Solving the Minimum Sum of L1 Distances Clustering Problem by Hyperbolic Smoothing and Partition into Boundary and Gravitational Regions

Adilson Elias Xavier
Vinicius Layter Xavier
Dept. of Systems Engineering and Computer Science
Graduate School of Engineering (COPPE)
Federal University of Rio de Janeiro
Rio de Janeiro, BRAZIL
e-mail: {adilson,vinicius}@cos.ufrj.br

Abstract

This article considers the minimum sum of distances clustering (MSDC) problem, where the distances are measured through the L1 or Manhattan metric. 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 uses also the method of partition of the set of observations into two non overlapping groups: "data in frontier" and "data in gravitational regions". The resulting combination of the two methotologies for the MSDC problem has interesting properties: complete differentiability and drastic simplification of computational tasks.

Keywords: Cluster Analysis, Min-Sum-Min Problems, Manhattan Metric, Nondifferentiable Programming, Smoothing

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) [Anderberg (1973), Hartingan (1975) and Späth (1980)].

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 (1976), Jain and Dubes (1988) and Hansen and Jaumard (1997).

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, a frequently adopted criterion is the minimum sum of L1 distances clustering (MSDC). This criterion corresponds to the minimization of the sum of distances of observations to their centroids, where the distances are measured through the L1 or Manhattan metric. As broadly recorded by the litterature, the L1 metric has the advantage of offering more stable solutions.

Similarly to others clustering formulations, it 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 Smothing Clustering Method (HSCM), Xavier (2010). Basically the method performs the smoothing of the nondifferentiable $min - sum - min$ problem engendered by the modeling of a broad class of clustering problems, including the minimum sum of L1 distances clustering (MSDC) formulation. This technique was developed through an adaptation of the hyperbolic penalty method originally introduced by Xavier (1982). 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 faster methodology applied to the specific considered problem. The basic idea is the partition of the set of observations into two non overlapping 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 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 for the specific minimum sum of L1 or Manhattan distances clustering (MSDC) formulation. The same partition scheme was presented first by Xavier and Xavier (2010) in order to solve the the specific minimum sum of distances clustering (MSSC) formulation.

This work is organized in the following way. A step-by-step definition of the minimum sum of L1 distances clustering problem is presented in the next section. The original smoothing 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. Brief conclusions are drawn in section 5.

2 The Minimum Sum of L1 Distances Clustering Problem

Let $S = \{s_1, \dots, s_m\}$ 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 $x_i, i = 1, \dots, q$ 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 L1 distance from s_j to the center in X that is nearest. This is given by

$$z_j = \min_{x_i \in X} \|s_j - x_i\|_1. \tag{1}$$

A frequent measurement of the quality of a clustering associated to a specific position of q centroids is provided by the sum of the L1 distances, which determines the MSDC problem:

$$\begin{aligned} & \mathit{minimize} \sum_{j=1}^{m} z_j \\ \text{subject to} \quad & z_j = \min_{i=1,\dots,q} \|s_j - x_i\|_1, \quad j = 1, \dots, m \end{aligned} \tag{2}$$

3 The Hyperbolic Smoothing Clustering Method

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

$$z_j - \|s_j - x_i\|_1 \le 0, \quad i = 1, \dots, q. \tag{3}$$

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

$$\begin{aligned} & \mathit{minimize} \sum_{j=1}^{m} z_j \\ \text{subject to} \quad & z_j - \|s_j - x_i\|_1 \le 0, \quad j = 1, \dots, m, \; i = 1, \dots, q. \end{aligned} \tag{4}$$

Since the variables z_j are not bounded from below, the optimum solution of the relaxed problem will be $z_j = 0,\ j = 1, \dots, m$. 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

$$\sum_{i=1}^{q} \varphi(z_j - \|s_j - x_i\|_1\ ) = 0, \qquad j = 1, \dots, m. \tag{5}$$

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

$$minimize \ \sum_{j=1}^{m} z_j \tag{6}$$

$$\text{subject to} \quad \sum_{i=1}^{q} \varphi(z_j - \|s_j - x_i\|_1\ ) \ \geq \ \varepsilon\ , \qquad j = 1, \dots, m$$

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:

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

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

Function ϕ has the following properties:

(a) $\phi(y, \tau) > \varphi(y), \quad \forall\, \tau > 0;$

(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

$$minimize \sum_{j=1}^{m} z_j \tag{8}$$

$$\text{subject to} \quad \sum_{i=1}^{q} \phi(z_j - \|s_j - x_i\|_1, \tau) \geq \varepsilon, \qquad j = 1, \ldots, m.$$

is produced.

