Pattern Recognition 44 (2011) 70–77

Contents lists available at ScienceDirect

Pattern Recognition

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

Solving the minimum sum-of-squares clustering problem by hyperbolic smoothing and partition into boundary and gravitational regions

Adilson Elias Xavier*, Vinicius Layter Xavier

Department of Systems Engineering and Computer Science, Graduate School of Engineering (COPPE), Federal University of Rio de Janeiro, P.O. Box 68511,
Rio de Janeiro RJ 21941-972, Brazil

ARTICLE INFO

Article history:
Received 11 January 2010
Received in revised form
20 May 2010
Accepted 6 July 2010

Keywords:
Cluster analysis
Min-sum-min problems
Nondifferentiable programming
Smoothing

ABSTRACT

This article considers the minimum sum-of-squares clustering (MSSC) problem. The mathematical modeling of this problem leads to a *min-sum-min* formulation which, in addition to its intrinsic bi-level nature, has the significant characteristic of being strongly nondifferentiable. To overcome these difficulties, the proposed resolution method, called hyperbolic smoothing, adopts a smoothing strategy using a special C^∞ differentiable class function. The final solution is obtained by solving a sequence of low dimension differentiable unconstrained optimization subproblems which gradually approach the original problem. This paper introduces the method of partition of the set of observations into two nonoverlapping groups: "data in frontier" and "data in gravitational regions". The resulting combination of the two methodologies for the MSSC problem has interesting properties, which drastically simplify the computational tasks.

1. Introduction

Cluster analysis deals with the problems of classification of a set of patterns or observations, in general represented as points in a multidimensional space, into clusters, following two basic and simultaneous objectives: patterns in the same clusters must be similar to each other (homogeneity objective) and different from patterns in other clusters (separation objective) [1–3].

Clustering is an important problem that appears in a broad spectrum of applications, whose intrinsic characteristics engender many approaches to this problem, as described by Dubes and Jain [4], Jain and Dubes [5] and Hansen and Jaumard [6].

Clustering analysis has been used traditionally in disciplines such as: biology, biometry, psychology, psychiatry, medicine, geology, marketing and finance. Clustering is also a fundamental tool in modern technology applications, such as: pattern recognition, data mining, web mining, image processing, machine learning and knowledge discovering.

In this paper, a particular clustering problem formulation is considered. Among many criteria used in cluster analysis, the most natural, intuitive and frequently adopted criterion is the minimum sum-of-squares clustering (MSSC). This criterion corresponds to the minimization of the sum-of-squares of distances of observations to their cluster means, or equivalently, to the minimization of within-group sum-of-squares. It is a criterion for both the homogeneity and the separation objectives. According to the Huygens Theorem, minimizing the within-cluster inertia of a partition (homogeneity within the cluster) is equivalent to maximizing the between-cluster inertia (separation between clusters).

The minimum sum-of-squares clustering (MSSC) formulation produces a mathematical problem of global optimization. It is both a nondifferentiable and a nonconvex mathematical problem, with a large number of local minimizers.

There are two main strategies for solving clustering problems: hierarchical clustering methods and partition clustering methods. Hierarchical methods produce a hierarchy of partitions of a set of observations. Partition methods, in general, assume a given number of clusters and, essentially, seek the optimization of an objective function measuring the homogeneity within the clusters and/or the separation between the clusters.

For the sake of completeness, we present first the Hyperbolic Smoothing Clustering Method (HSCM), Xavier [7]. Basically the method performs the smoothing of the nondifferentiable *min-sum-min* clustering formulation. This technique was developed through an adaptation of the hyperbolic penalty method originally introduced by Xavier [8]. By smoothing, we fundamentally mean the substitution of an intrinsically nondifferentiable two-level problem by a C^∞ unconstrained differentiable single-level alternative.

Additionally, the paper presents a new, faster, methodology. The basic idea is the partition of the set of observations into two nonoverlapping parts. By using a conceptual presentation, the first set corresponds to the observation points relatively close to two or more centroids. This set of observations, named boundary

* Corresponding author.
E-mail addresses: adilson@cos.ufrj.br (A.E. Xavier), vinicius@cos.ufrj.br (V.L. Xavier).

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.07.004

band points, can be managed by using the previously presented smoothing approach. The second set corresponds to observation points significantly closer to a single centroid in comparison with others. This set of observations, named gravitational points, is managed in a direct and simple way, offering much faster performance.

This work is organized in the following way. A step-by-step definition of the minimum sum-of-squares clustering problem is presented in the next section. The original hyperbolic smoothing approach and the derived algorithm are presented in Section 3. The boundary and gravitational regions partition scheme and the new derived algorithm are presented in Section 4. Computational results are presented in Section 5. Brief conclusions are drawn in Section 6.

