An Incremental DC Algorithm for the Minimum Sum-of-Squares Clustering

A. M. Bagirov

Here, an algorithm is presented for solving the minimum sum-of-squares clustering problems using their difference of convex representations. The proposed algorithm is based on an incremental approach and applies the well known DC algorithm at each iteration. The proposed algorithm is tested and compared with other clustering algorithms using large real world data sets.

Keywords: Clustering, Nonsmooth optimization, Nonconvex optimization, Incremental algorithms.

Manuscript was received on 27/04/2015 and accepted for publication on 16/06/2015.†

1. Introduction

Clustering is an unsupervised technique dealing with problems of organizing a collection of patterns into clusters based on similarity. Most clustering algorithms are based on hierarchical and partitional approaches. Algorithms based on an hierarchical approach generate a dendrogram representing the nested grouping of patterns and similarity levels at which groupings change [19]. Partitional clustering algorithms find the partition that optimizes a clustering criterion [19]. In this paper, we develop a partitional clustering algorithm for solving the minimum sum-of-squares clustering (MSSC) problems.

To date various heuristics such as the $k$-means algorithm and its variations have been developed to solve the MSSC problem (see, for example, [20, 21] and references therein). The global $k$-means algorithm and its various modifications are among the most efficient algorithms for solving the MSSC problem [5, 10, 12, 22, 24, 25, 27].

The MSSC can be formulated as an optimization problem [8, 11, 30] and different deterministic optimization techniques and metaheuristics have been applied to solve it. These techniques include branch and bound [15] and interior point methods [16], nonsmooth optimization algorithms [6, 9, 11], algorithms based on the hyperbolic smoothing technique [7, 32, 33], the variable neighborhood search [17], simulated annealing [28], tabu search [1] and genetic algorithms [26]. Most of these algorithms are not efficient for solving the MSSC problem with very large data sets.

The objective function of MSSC problems can be represented by a difference of convex (DC) functions. There are several papers in which this representation is used to design algorithms. In [14], the truncated codifferential method is applied to solve the MSSC problem using its DC representation. The branch and bound method was modified for such problems in [31]. In [2], an algorithm based on

† Invited Paper

1 Faculty of Science and Technology, Federation University Australia, Victoria, Australia.

Email: a.bagirov@federation.edu.au
DC Algorithms (DCAs) was introduced. In [3], a DCA and a Gaussian kernel were applied to design a clustering algorithm.

Here, we develop a new algorithm for solving the MSSC problem using its DC representation. This algorithm is based on an incremental approach and applies the DCA to solve optimization problems at each iteration of the incremental algorithm. Computational results on some real world data sets are reported and the algorithm is compared with two other clustering algorithms. It is demonstrated that the proposed algorithm is especially efficient for solving the MSSC problems with very large data sets.

The rest of the paper is organized as follows. Section 2 provides some preliminaries on DC functions and nonsmooth analysis. DC representations of cluster functions are given in Section 3. Section 4 presents an algorithm for solving clustering problems. An incremental algorithm is discussed in Section 5. Numerical results are reported in Section 6 and Section 7 contains the concluding remarks.

2. Preliminaries

In this section we give some results on nonsmooth analysis and DC functions used throughout the paper. In what follows, we denote by \( \mathbb{R}^n \) the \( n \)-dimensional Euclidean space with the inner product \( \langle u, v \rangle = \sum_{i=1}^{n} u_i v_i \) and the associated norm \( |u| = \langle u, u \rangle^{1/2}, u, v \in \mathbb{R}^n \). The set \( B(x, \varepsilon) = \{ y \in \mathbb{R}^n : |y - x| < \varepsilon \} \) is the open ball centered at \( x \) with the radius \( \varepsilon > 0 \).

Let \( f: \mathbb{R}^n \to \mathbb{R} \) be a convex function. Its subdifferential at \( x \in \mathbb{R}^n \) is defined as

\[ \partial_c f(x) = \{ \xi \in \mathbb{R}^n : f(y) - f(x) \geq \langle \xi, y - x \rangle \ \forall y \in \mathbb{R}^n \}. \]

A function \( f: \mathbb{R}^n \to \mathbb{R} \) is called locally Lipschitz on \( \mathbb{R}^n \) if for any bounded subset \( X \subset \mathbb{R}^n \) there exists \( L > 0 \) such that

\[ |f(x) - f(y)| \leq L|x - y| \ \forall x, y \in X. \]

The generalized derivative of a locally Lipschitz function \( f \) at a point \( x \) with respect to a direction \( u \in \mathbb{R}^n \) is defined as [13]

\[ f^0(x, u) = \limsup_{\alpha \downarrow 0, \ y \to x} \frac{f(y + \alpha u) - f(y)}{\alpha}. \]

The subdifferential \( \partial f(x) \) of the function \( f \) at \( x \) is

\[ \partial f(x) = \{ \xi \in \mathbb{R}^n : f^0(x, u) \geq \langle \xi, u \rangle \ \forall u \in \mathbb{R}^n \}. \]

Each vector \( \xi \in \partial f(x) \) is called a subgradient. For a convex function \( f: \mathbb{R}^n \to \mathbb{R} \), one has \( \partial f(x) = \partial_c f(x), x \in \mathbb{R}^n \). From now on, we will use the notation \( \partial f \) for subdifferential of convex function.