To obtain a differentiable problem, it is still necessary to smooth the L1 distance $\|s_j - x_i\|_1$. For this purpose, let us define the function

$$\theta_1(\, s_j \,,\, x_i \,,\, \gamma\,) \;=\; \sqrt{\sum_{l=1}^{n} \|s_j^l - x_i^l\|^2 \,+\, \gamma^2} \tag{9}$$

for $\gamma > 0$.

Function θ_1 has the following properties:

(a) $\lim_{\gamma \to 0} \theta_1(\, s_j \,,\, x_i \,,\, \gamma\,) \;=\; \|s_j - x_i\|_1 \,;$

(b) θ_1 is a C^∞ function.

By using function θ_1 in place of the distance $\|s_j - x_i\|_1$, the following completely differentiable problem is now obtained:

$$\begin{aligned} & \textit{minimize} \ \sum_{j=1}^{m} z_j && (10) \\ & \text{subject to} \quad \sum_{i=1}^{q} \phi(z_j - \theta_1(s_j, x_i, \gamma), \tau) \geq \varepsilon, \quad j = 1, \ldots, m. \end{aligned}$$

So, the properties of functions ϕ and θ_1 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 $\gamma \rightarrow 0$, $\tau \rightarrow 0$, and $\varepsilon \rightarrow 0$.

Since $z_j \geq 0, j = 1, \ldots, m$, the objective function minimization process will work for reducing these values to the utmost. On the other hand, given any set of centroids $x_i, i = 1, \ldots, q$, due to property (c) of the hyperbolic smoothing function ϕ, the constraints of problem (10) are a monotonically increasing function in z_j. So, these constraints will certainly be active and problem (10) will at last be equivalent to problem:

$$\begin{aligned} & \textit{minimize} \ \sum_{j=1}^{m} z_j && (11) \\ & \text{subject to} \quad h_j(z_j, x) = \sum_{i=1}^{q} \phi(z_j - \theta_1(s_j, x_i, \gamma), \tau) - \varepsilon = 0, \quad j = 1, \ldots, m. \end{aligned}$$

The dimension of the variable domain space of problem (11) is $(nq+m)$. As, in general, the value of the parameter m, the cardinality of the set S of the observations s_j, is large, problem (11) has a large number of variables. However, it has a separable structure, because each variable z_j appears only in one equality constraint. Therefore, as the partial derivative of $h(z_j, x)$ with respect to $z_j, j = 1, \ldots, m$ is not equal to zero, it is possible to use the Implicit Function Theorem to calculate each component $z_j, j = 1, \ldots, m$ as a function of the centroid variables $x_i, i = 1, \ldots, q$. In this way, the unconstrained problem

$$\textit{minimize} \ f(x) = \sum_{j=1}^{m} z_j(x) \tag{12}$$

is obtained, where each $z_j(x)$ results from the calculation of a zero of each equation

$$h_j(z_j, x) = \sum_{i=1}^{q} \phi(z_j - \theta_1(s_j, x_i, \gamma), \tau) - \varepsilon = 0, \quad j = 1, \dots, m. \tag{13}$$

Due to property (c) of the hyperbolic smoothing function, each term ϕ above is strictly increasing with variable z_j and therefore the equation has a single zero.

Again, due to the Implicit Function Theorem, the functions $z_j(x)$ have all derivatives with respect to the variables $x_i,\ i = 1, \dots, q,$ and therefore it is possible to calculate the gradient of the objective function of problem (12),

$$\nabla f(x) = \sum_{j=1}^{m} \nabla z_j(x) \tag{14}$$

where

$$\nabla z_j(x) = -\nabla h_j(z_j, x) \;/\; \frac{\partial h_j(z_j, x)}{\partial z_j}, \tag{15}$$

while $\nabla h_j(z_j, x)$ and $\partial h_j(z_j, x)/\partial z_j$ are obtained from equations (7), (9) and (13).