2. The minimum sum-of-squares clustering problem

Let $S=\{s_1,\ldots,s_m\}$ denote a set of m patterns or observations from an Euclidean n-space, to be clustered into a given number q of disjoint clusters. To formulate the original clustering problem as a *min-sum-min* problem, we proceed as follows. Let x_i, $i=1,\ldots,q$ be the centroids of the clusters, where each $x_i \in \mathbb{R}^n$. The set of these centroid coordinates will be represented by $X \in \mathbb{R}^{nq}$. Given a point s_j of S, we initially calculate the Euclidian distance from s_j to the center in X that is nearest. This is given by

$$z_j = \min_{i=1,\ldots,q} \|s_j - x_i\|_2. \tag{1}$$

The most frequent measurement of the quality of a clustering associated to a specific position of q centroids is provided by the sum of the squares of these distances, which determines the MSSC problem:

$$\begin{aligned} &\text{minimize} \quad \sum_{j=1}^{m} z_j^2 \\ &\text{subject to} \quad z_j = \min_{i=1,\ldots,q} \|s_j - x_i\|_2, \quad j=1,\ldots,m \end{aligned} \tag{2}$$

3. The hyperbolic smoothing clustering method

Considering its definition, each z_j must necessarily satisfy the following set of inequalities:

$$z_j - \|s_j - x_i\|_2 \le 0, \quad i=1,\ldots,q. \tag{3}$$

Substituting these inequalities for the equality constraints of problem (2), the relaxed problem is produced:

$$\begin{aligned} &\text{minimize} \quad \sum_{j=1}^{m} z_j^2 \\ &\text{subject to} \quad z_j - \|s_j - x_i\|_2 \le 0, \quad j=1,\ldots,m, \; i=1,\ldots,q. \end{aligned} \tag{4}$$

Since the variables z_j are not bounded from below, the optimum solution of the relaxed problem will be $z_j=0$, $j=1, \ldots, m$. In order to obtain the desired equivalence, we must, therefore, modify problem (4). We do so by first letting $\varphi(y)$ denote $\max\{0, y\}$ and then observing that, from the set of inequalities in (4), it follows that

$$\sum_{i=1}^{q} \varphi(z_j - \|s_j - x_i\|_2) = 0, \quad j=1,\ldots,m. \tag{5}$$

Using (5) in place of the set of inequality constraints in (4), we would obtain an equivalent problem maintaining the undesirable property that $z_j, j=1, \ldots, m$ still has no lower bound. Considering, however, that the objective function of problem (4) will force each z_j, $j=1,\ldots,m$, downward, we can think of bounding the latter variables from below by including an ε perturbation in (5). So, the following modified problem is obtained:

$$\begin{aligned} &\text{minimize} \quad \sum_{j=1}^{m} z_j^2 \\ &\text{subject to} \quad \sum_{i=1}^{q} \varphi(z_j - \|s_j - x_i\|_2) \ge \varepsilon, \quad j=1,\ldots,m \end{aligned} \tag{6}$$

for $\varepsilon > 0$. Since the feasible set of problem (2) is the limit of that of (6) when $\varepsilon \to 0_+$, we can then consider solving (2) by solving a sequence of problems like (6) for a sequence of decreasing values for ε that approaches 0.

Analyzing the problem (6), the definition of function φ endows it with an extremely rigid nondifferentiable structure, which makes its computational solution very hard. In view of this, the numerical method we adopt for solving problem (1), takes a smoothing approach. From this perspective, let us define the function:

$$\phi(y,\tau) = (y + \sqrt{y^2 + \tau^2})/2 \tag{7}$$

for $y \in \mathbb{R}$ and $\tau > 0$.

Function ϕ has the following properties:

(a) $\phi(y,\tau) > \varphi(y), \forall \tau > 0$;
(b) $\lim_{\tau \to 0} \phi(y,\tau) = \varphi(y)$;
(c) $\phi(y,\tau)$ is an increasing convex C^∞ function in variable y.

By using function ϕ in the place of function φ, the problem

$$\begin{aligned} &\text{minimize} \quad \sum_{j=1}^{m} z_j^2 \\ &\text{subject to} \quad \sum_{i=1}^{q} \phi(z_j - \|s_j - x_i\|_2, \tau) \ge \varepsilon, \quad j=1,\ldots,m \end{aligned} \tag{8}$$

is produced.

Now, the Euclidean distance $\|s_j - x_i\|_2$ is the single nondifferentiable component on problem (8). So, to obtain a completely differentiable problem, it is still necessary to smooth it. For this purpose, let us define the function

$$\theta(s_j, x_i, \gamma) = \sqrt{\sum_{l=1}^{n} (s_j^l - x_i^l)^2 + \gamma^2} \tag{9}$$

for $\gamma > 0$.

Function θ has the following properties:

(a) $\lim_{\gamma \to 0} \theta(s_j, x_i, \gamma) = \|s_j - x_i\|_2$;
(b) θ is a C^∞ function.

By using function θ in place of the distance $\|s_j - x_i\|_2$, the following completely differentiable problem is now obtained:

$$\begin{aligned} &\text{minimize} \quad \sum_{j=1}^{m} z_j^2 \\ &\text{subject to} \quad \sum_{i=1}^{q} \phi(z_j - \theta(s_j, x_i, \gamma), \tau) \ge \varepsilon, \quad j=1,\ldots,m. \end{aligned} \tag{10}$$

So, the properties of functions ϕ and θ allow us to seek a solution to problem (6) by solving a sequence of subproblems like problem (10), produced by the decreasing of the parameters $\gamma \to 0$, $\tau \to 0$, and $\varepsilon \to 0$.

Since $z_j \ge 0, j=1,\ldots,m$, the objective function minimization process will work for reducing these values to the utmost. On the