\( f: \mathbb{R}^n \to \mathbb{R} \) is called a DC function if there exist convex functions \( g, h: \mathbb{R}^n \to \mathbb{R} \) such that [18]

\[ f(x) = g(x) - h(x), \quad x \in \mathbb{R}^n. \]
An Incremental DC Algorithm for the Minimum Sum-of-Squares Clustering

Here, $g - h$ is called a DC decomposition of $f$ while $g$ and $h$ are DC components of $f$. Note that a DC function has infinitely many DC decompositions.

An unconstrained DC program is an optimization problem of the form

$$\text{minimize } f(x) = g(x) - h(x) \quad \text{subject to } x \in \mathbb{R}^n. $$  (1)

One can define different types of stationary points for problem (1). Such stationary points include critical, Clarke stationary and inf-stationary points. A point $x^*$ is called inf-stationary for problem (1) if

$$\partial h(x^*) \subseteq \partial g(x^*).$$  (2)

A point $x^*$ is called Clarke stationary for problem (1) if

$$0 \in \partial f(x^*).$$  (3)

Finally, a point $x^*$ is called a critical point of problem (1) if

$$\partial h(x^*) \cap \partial g(x^*) \neq \emptyset.$$  (4)

In general, any inf-stationary point is also Clarke stationary and a critical point. Furthermore, any Clarke stationary point is a critical point.

3. DC Programming Approach to Clustering Problems

Let $A$ be a finite set of points in $\mathbb{R}^n$, that is, $A = \{a^1, \ldots, a^m\}$, $a^i \in \mathbb{R}^n$, $i = 1, \ldots, m$. The hard clustering problem is the distribution of the points of the set $A$ into a given number $k$ of disjoint subsets $A^j$, $j = 1, \ldots, k$, such that

1. $A^j \neq \emptyset$ and $A^j \cap A^l = \emptyset$, $j, l = 1, \ldots, k, j \neq l$.

2. $A = \bigcup_{j=1}^k A^j$.

The sets $A^j$, $j = 1, \ldots, k$, are called clusters and each cluster $A^j$ can be identified by its center $x^j \in \mathbb{R}^n$, $j = 1, \ldots, k$. The problem of finding these centers is called the $k$-clustering (or $k$-partition) problem. In order to formulate the clustering problem one needs to define the similarity (or dissimilarity) measure. Here, the similarity measure is defined using the $L_2$ norm as follows:

$$d_2(x, a) = \sum_{i=1}^n (x_i - a_i)^2.$$

The nonsmooth optimization formulation of the MSSC problem is [8, 11]:

$$\text{minimize } f_k(x) \quad \text{subject to } x = (x^1, \ldots, x^k) \in \mathbb{R}^{nk},$$  (5)

where
The objective function $f_k$ in problem (5) can be represented as a DC function,

$$f_k(x) = f_{k1}(x) - f_{k2}(x), \ x = (x^1, \ldots, x^k) \in \mathbb{R}^{nk}. \tag{7}$$

where

$$f_{k1}(x) = \frac{1}{m} \sum_{a \in A} \sum_{j=1}^{k} d_2(x^j, a), \quad f_{k2}(x) = \frac{1}{m} \sum_{a \in A} \max_{j=1, \ldots, k} \sum_{s=1, s \neq j}^{k} d_2(x^s, a).$$

Since the function $d_2$ is convex in $x$, the function $f_{k1}$, as a sum of convex functions, is also convex. The function $f_{k2}$ is a sum of maxima of sum of convex functions. Since the sum of convex functions is convex, the functions under maximum are also convex. Furthermore, since the maximum of a finite number of convex functions is convex, the function $f_{k2}$ being a sum of convex functions is also convex.

Next, we derive expressions for subdifferentials of the functions $f_{k1}$ and $f_{k2}$. The function $f_{k1}$ is differentiable and its gradient is

$$\nabla f_{k1}(x) = 2(x - \bar{A}). \tag{8}$$

Here, $\bar{A} = (\bar{A}_1, \ldots, \bar{A}_k), \ ar{A}_1 = \cdots = \bar{A}_k = (\bar{a}_1, \ldots, \bar{a}_n)$ and

$$\bar{a}_t = \frac{1}{m} \sum_{a \in A} a_t.$$

This means that the subdifferential $\partial f_{k1}(x)$ is a singleton for any $x \in \mathbb{R}^n$.