In this way, it is easy to solve problem (12) by making use of any method based on first order derivative information. At last, it must be emphasized that problem (12) is defined on a $(nq)-$dimensional space, so it is a small problem, since the number of clusters, q, is, in general, very small for real applications.

The solution of the original clustering problem can be obtained by using the Hyperbolic Smoothing Clustering Algorithm, described below in a simplified form.

The Simplified HSC Algorithm

Initialization Step: Choose initial values: x^0, γ^1 , τ^1 , ε^1.

Choose values $0 < \rho_1 < 1$, $0 < \rho_2 < 1$, $0 < \rho_3 < 1$; let $k = 1$.

Main Step: Repeat until a stopping rule is attained

Solve problem (12) with $\gamma = \gamma^k$, $\tau = \tau^k$ and $\varepsilon = \varepsilon^k$, starting at the initial point x^{k-1} and let x^k be the solution obtained.

Let $\gamma^{k+1} = \rho_1 \gamma^k$, $\tau^{k+1} = \rho_2 \tau^k$, $\varepsilon^{k+1} = \rho_3 \varepsilon^k$, $k := k + 1$. ■

Just as in other smoothing methods, the solution to the clustering problem is obtained, in theory, by solving an infinite sequence of optimization problems. In the HSC algorithm, each problem to be minimized is unconstrained and of low dimension.

Notice that the algorithm causes τ and γ to approach 0, so the constraints of the subproblems as given in (10) tend to those of (6). In addition, the algorithm causes ε to approach 0, so, in a simultaneous movement, the solved problem (6) gradually approaches the original MSDC problem (2).

4 The Accelerated Hyperbolic Smoothing Clustering Method

The calculation of the objective function of the problem (12) demands the determination of the zeros of m equations (13), one equation for each observation point. This is a relevant computational task associated to HSC Algorithm.

In this section, it is presented a faster procedure. The basic idea is the partition of the set of observations into two non overlapping regions. By using a conceptual presentation, the first region corresponds to the observation points that are relatively close to two or more centroids. The second region corresponds to the observation points that are significantly close to a unique centroid in comparison with the other ones.

So, the first part J_B is the set of boundary observations and the second is the set J_G of gravitational observations. Considering this partition, equation (12) can be expressed in the following way:

$$minimize\ f(x) = \sum_{j=1}^{m} z_j(x) = \sum_{j\in J_B} z_j(x) + \sum_{j\in J_G} z_j(x), \tag{16}$$

so that the objective function can be presented in the form:

$$minimize\ f(x) = f_B(x) + f_G(x), \tag{17}$$

where the two components are completely independent.

The first part of expression (17), associated with the boundary observations, can be calculated by using the previous presented smoothing approach, see (12) and (13):

$$minimize\ f_B(x) = \sum_{j\in J_B} z_j(x), \tag{18}$$

where each $z_j(x)$ results from the calculation of a zero of each equation

$$h_j(z_j, x) = \sum_{i=1}^{q} \phi(z_j - \theta_1(s_j, x_i, \gamma), \tau) - \varepsilon = 0, \quad j \in J_B. \tag{19}$$

The second part of expression (17) can be calculated by using a faster procedure, as we will show right away.

Let us define the two parts in a more rigorous form. Let be $\overline{x}_i,\ i = 1, \dots, q$ be a referential position of centroids of the clusters taken in the iterative process.

The boundary concept in relation to the referential point $\overline{x}$ can be easily specified by defining a δ band zone between neighboring centroids. For a generic point $s \in \mathbb{R}^n$, we define the first and second nearest distances from s to the centroids:

$$d_1(s, \overline{x}) = \| s - \overline{x}_{i_1} \| = \min_i \| s - \overline{x}_i \| \tag{20}$$

$$d_2(s,\overline{x}) \;=\; \|\, s - \overline{x}_{i_2} \,\| \;=\; \min_{i \neq i_1} \|\, s - \overline{x}_i \,\| \;, \tag{21}$$

where i_1 and i_2 are the labeling indexes of these two nearest centroids.

