearization formulation

Formulation F_1 can be linearized quite straightforwardly by introducing the new variables

$$w_{ijk} = x_{ij} z_k, \; i,j = 1, \ldots, n, \; k = 1, \ldots, m,$$

plus some well-definition constraints, which is a standard procedure in the literature. The linearized model is

$$(F_2)\begin{cases} \text{Min.} & \displaystyle\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{m} d_{ijk} w_{ijk} & & \\ \text{s.t.} & x_{ij} \le y_j, & i,j = 1, \ldots, n, & \\ & \displaystyle\sum_{j=1}^{n} x_{ij} = 1, & i = 1, \ldots, n, & \\ & \displaystyle\sum_{j=1}^{n} y_j = p, & & \\ & \displaystyle\sum_{k=1}^{m} z_k = q, & & \\ & w_{ijk} \ge x_{ij} + z_k - 1, & i,j = 1, \ldots, n, \; k = 1, \ldots, m, & \\ & w_{ijk} \le x_{ij}, & i,j = 1, \ldots, n, \; k = 1, \ldots, m, & \text{(a)} \\ & w_{ijk} \le z_k, & i,j = 1, \ldots, n, \; k = 1, \ldots, m, & \text{(b)} \\ & w_{ijk} \ge 0, & i,j = 1, \ldots, n, \; k = 1, \ldots, m, & \\ & x_{ij} \ge 0, & i,j = 1, \ldots, n, & \\ & y_j \in \{0,1\}, & j = 1, \ldots, n, & \\ & z_k \in \{0,1\}, & k = 1, \ldots, m. & \end{cases} \quad (1)$$

We are not imposing that variables x_{ij} must be binary, but only positive because we have a minimization problem in which distances d_{ijk} are positive, and variables z_k and y_j are binary, meaning that there is an optimal solution where all variables x_{ij} are binary. Besides, inequalities (1a) and (1b) can be dropped out because they are satisfied at every optimal solution.