In general, the function $f_{k2}$ is nonsmooth. To compute its subdifferential, consider the following function and a set as follows [25]:

$$\varphi_a(x) = \max_{j=1, \ldots, k} \sum_{s=1, s \neq j}^{k} d_2(x^s, a), \tag{9}$$

and

$$R_a(x) = \left\{ j \in \{1, \ldots, k\} : \sum_{s=1, s \neq j}^{k} d_2(x^s, a) = \varphi_a(x) \right\}. \tag{10}$$

The subdifferential $\partial \varphi_a(x)$ of the function $\varphi_a$ at $x$ is

$$\partial \varphi_a(x) = \text{conv}\{V \in \mathbb{R}^{nk} : V = 2(\bar{x}^j - \bar{A}_t^j), j \in R_a(x)\}, \tag{11}$$

where

$$\bar{x}^j = (x^1, \ldots, x^{j-1}, 0_n, x^{j+1}, \ldots, x^k), \quad \bar{A}_t^j = (\bar{A}_1^j, \ldots, \bar{A}_t^j) \in \mathbb{R}^{nk},$$

and
\[
\bar{A}^j_t = a^j, \quad t = 1, ..., k, t \neq j, \quad \bar{A}_{ij} = 0, \quad i = 1, ..., n.
\]

Then, the subdifferential \( \partial f_k^2(x) \) can be expressed by

\[
\partial f_k^2(x) = \frac{1}{m} \sum_{a \in A} \partial \varphi_a(x).
\] (12)

Problem (5) is a global optimization problem. It has many local minimizers and only global or near global solutions provide the best clustering structure of a data set. To find such solutions using a local search method, it is crucial to apply a special procedure to generate good starting cluster centers. We apply an algorithm introduced in [25]. This algorithm involves the solution of the so-called auxiliary clustering problem. Next, we briefly describe this problem.

Assume that the solution \( x^1, ..., x^{k-1}, k \geq 2 \), to the \((k-1)\)-clustering problem is known. Denote by \( r_{k-1}^a \), the distance between the data point \( a \in A \) and the closest cluster center among the \( k-1 \) centers \( x^1, ..., x^{k-1} \):

\[
r_{k-1}^a = \min\{d_2(x^1, a), ..., d_2(x^{k-1}, a)\}. \tag{13}
\]

The \( k \)-th auxiliary cluster function is defined as [5, 25]

\[
\tilde{f}_k(y) = \frac{1}{m} \sum_{a \in A} \min\{r_{k-1}^a, d_2(y, a)\}, \quad y \in \mathbb{R}^n.
\] (14)

This function is nonsmooth, locally Lipschitz, directionally differentiable and as a sum of minima of convex functions it is, in general, nonconvex. It is obvious that

\[
\tilde{f}_k(y) = f_k(x^1, ..., x^{k-1}, y), \quad \forall y \in \mathbb{R}^n.
\]

A problem,

minimize \( \tilde{f}_k(y) \) subject to \( y \in \mathbb{R}^n \), \tag{15}

is called the \( k \)-th auxiliary clustering problem [5, 25]. The DC representation of the function \( \tilde{f}_k \) is as follows:

\[
\tilde{f}_k(y) = \tilde{f}_{k1}(y) - \tilde{f}_{k2}(y), \tag{16}
\]

where

\[
\tilde{f}_{k1}(y) = \frac{1}{m} \sum_{a \in A} (r_{k-1}^a + d_2(y, a)), \quad \tilde{f}_{k2}(y) = \frac{1}{m} \sum_{a \in A} \max\{r_{k-1}^a, d_2(y, a)\}.
\]

To express subdifferentials of the functions \( \tilde{f}_{k1} \) and \( \tilde{f}_{k2} \) at a given point \( y \in \mathbb{R}^n \), define the following sets:

\[
\bar{A}_1(y) = \{a \in A: r_{k-1}^a > d_2(y, a)\}, \quad \bar{A}_2(y) = \{a \in A: r_{k-1}^a < d_2(y, a)\},
\]

\[
\bar{A}_3(y) = \{a \in A: r_{k-1}^a = d_2(y, a)\}.
\]
Then, the function \( \tilde{f}_{k2} \) at \( y \) can be rewritten as

\[
\tilde{f}_{k2}(y) = \frac{1}{m} \left( \sum_{a \in A_1(y)} r_{k-1}^a + \sum_{a \in A_2(y)} d_2(y, a) + \sum_{a \in A_3(y)} \max \{ r_{k-1}^a, d_2(y, a) \} \right).
\]

The function \( \tilde{f}_{k1} \) is continuously differentiable on \( \mathbb{R}^n \) and its gradient at \( y \in \mathbb{R}^n \) is

\[
\nabla \tilde{f}_{k1}(y) = \frac{2}{m} \sum_{a \in A} (y - a).
\] (17)

The function \( \tilde{f}_{k2} \), in general, is nonsmooth and its subdifferential at \( y \in \mathbb{R}^n \) is

\[
\partial \tilde{f}_{k2}(y) = \frac{2}{m} \left( \sum_{a \in A_2(y)} (y - a) + \sum_{a \in A_3(y)} \text{conv}\{0, (y - a)\} \right).
\] (18)

### 4. An Algorithm for Solving Optimization Problems

In this section we describe the DCA for solving both the clustering problem (5) and the auxiliary clustering problem (15). It is easy to observe that both problems can be formulated as the following unconstrained DC programming problem:

\[
\text{minimize } f(x) \text{ subject to } x \in \mathbb{R}^n,
\] (19)

where \( f(x) = f_1(x) - f_2(x) \), the function \( f_1 \) is continuously differentiable convex and the function \( f_2 \) is, in general, a nonsmooth convex function.