By using the above definitions, let us define precisely the δ boundary band zone:

$$Z_\delta(\overline{x}) \;=\; \{ s \in \mathbb{R}^n \mid d_2(s,\overline{x}) \;-\; d_1(s,\overline{x}) \;<\; 2\,\delta \,\} \tag{22}$$

and the gravity region, this is the complementary space:

$$G_\delta(\overline{x}) \;=\; \{ s \in \mathbb{R}^n \;-\; Z_\delta(\overline{x}) \,\} \,. \tag{23}$$

Figure 1 illustrates in $\mathbb{R}^2$ the $Z_\delta(\overline{x})$ and $G_\delta(\overline{x})$ partitions. The central lines form the Voronoy polygon associated with the referential centroids $\overline{x}_i$, $i = 1, \ldots, q$. The region between two parallel lines to Voronoy lines constitutes the boundary band zone $Z_\delta(\overline{x})$.

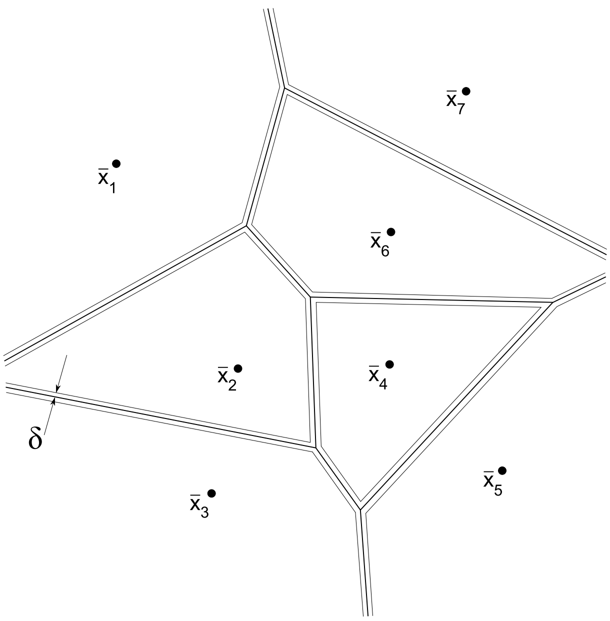

Figure 1: The $Z_\delta(\overline{x})$ and $G_\delta(\overline{x})$ partitions.

Now, the sets J_B and J_G can be defined in a precise form:

$$J_B(\overline{x}) = \{j = 1, \ldots, m \,|\, s_j \in Z_\delta(\overline{x})\}\,, \tag{24}$$

$$J_G(\overline{x}) = \{j = 1, \ldots, m \,|\, s_j \in G_\delta(\overline{x})\}\,. \tag{25}$$

Proposition 1:

Let s be a generic point belonging to the gravity region $G_\delta(\overline{x})$, with nearest centroid i_1. Let x be the current position of the centroids. Let $\Delta\,x = \max_i \|x_i - \overline{x}_i\|$ be the maximum displacement of the centroids.

If $\Delta\,x < \delta$ then s will continue to be nearest to centroid x_{i_1} than to any other one,so

$$\min_{i \neq i_1} \|\, s - x_i \,\| \;-\; \|\, s - x_{i_1} \,\| \;\geq\; 0. \tag{26}$$

Proof

$$\min_{i \neq i_1} \|\, s - x_i \,\| - \|\, s - x_{i_1} \,\| \;=\; \min_{i \neq i_1} \|\, s - \overline{x}_i + \overline{x}_i - x_i \,\| - \|\, s - \overline{x}_{i_1} + \overline{x}_{i_1} - x_{i_1} \,\| \;\geq \tag{27}$$

$$\min_{i \neq i_1} \|\, s - \overline{x}_i\| \;-\; \|\overline{x}_i - x_i \,\| \;-\; \|\, s - \overline{x}_{i_1} \,\| \;-\; \|\overline{x}_{i_1} - x_{i_1} \,\| \;\geq \tag{28}$$

$$2\,\delta \;-\; 2\,\Delta\,x \;\geq\; 0 \qquad \blacksquare \tag{29}$$

Since $\delta \geq \Delta\,x$, Proposition 1 makes it possible to calculate exactly expression (16) in a very fast way. First, let us define the subsets of gravity observations associated with each referential centroid:

$$J_i(\overline{x}) \;=\; \left\{ j \in J_G \,|\, \min_{l=1,\ldots,q} \|\, s_j - \overline{x}_l \,\| \;=\; \|\, s_j - \overline{x}_i \,\| \right\} \tag{30}$$

