Available online at www.sciencedirect.com
SCIENCE DIRECT®
European Journal of Operational Research 170 (2006) 578–596
EUROPEAN JOURNAL OF OPERATIONAL RESEARCH
www.elsevier.com/locate/ejor

Computing, Artificial Intelligence and Information Technology

A new nonsmooth optimization algorithm for minimum sum-of-squares clustering problems

Adil M. Bagirov, John Yearwood *

Centre for Informatics and Applied Optimization, School of Information Technology and Mathematical Sciences, University of Ballarat, P.O. Box 663, Vic. 3353, Australia

Received 25 November 2002; accepted 17 June 2004
Available online 26 August 2004

Abstract

The minimum sum-of-squares clustering problem is formulated as a problem of nonsmooth, nonconvex optimization, and an algorithm for solving the former problem based on nonsmooth optimization techniques is developed. The issue of applying this algorithm to large data sets is discussed. Results of numerical experiments have been presented which demonstrate the effectiveness of the proposed algorithm.

Keywords: Nonsmooth optimization; Cluster analysis; Minimum sum-of-squares clustering

1. Introduction

Clustering is the *unsupervised* classification of the patterns. Cluster analysis deals with the problems of organization of a collection of patterns into clusters based on similarity. It has found many applications, including information retrieval, document extraction, image segmentation, etc.

In cluster analysis we assume that we have been given a set X of a finite number of points of d-dimensional space $\mathbb{R}^d$, that is

$$X = \{x^1, \ldots, x^n\}, \quad \text{where } x^i \in \mathbb{R}^d, \ \ i = 1, \ldots, n.$$

The subject of cluster analysis is the partition of the set X into a given number q of overlapping or disjoint subsets C_i, $i = 1,\ldots,q$, with respect to predefined criteria such that

* Corresponding author. Tel.: +61 3 5327 9272; fax: +61 3 5327 9966.
E-mail address: j.yearwood@ballarat.edu.au (J. Yearwood).

0377-2217/$ - see front matter
doi:10.1016/j.ejor.2004.06.014

$$X = \bigcup_{i=1}^{q} C_i.$$

The sets C_i, $i = 1, \ldots, q$, are called clusters. The clustering problem is said to be *hard clustering* if every data point belongs to one and only one cluster. Unlike hard clustering in the *fuzzy clustering* problem the clusters are allowed to overlap and instances have degrees of appearance in each cluster. In this paper we will exclusively consider the hard unconstrained clustering problem, that is we additionally assume that

$$C_i \cap C_k = \emptyset, \quad \forall i, k = 1, \ldots, q, \ \ i \neq k,$$

and no constraints are imposed on the clusters C_i, $i = 1, \ldots, q$. Thus every point $x \in X$ is contained in exactly one and only one set C_i.

Each cluster C_i can be identified by its center (or centroid). Then the clustering problem can be reduced to the following optimization problem (see [7,8,38]):

$$\begin{aligned} &\text{minimize} \quad \varphi(C, a) = \frac{1}{n} \sum_{i=1}^{q} \sum_{x \in C_i} \|a^i - x\|^2 \\ &\text{subject to} \quad C \in \overline{C}, \quad a = (a^1, \ldots, a^q) \in \mathbb{R}^{d \times q}, \end{aligned} \tag{1}$$

where $\|\cdot\|$ denotes the Euclidean norm, $C = \{C_1, \ldots, C_q\}$ is a set of clusters, $\overline{C}$ is a set of all possible q-partitions of the set X, a^i is the center of the cluster C_i, $i = 1, \ldots, q$,

$$a^i = \frac{1}{|C_i|} \sum_{x \in C_i} x,$$

and $|C_i|$ is a cardinality of the set C_i, $i = 1, \ldots, q$. The problem (1) is also known as the minimum sum-of-squares clustering. The combinatorial formulation (1) of the minimum sum-of-squares clustering is not suitable for direct application of mathematical programming techniques. The problem (1) can be rewritten as the following mathematical programming problem:

$$\begin{aligned} &\text{minimize} \quad \psi(a, w) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{q} w_{ij} \|a^j - x^i\|^2 \\ &\text{subject to} \quad \sum_{j=1}^{q} w_{ij} = 1, \ \ i = 1, \ldots, n, \end{aligned} \tag{2}$$

and

$$w_{ij} \in \{0, 1\}, \quad i = 1, \ldots, n, \ \ j = 1, \ldots, q.$$

Here

$$a^j = \frac{\sum_{i=1}^{n} w_{ij} x^j}{\sum_{i=1}^{n} w_{ij}}, \quad j = 1, \ldots, q,$$

and w_{ij} is the association weight of pattern x^i with cluster j (to be found), given by

$$w_{ij} = \begin{cases} 1 & \text{if pattern } i \text{ is allocated to cluster } j \ \forall i = 1, \ldots, n, \ \ j = 1, \ldots, q, \\ 0 & \text{otherwise,} \end{cases}$$

w is an $n \times q$ matrix.