DCA is an algorithm for solving DC programming problems. A detailed study of this algorithm can be found in [4, 29]. The generic DCA scheme for solving the problem (19) is shown in Algorithm 1 below.

**Algorithm 1.** DCA scheme for problem (19).

**Step 1:** Select any starting point \( x^1 \in \mathbb{R}^n \) and set \( j := 1 \).

**Step 2:** Compute \( \xi^i \in \partial f_2(x^i) \).

**Step 3:** If \( \xi^i = \nabla f_1(x^i) \) then stop.

**Step 4:** Find the solution \( x^{i+1} \) to the following convex optimization problem:

\[
\text{minimize } f_1(x) - \langle \xi^i, x - x^i \rangle \text{ subject to } x \in \mathbb{R}^n.
\] (20)

**Step 5:** Set \( j := j + 1 \) and go to Step 2.

**Proposition 4.1.** All accumulation points of the sequence \( \{x^k\} \) generated by Algorithm 1 are Clarke stationary points of the problem (19).
Proof. It is well known that accumulation points of the sequence \( \{x^k\} \) are critical points of the problem (19). Since the function \( f_j \) in (19) is continuously differentiable, the sets of critical points and Clarke stationary points of the problem (19) coincide.

In order to apply DCA to solve problem (15), the subgradient \( \xi^j \) in Step 2 is computed as (see (18))
\[
\xi^j = \frac{2}{m} \sum_{a \in A_2(x^j)} (x^j - a), \quad x^j \in \mathbb{R}^n.
\]
Then, the solution \( x^{j+1} \) to the problem (20) in Step 4 can be expressed as follows:
\[
x^{j+1} = \frac{1}{m} \left( |\bar{A}_2(x^j)| x^j + \sum_{a \in \bar{A}_1(x^j) \cup \bar{A}_2(x^j)} a \right).
\]
Applying (17), the stopping criterion in Step 3 can be given by
\[
\sum_{a \in \bar{A}_1(x^j) \cup \bar{A}_2(x^j)} (x^j - a) = 0.
\]

Now, we describe an application of DCA to solve the clustering problem (5). Let \( x^j = (x_j^1, \ldots, x_j^k) \in \mathbb{R}^{nk} \) be a vector of cluster centers at the iteration \( j \) and \( A^1, \ldots, A^k \) be the cluster partition of the data set \( A \) given by these centers.

In order to compute the subgradient \( \xi^j \) in Step 2, for each \( a \in A \) we compute the set \( R_a(x^j) \) given by (10), take any \( p \in R_a(x^j) \) and compute the subgradient \( v^j_a \in \partial \varphi_a(x^j) \) using (11). Then, we apply (12) to compute the subgradient \( \xi^j \). Thus, we get the following formula for the subgradient \( \xi^j \):
\[
\xi^j = \frac{2}{m} \left( \sum_{a \in A \setminus A^1} (x_j^1 - a), \ldots, \sum_{a \in A \setminus A^k} (x_j^k - a) \right)
= \frac{2}{m} \left( (m - |A^1|)x_j^1 - (m\bar{a} - |A^1|\bar{a}_1), \ldots, (m - |A^k|)x_j^k - (m\bar{a} - |A^k|\bar{a}_k) \right),
\]
where \( \bar{a}_l \) is the center of the cluster \( A^l, l = 1, \ldots, k \), and \( \bar{a} \) is the center of the whole set \( A \).

The solution \( x^{j+1} = (x_{j+1}^1, \ldots, x_{j+1}^k) \) to the problem (20) in Step 4 is given by
\[
x_{j+1}^t = \left( 1 - \frac{|A^t|}{m} \right) x_j^t + \frac{|A^t|}{m} \bar{a}_t, \quad t = 1, \ldots, k.
\]
Finally, the stopping criterion in Step 3 can be given by
\[
x_j^t = \left( 1 - \frac{|A^t|}{m} \right) x_j^t + \frac{|A^t|}{m} \bar{a}_t, \quad t = 1, \ldots, k.
\]
These results demonstrate we do not need to apply any optimization algorithm to solve problem (20) for both clustering problem (5) and auxiliary clustering problem (15). In both cases, solutions can be expressed explicitly.