Let us consider the second sum in expression (16).

$$minimize \;\; f_G(x) \;=\; \sum_{j \in J_G} z_j(x) \;=\; \sum_{i=1}^{q} \sum_{j \in J_i} \|\, s_j - x_i \,\|_1 \;= \tag{31}$$

$$\sum_{i=1}^{q} \sum_{j \in J_i} \sum_{l=1}^{n} |\, s_j^l - x_i^l \,| = \tag{32}$$

$$\sum_{i=1}^{q} \sum_{l=1}^{n} \sum_{j \in J_i} |\, s_j^l - x_i^l \,| \tag{33}$$

Let us now perform the partition of each set J_i into 3 subsets for each component in the following form:

$$J_{il}^{+}(\overline{x}) \;=\; \left\{ j \in J_i(\overline{x}) \,|\, s_j^l - x_i^l \;\geq\; \delta \right\} \tag{34}$$

$$J_{il}^{-}(\overline{x}) \;=\; \left\{ j \in J_i(\overline{x}) \,|\, s_j^l - x_i^l \;\leq\; -\,\delta \right\} \tag{35}$$

$$J_{il}^{0}(\overline{x}) \;=\; \left\{ j \in J_i(\overline{x}) \,|\, -\;\delta \;<\; s_j^l - x_i^l \;<\; \delta \right\} \tag{36}$$

By using the defined subsets, it is obtained:

$$minimize\;\; f_G(x) = \sum_{i=1}^{q} \sum_{l=1}^{n} \left[\sum_{j \in J_{il}^{+}} |\, s_j^l - x_i^l \,| + \sum_{j \in J_{il}^{-}} |\, s_j^l - x_i^l \,| + \sum_{j \in J_{il}^{0}} |\, s_j^l - x_i^l \,| \right] =$$

$$\sum_{i=1}^{q} \sum_{l=1}^{n} \left[\sum_{j \in J_{il}^{+}} |\, s_j^l - \overline{x}_1^l + \overline{x}_i^l - x_i^l \,| + \sum_{j \in J_{il}^{-}} |\, s_j^l - \overline{x}_i^l + \overline{x}_i^l - x_i^l \,| + \sum_{j \in J_{il}^{0}} |\, s_j^l - x_i^l \,| \right]$$

Let us define the component displacement of centroid $\Delta\, x_i^l \;=\; x_i \;-\; \overline{x}_i^l$. Since $|\,\Delta\, x_i^l\,| \;<\delta,$ from the above definitions of the subsets, it follows that:

$$|\, s_j^l - x_i^l \,| \;=\; |\, s_j^l - \overline{x}_i^l \,| \;-\; \Delta\, x_i^l \quad \text{for} \;\; j \in J_{il}^{+} \tag{37}$$

$$|\, s_j^l - x_i^l \,| \;=\; |\, s_j^l - \overline{x}_i^l \,| \;+\; \Delta\, x_i^l \quad \text{for} \;\; j \in J_{il}^{-}$$

So, it follows:

$$minimize\ f_G(x) =$$

$$\sum_{i=1}^{q}\sum_{l=1}^{n}\left[\sum_{j\in J_{il}^{+}}(\,|\,s_j^l-\overline{x}_i^l\,|\,-\,\Delta\,x_i^l\,)\,+\,\sum_{j\in J_{il}^{-}}(\,|\,s_j^l-\overline{x}_i^l\,|\,+\,\Delta\,x_i^l\,)\,+\,\sum_{j\in J_{il}^{0}}|\,s_j^l-x_i^l\,|\right] =$$

$$\sum_{i=1}^{q}\sum_{l=1}^{n}\left[\sum_{j\in J_{il}^{+}}|\,s_j^l-\overline{x}_i^l|\,-\,|\,J_{il}^{+}\,|\,\Delta\,x_i^l\,+\,\sum_{j\in J_{il}^{-}}|\,s_j^l-\overline{x}_i^l|\,+\,|\,J_{il}^{-}\,|\,\Delta\,x_i^l\,+\,\sum_{j\in J_{il}^{0}}|\,s_j^l-x_i^l\,|\right] \tag{38}$$

where $|\,J_{il}^{+}\,|$ and $|\,J_{il}^{-}\,|$ are the cardinalities of two first subsets.