5. Incremental Algorithm

We propose an incremental algorithm to solve problems (5) and (15). This algorithm starts with computing the center of the whole data set. Then, it solves the $k$-clustering problem ($k > 1$) by gradually adding one cluster center at each iteration. At the $l$th, $1 < l \leq k$, iteration a special procedure is applied to find the starting cluster centers using the solution of the $(l - 1)$-clustering problem. One such procedure was introduced in [25] (see also [7]) which we will use in the incremental algorithm given below. At each iteration of the incremental algorithm, DCA is applied to solve the optimization problems (5) and (15). The proposed incremental algorithm is called the Incremental DCA (IncDCA). It is easy to see that this algorithm, in addition to the $k$-partition problem, also solves all intermediate $l$-partition problems, for $l = 1, ..., k - 1$.

Algorithm 2. An incremental clustering algorithm (IncDCA).

**Step 1:** (Initialization) Compute the center $x^1 \in \mathbb{R}^n$ of the set $A$. Set $l := 1$.

**Step 2:** (Stopping criterion) Set $l := l + 1$. If $l > k$ then stop the $k$-partition problem has been solved.

**Step 3:** (Computation of a set of starting points for the auxiliary clustering problem) Apply the procedure from [25] to find the set $S_{1l} \subset \mathbb{R}^n$ of the starting points for solving the auxiliary clustering problem (15) for $k = l$.

**Step 4:** (Computation of a set of starting points for the $l$th cluster center) Apply Algorithm 1 to solve problem (15) starting from each point $y \in S_{1l}$ and generate a set $S_{2l} \subset \mathbb{R}^n$ of starting points for the $l$th cluster center.

**Step 5:** (Computation of a set of cluster centers) For each $\tilde{y} \in S_{2l}$, apply Algorithm 1 to solve problem (5) starting from the point $(x^1, ..., x^{l-1}, \tilde{y})$ and find a solution $(\hat{y}^1, ..., \hat{y}^l)$. Denote by $S_{3l} \subset \mathbb{R}^{nl}$, the set of all such solutions.

**Step 6:** (Computation of the best solution) Compute

$$f_{l}^{\min} = \min \{ f_{l}(\hat{y}^1, ..., \hat{y}^l) : (\hat{y}^1, ..., \hat{y}^l) \in S_{3l} \}$$

and the collection of cluster centers $(\hat{y}^1, ..., \hat{y}^l)$ such that $f_{l}(\hat{y}^1, ..., \hat{y}^l) = f_{l}^{\min}$.

**Step 5:** (Solution to the $l$-partition problem) Set $x^{l} := \hat{y}^j, j = 1, ..., l$, as a solution to the $l$th partition problem and go to Step 2.

Algorithm 2 contains a special procedure to generate starting cluster centers (Step 3) which is described in detail in [25]. The values of the parameters in this procedure were chosen according to the recommendations given in [7].

6. Numerical Results

To test IncDCA and compare it with other clustering algorithms, numerical experiments with a number of real-world data sets were carried out. Algorithms were implemented in Fortran 95 and compiled using the gfortran compiler. Computational results were obtained on a PC with the CPU Intel(R) Core(TM) i5-3470S 2.90 GHz and RAM 8 GB. Eight data sets were used in the numerical
experiments. Their brief descriptions are given in Table 1. The detailed descriptions can be found in [23]. All data sets contain only numeric features and they do not have missing values. The number of attributes in these data sets ranges from 3 to 128 and the number of data points ranges from tens of thousands (smallest: 13,910) to hundred of thousands (largest: 434,874).

The IncDCA’s performance was compared with two other incremental algorithms: the global $k$-means algorithm (GKM) [24] and the multi-start modified global $k$-means algorithm (MS-MGKM) [25]. Implementations of the GKM and MS-MGKM algorithms were discussed in [24] and [25], respectively.

<table>
<thead>
<tr>
<th>Data sets</th>
<th>Number of instances</th>
<th>Number of attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gas Sensor Array Drift</td>
<td>13910</td>
<td>128</td>
</tr>
<tr>
<td>EEG Eye State</td>
<td>14980</td>
<td>14</td>
</tr>
<tr>
<td>Letter Recognition</td>
<td>20000</td>
<td>16</td>
</tr>
<tr>
<td>KEGG Metabolic Relation Network</td>
<td>53413</td>
<td>20</td>
</tr>
<tr>
<td>Shuttle Landing Control</td>
<td>58000</td>
<td>9</td>
</tr>
<tr>
<td>Localization Data for Person Activity</td>
<td>164860</td>
<td>3</td>
</tr>
<tr>
<td>Skin Segmentation</td>
<td>245057</td>
<td>3</td>
</tr>
<tr>
<td>3D Road Network</td>
<td>434874</td>
<td>3</td>
</tr>
</tbody>
</table>

We computed up to 25 clusters in all the data sets. The CPU times used by the algorithms were limited to 20 hours. Results are presented in tables 2-4. In these tables, we use the following notations:

- $k$ is the number of clusters;
- $f_{\text{best}}$ (multiplied by the number shown after name of the data set) is the best value of the cluster function (6) (multiplied by $m$) obtained by the three algorithms;
- $E$ is the percentage error calculated as follows:

$$E = \frac{\bar{f} - f_{\text{best}}}{f_{\text{best}}} \times 100\%,$$

where $\bar{f}$ is the value of the clustering function obtained by the corresponding algorithm;

- $N_{\text{best}}$ is the least number of distance function evaluations used by the three algorithms;
- $r_N$ is the ratio of the number of distance function evaluations ($N$) used by an algorithm and the corresponding $N_{\text{best}}$: $r_N = N/N_{\text{best}}$;
- $t_{\text{best}}$ is the least CPU time required by three algorithms;
- $r_t$ is the ratio of the CPU time ($t$) required by an algorithm and the corresponding least CPU time: $r_t = t/t_{\text{best}}$;
- The sign “-” in the tables shows that an algorithm failed to compute clusters in the given time limit.

Results for the best cluster function values found by different algorithms are presented in Table 2. These results show that in most cases all the three algorithms achieved high accuracy. The IncDCA failed to find highly accurate solutions over the Localization Data and the GKM and MS-MGKM algorithms over the KEGG Network data set. It is well-known that both the GKM and MS-MGKM algorithms are able to find either global or near global solutions to the clustering problem. Results from Table 2 allow us to claim that the IncDCA algorithm is also able to find either global or near global solutions of the clustering problems.
Results for the number of distance function evaluations by different algorithms are shown in Table 3. One can see that the MS-MGKM algorithm uses the least number of distance function evaluations among the three algorithms in almost all cases. The IncDCA algorithm requires less distance function evaluations than the GKM algorithm. We can also see that the IncDCA and MS-MGKM algorithms use similar number of distance function evaluations over the three largest data sets: Localization Data, Skin Segmentation, and 3D Road Network. In these data sets the GKM algorithm uses significantly more function evaluations than the other two algorithms.

Table 2. Cluster function values obtained by the algorithms

<table>
<thead>
<tr>
<th>Dataset</th>
<th>IncDCA</th>
<th>InDCA</th>
<th>GKM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shuttle Landing Control</td>
<td>1.30</td>
<td>1.25</td>
<td>1.10</td>
</tr>
<tr>
<td>Letter Recognition (×10^{-5})</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Localization Data (×10^{12})</td>
<td>0.34</td>
<td>0.33</td>
<td>0.31</td>
</tr>
<tr>
<td>Skin Segmentation (×10^{15})</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>3D Road Network (×10^{15})</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Gas Sensor Array Drift (×10^{-9})</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>EEG Eye State (×10^{10})</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>KEGG Network (×10^{13})</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Localization Data (×10^{12})</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Skin Segmentation (×10^{15})</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>3D Road Network (×10^{15})</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Gas Sensor Array Drift (×10^{-9})</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>EEG Eye State (×10^{10})</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>KEGG Network (×10^{13})</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>
### Table 3. Number of distance function evaluations