When the position of centroids $x_i\,, i=1,\ldots,q$ moves within the iterative process, the value of the first two sums of (38) assumes a constant value, since the values s_j^l and $\overline{x}_i^l$ are fixed. So, to evaluate $f_G(x)$ it is only necessary to calculate the displacements $\Delta\,x_i^l\,,\ i=1,\ldots,q\,,\ l=1,\ldots,n\,,$ and evaluate the last sum, that normally has only a few number of terms because δ assumes in general a relatively small value.

The function $f_G(x)$ above specified is nondifferentiable due the last sum, so in order to use gradient information, it is necessary to use a smooth approximation:

$$minimize\ f_G(x) =$$

$$\sum_{i=1}^{q}\sum_{l=1}^{n}\left[\sum_{j\in J_{il}^{+}}|\,s_j^l-\overline{x}_i^l\,|\,-\,|\,J_{il}^{+}\,|\,\Delta\,x_i^l\,+\,\sum_{j\in J_{il}^{-}}|\,s_j^l-\overline{x}_i^l\,|\,+\,|\,J_{il}^{-}\,|\,\Delta\,x_i^l\,+\,\sum_{j\in J_{il}^{0}}\theta_1(s_j^l,x_i^l,\gamma)\right] \tag{39}$$

where $\theta_1(s_j^l,x_i^l,\gamma)\,=\,((\,s_j^l-x_i^l\,)^2+\gamma^2)^{1/2}\,.$

So, the gradient of the smoothed second part of objective function is easily calculated by:

$$\nabla f_G(x) \;=\; \sum_{i=1}^{q} \sum_{l=1}^{n} \left[\; -\mid J_{il}^{+} \mid + \mid J_{il}^{-} \mid + \sum_{j \in J_{il}^{0}} -(s_j^l - x_i^l) \,/\, \theta_1(s_j^l, x_i^l, \gamma) \right] \; e_{il} \,, \tag{40}$$

where e_{il} stands for a unitary vector with the component l of centroid i equal to 1.

Therefore, if $\delta \geq \Delta\, x$ was observed within the iterative process, the calculation of the expression $\sum_{j \in J_G} z_j(x)$ and its gradient can be exactly performed by very fast procedures, equations (39) and (40).

By using the above results, it is possible to construct a specific method, the Accelerated Hyperbolic Smoothing Method Applied to the Minimum of Sum of L1 Distances Clustering Problem, which has conceptual properties to offer a faster computational performance for solving this specific clustering problem given by formulation (17), since the calculation of the second sum is very simple.

A fundamental question is the proper choice of the boundary parameter δ. Moreover, there are two main options for updating the boundary parameter δ, inside the internal minimization procedure or after it. For simplicity sake, the HSC method connected with the partition scheme presented below adopts the second option, which offers a better computational performance, in spite of an eventual violation of the $\delta \geq \Delta\, x$ condition, which is corrected in the next partition update.

The Simplified AHSC-L1 Algorithm

Initialization Step:

Choose initial start point: x^0;

Choose parameter values: γ^1 , τ^1 , ε^1;

Choose reduction factors: $0 < \rho_1 < 1$, $0 < \rho_2 < 1$, $0 < \rho_3 < 1$;

Specify the boundary band width: δ^1;

Let $k = 1$.

Main Step: Repeat until an arbitrary stopping rule is attained

For determining the $Z_\delta(\overline{x})$ and $G_\delta(\overline{x})$ partitions, given by (22) and (23), use $\overline{x} = x^{k-1}$ and $\delta = \delta^k$.

Determine the subsets J^+_{il} , J^-_{il} and J^+_{il} and calculate the cardinalities of two first sets: $|J^+_{il}|$ and $|J^-_{il}|$.