<table>
<thead>
<tr>
<th>k</th>
<th>$N_{\text{best}}$</th>
<th>IncDCA $r_N$</th>
<th>MGKM $r_N$</th>
<th>GKM $r_N$</th>
<th>$\tilde{N}_{\text{best}}$</th>
<th>IncDCA $r_N$</th>
<th>MGKM $r_N$</th>
<th>GKM $r_N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gas Sensor Array Drift ($\times 10^5$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>EEG Eye State ($\times 10^6$)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.457</td>
<td>1.11</td>
<td>1.00</td>
<td>4.24</td>
<td>0.180</td>
<td>1.00</td>
<td>1.08</td>
<td>1248.83</td>
</tr>
<tr>
<td>3</td>
<td>1.302</td>
<td>1.12</td>
<td>1.00</td>
<td>2.98</td>
<td>0.375</td>
<td>1.00</td>
<td>1.08</td>
<td>1198.92</td>
</tr>
<tr>
<td>5</td>
<td>3.185</td>
<td>2.42</td>
<td>1.00</td>
<td>2.44</td>
<td>28.477</td>
<td>1.00</td>
<td>1.00</td>
<td>31.54</td>
</tr>
<tr>
<td>10</td>
<td>8.487</td>
<td>2.60</td>
<td>1.00</td>
<td>2.07</td>
<td>354.547</td>
<td>1.36</td>
<td>1.00</td>
<td>5.74</td>
</tr>
<tr>
<td>15</td>
<td>14.126</td>
<td>3.01</td>
<td>1.00</td>
<td>1.94</td>
<td>795.978</td>
<td>2.28</td>
<td>1.00</td>
<td>4.01</td>
</tr>
<tr>
<td>20</td>
<td>19.867</td>
<td>3.62</td>
<td>1.00</td>
<td>1.87</td>
<td>1274.649</td>
<td>3.09</td>
<td>1.00</td>
<td>3.45</td>
</tr>
<tr>
<td>25</td>
<td>25.505</td>
<td>3.58</td>
<td>1.00</td>
<td>1.85</td>
<td>1793.991</td>
<td>3.25</td>
<td>1.00</td>
<td>3.11</td>
</tr>
<tr>
<td>Letter Recognition ($\times 10^3$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>KEGG Network ($\times 10^4$)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.172</td>
<td>1.01</td>
<td>1.00</td>
<td>2.34</td>
<td>0.157</td>
<td>1.00</td>
<td>1.00</td>
<td>18.17</td>
</tr>
<tr>
<td>3</td>
<td>0.341</td>
<td>1.01</td>
<td>1.00</td>
<td>2.35</td>
<td>0.389</td>
<td>1.11</td>
<td>1.00</td>
<td>14.68</td>
</tr>
<tr>
<td>5</td>
<td>0.673</td>
<td>1.06</td>
<td>1.00</td>
<td>2.39</td>
<td>1.292</td>
<td>1.46</td>
<td>1.00</td>
<td>8.84</td>
</tr>
<tr>
<td>10</td>
<td>1.518</td>
<td>1.18</td>
<td>1.00</td>
<td>2.40</td>
<td>4.63</td>
<td>3.32</td>
<td>1.00</td>
<td>5.55</td>
</tr>
<tr>
<td>15</td>
<td>2.398</td>
<td>1.34</td>
<td>1.00</td>
<td>2.37</td>
<td>9.69</td>
<td>3.31</td>
<td>1.00</td>
<td>4.13</td>
</tr>
<tr>
<td>20</td>
<td>3.319</td>
<td>1.52</td>
<td>1.00</td>
<td>2.34</td>
<td>15.127</td>
<td>2.98</td>
<td>1.00</td>
<td>3.60</td>
</tr>
<tr>
<td>25</td>
<td>4.251</td>
<td>1.69</td>
<td>1.00</td>
<td>2.31</td>
<td>20.873</td>
<td>3.64</td>
<td>1.00</td>
<td>3.30</td>
</tr>
<tr>
<td>Shuttle Landing Control ($\times 10^5$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Localization Data ($\times 10^{10}$)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.122</td>
<td>1.00</td>
<td>1.00</td>
<td>274.91</td>
<td>0.948</td>
<td>1.00</td>
<td>1.00</td>
<td>2.87</td>
</tr>
<tr>
<td>3</td>
<td>0.264</td>
<td>1.00</td>
<td>1.00</td>
<td>254.97</td>
<td>1.997</td>
<td>1.00</td>
<td>1.00</td>
<td>2.72</td>
</tr>
<tr>
<td>5</td>
<td>2.247</td>
<td>38.48</td>
<td>1.00</td>
<td>59.89</td>
<td>3.857</td>
<td>1.00</td>
<td>1.00</td>
<td>2.82</td>
</tr>
<tr>
<td>10</td>
<td>12.003</td>
<td>15.17</td>
<td>1.00</td>
<td>25.23</td>
<td>8.443</td>
<td>1.00</td>
<td>1.00</td>
<td>2.90</td>
</tr>
<tr>
<td>15</td>
<td>29.131</td>
<td>33.64</td>
<td>1.00</td>
<td>16.18</td>
<td>13.162</td>
<td>1.00</td>
<td>1.00</td>
<td>2.90</td>
</tr>
<tr>
<td>20</td>
<td>59.699</td>
<td>29.69</td>
<td>1.00</td>
<td>10.71</td>
<td>18.041</td>
<td>1.01</td>
<td>1.00</td>
<td>2.87</td>
</tr>
<tr>
<td>25</td>
<td>103.658</td>
<td>25.30</td>
<td>1.00</td>
<td>7.80</td>
<td>23.018</td>
<td>1.00</td>
<td>1.00</td>
<td>2.85</td>
</tr>
<tr>
<td>Skin Segmentation ($\times 10^{13}$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3D Road Network ($\times 10^{13}$)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.231</td>
<td>1.00</td>
<td>1.00</td>
<td>2.61</td>
<td>0.594</td>
<td>1.00</td>
<td>1.00</td>
<td>3.19</td>
</tr>
<tr>
<td>3</td>
<td>0.452</td>
<td>1.00</td>
<td>1.00</td>
<td>2.65</td>
<td>1.394</td>
<td>1.00</td>
<td>1.00</td>
<td>2.71</td>
</tr>
<tr>
<td>5</td>
<td>0.869</td>
<td>1.00</td>
<td>1.00</td>
<td>2.76</td>
<td>2.963</td>
<td>1.00</td>
<td>1.00</td>
<td>2.55</td>
</tr>
<tr>
<td>10</td>
<td>1.793</td>
<td>1.01</td>
<td>1.00</td>
<td>3.02</td>
<td>7.038</td>
<td>1.00</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>15</td>
<td>2.618</td>
<td>1.03</td>
<td>1.00</td>
<td>3.21</td>
<td>11.178</td>
<td>1.00</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>20</td>
<td>3.424</td>
<td>1.05</td>
<td>1.00</td>
<td>3.33</td>
<td>15.464</td>
<td>1.00</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>25</td>
<td>4.218</td>
<td>1.11</td>
<td>1.00</td>
<td>3.42</td>
<td>19.771</td>
<td>1.00</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Results for the CPU times required by the algorithms are given in Table 4. These results show that the GKM algorithm is the most time consuming among the three algorithms as the size of a data set increases. The MS-MGKM algorithm requires the least CPU time over five data sets (except the cases, when $k = 2, 3$): Gas Sensor Array Drift, EEG Eye State, Letter Recognition, KEGG Network and Shuttle Control. However, in three largest data sets the IncDCA algorithm requires the least CPU time. Moreover, the GKM and MS-MGKM algorithms computed only 5 clusters in 3D Road Network data set within the given time limit. These results demonstrate that for largest data sets used in this paper, the IncDCA algorithm outperforms the other two algorithms, that is, the IncDCA algorithm is more efficient for solving clustering problems with hundreds of thousands of data points.