Solve problem (17) starting at the initial point x^{k-1} and let x^k be the solution obtained:

For solving the equations (19), associated to the first part given by (19), take the smoothing parameters: $\gamma = \gamma^k$, $\tau = \tau^k$ and $\varepsilon = \varepsilon^k$;

For solving the second part, given by (39), use the above determined subsets and their cardinalities.

Updating procedure:

Let $\gamma^{k+1} = \rho_1\, \gamma^k$, $\tau^{k+1} = \rho_2\, \tau^k$, $\varepsilon^{k+1} = \rho_3\, \varepsilon^k$

Redefine the boundary value: δ^{k+1}

Let $k := k+1$. ∎

Comments:

The above algorithm does not include any procedure for considering the occurrence of empty gravitational regions. This possibility can be overcome by simply moving the centroids.

The efficiency of the AHSC-L1 algorithm depends strongly on the parameter δ. A choice of a small value for it will imply an improper definition of the set $G_\delta(\overline{x})$, and frequent violation of the basic condition $\Delta\, x < \delta$, for the validity of Proposition 1. Otherwise, a choice of a large value will imply a decrease in the number of gravitational observation points and, therefore, the computational advantages given by formulation (39) will be reduced.

As a general strategy, within first iterations, larger δ values must be used, because the centroid displacements are more expressive. The δ values must be gradually decreased in the same proportion of the decrease of these displacements.

5 Conclusions

In this paper, a new method for the solution of the minimum sum of L1 distances clustering problem is proposed. It is a natural development of the original HSC method and of the descending AHSC-L2 method, linked to the the minimum sum of distances clustering (MSSC) formulation, presented respectively by Xavier (2010) an by Xavier and Xavier (2010). We hope for a good computational performance in a similar way of the two precedent methods. Such expectation is based on the complete differentiability of the approach and on the partition of the set of observations into two non overlapping parts, which offers a drastic simplification of computational tasks.

Acknowledgments

The author would like to thank Cláudio Joaquim Martagão Gesteira of Federal University of Rio de Janeiro for the helpful review of the work and constructive comments.

References

ANDERBERG, M. R. (1973) "Cluster Analysis for Applications", Academic Press Inc., New York.

BAGIROV, A. M. (2008) "Modified Global k-means Algorithm for Minimum Sum-of-Squares Clustering Problems", Pattern Recognition, Vol 41 Issue 10 pp. 3192-3199.

DUBES, R. C. and JAIN, A. K. (1976) "Cluster Techniques: The User's Dilemma", Pattern Recognition No. 8 pp. 247-260.

HANSEN, P. and JAUMARD, B. (1997) "Cluster Analysis and Mathematical Programming", Mathematical Programming No. 79 pp. 191-215.

HARTINGAN, J. A. (1975) "Clustering Algorithms", John Wiley and Sons, Inc., New York, NY.

JAIN, A. K. and DUBES, R. C. (1988) "Algorithms for Clustering Data", Prentice-Hall Inc., Upper Saddle River, NJ.

LIKAS, A., VLASSIS, M., and VERBEEK, J. (2003) "The Global k-means Clustering Algorithm", Pattern Recognition, Vol 36 pp. 451-461.

REINELT, G. (1991) "TSP-LIB A Traveling Salesman Library", ORSA J. Comput. pp. 376-384

SPÄTH, H. (1980) "Cluster Analysis Algorithms for Data Reduction and Classification", Ellis Horwood, Upper Saddle River, NJ.

WU, X. et alli (2008) "Top 10 algorithms in data mining", Knowledge and Information Systems, Springer, 14, pp. 1-37.

XAVIER, A.E. (1982) "Penalização Hiperbólica: Um Novo Método para Resolução de Problemas de Otimização", M.Sc. Thesis - COPPE - UFRJ, Rio de Janeiro.

XAVIER, A.E. (2010) "The Hyperbolic Smothing Clustering Method", Pattern Recognition, Vol 43, pp. 731-737.

XAVIER, A.E. and XAVIER, V.L. (2010) " Solving the Minimum Sum-of-Squares Clustering Problem by Hyperbolic Smoothing and Partition into Boundary and Gravitational Regions ", Technical Report PESC/UFRJ , Submitted to Pattern Recognition.