Table 4. CPU times used by the algorithms

<table>
<thead>
<tr>
<th>$k$</th>
<th>$t_{best}$</th>
<th>IncDCA $t_r$</th>
<th>MGKM $t_r$</th>
<th>GKM $t_r$</th>
<th>$t_{best}$</th>
<th>IncDCA $t_r$</th>
<th>MGKM $t_r$</th>
<th>GKM $t_r$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Gas Sensor Array Drift</td>
<td>EEG Eye State</td>
<td></td>
<td></td>
<td>Letter Recognition</td>
<td>KEGG Network</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>29.87</td>
<td>1.03</td>
<td>1.00</td>
<td>3.76</td>
<td>0.02</td>
<td>1.00</td>
<td>19.50</td>
<td>630.81</td>
</tr>
<tr>
<td>3</td>
<td>83.46</td>
<td>1.04</td>
<td>1.00</td>
<td>2.70</td>
<td>0.03</td>
<td>1.00</td>
<td>20.13</td>
<td>644.65</td>
</tr>
<tr>
<td>5</td>
<td>199.00</td>
<td>2.36</td>
<td>1.00</td>
<td>2.28</td>
<td>1.28</td>
<td>1.00</td>
<td>1.96</td>
<td>30.97</td>
</tr>
<tr>
<td>10</td>
<td>518.78</td>
<td>2.45</td>
<td>1.00</td>
<td>1.99</td>
<td>22.59</td>
<td>1.00</td>
<td>1.34</td>
<td>4.47</td>
</tr>
<tr>
<td>15</td>
<td>853.53</td>
<td>2.73</td>
<td>1.00</td>
<td>1.85</td>
<td>64.18</td>
<td>1.31</td>
<td>1.00</td>
<td>2.50</td>
</tr>
<tr>
<td>20</td>
<td>1192.04</td>
<td>3.18</td>
<td>1.00</td>
<td>1.79</td>
<td>94.07</td>
<td>1.92</td>
<td>1.00</td>
<td>2.33</td>
</tr>
<tr>
<td>25</td>
<td>1523.23</td>
<td>3.13</td>
<td>1.00</td>
<td>1.76</td>
<td>123.41</td>
<td>2.14</td>
<td>1.00</td>
<td>2.20</td>
</tr>
</tbody>
</table>

|     |            | Skin Segmentation ($\times 10^8$) | 3D Road Network ($\times 10^{13}$) |
| 2   | 562.15     | 1.00         | 3.69       | 4.83      | 1487.22    | 1.00       | 7.09       | 9.73      |
| 3   | 1103.96    | 1.00         | 3.60       | 4.78      | 3452.60    | 1.00       | 5.33       | 7.45      |
| 5   | 2088.68    | 1.00         | 2.82       | 4.21      | 7225.06    | 1.00       | 5.53       | 7.79      |
| 10  | 4207.89    | 1.00         | 2.20       | 3.84      | 16583.94   | 1.00       | -          | -         |
| 15  | 6150.09    | 1.00         | 1.99       | 3.72      | 25867.28   | 1.00       | -          | -         |
| 20  | 8137.20    | 1.00         | 1.81       | 3.55      | 35359.02   | 1.00       | -          | -         |
| 25  | 10309.79   | 1.00         | 1.65       | 3.37      | 44701.26   | 1.00       | -          | -         |
7. Conclusion

An algorithm for solving the minimum sum-of-squares clustering problems was developed. The algorithm exploited DC representation of the clustering problems, was based on the incremental approach and applied the DC algorithm to solve optimization problems. The proposed algorithm was tested using, in particular, real world data sets with hundreds of thousands of data points. Results demonstrated that the algorithm was able to find global or near global solutions of the clustering problems with very large data sets in a reasonable time, outperforming two other state-of-the-art incremental clustering algorithms.

Acknowledgement

This research by Dr. Adil Bagirov was supported under Australian Research Council’s Discovery Projects funding scheme (Project No. DP140103213).

References


