Fast global \(k\)-means clustering based on local geometrical information

Liang Bai \(^{a}\), Jiye Liang \(^{a, *}\), Chao Sui \(^{b}\), Chuangyin Dang \(^{c}\)

\(^{a}\) School of Computer and Information Technology, Shanxi University, Taiyuan, 030006 Shanxi, China
\(^{b}\) Shanxi Academy of Agricultural Sciences, Taiyuan, 030006 Shanxi, China
\(^{c}\) Department of System Engineering and Engineering Management, City University of Hong Kong, Hong Kong

**Abstract**

The fast global \(k\)-means (FGKM) clustering algorithm is one of the most effective approaches for resolving the local convergence of the \(k\)-means clustering algorithm. Numerical experiments show that it can effectively determine a global or near global minimizer of the cost function. However, the FGKM algorithm needs a large amount of computational time or storage space when handling large data sets. To overcome this deficiency, a more efficient FGKM algorithm, namely FGKM+A, is developed in this paper. In the development, we first apply local geometrical information to describe approximately the set of objects represented by a candidate cluster center. On the basis of the approximate description, we then propose an acceleration mechanism for the production of new cluster centers. As a result of the acceleration, the FGKM+A algorithm not only yields the same clustering results as that of the FGKM algorithm but also requires less computational time and fewer distance calculations than the FGKM algorithm and its existing modifications. The efficiency of the FGKM+A algorithm is further confirmed by experimental studies on several UCI data sets.

**1. Introduction**

Clustering is an important problem in statistical multivariate analysis, data mining and machine learning [12]. The goal of clustering is to group a set of objects into clusters so that the objects in the same cluster are highly similar but remarkably dissimilar with objects in other clusters [20]. To tackle this problem, various types of clustering algorithms have been developed in the literature (e.g., [14] and references therein). Among them, the \(k\)-means clustering algorithm [17] is one of the most efficient clustering algorithms for large-scale spherical data sets. It has extensive applications in such domains as financial fraud, medical diagnosis, image processing, information retrieval, and bioinformatics.

The \(k\)-means clustering algorithm uses the alternating minimization method to solve a nonconvex optimization problem in finding cluster solutions [14]. However, the obtained clustering results guarantee local optimization solutions only [20]. To solve this problem, several techniques have been developed based on different global search methods, such as simulated annealing [6,11], genetic algorithms [3,15,19,23], colony optimization [2,27], particle swarm optimization [1,16], stochastic optimization [9], and black hole algorithm [13]. Among these methods, the fast global \(k\)-means clustering (FGKM) algorithm proposed by Likas et al. [22] is a very effective search approach, which uses the incremental learning technique to solve the local minimum problem. The numerical experiment results [4] have shown that the FGKM algorithm can determine a global
or near global minimizer of the $k$-means objective function. Nevertheless, the FGKM algorithm requires calculating the distances between any two data objects in each iteration. For a small data set, we can use $O(n^2)$ memory space to store distances and avoid repeated computations, where $n$ is the number of data objects. However, for a large data set, storing these distances is unfeasible. For instance, if a data set has $n = 10^6$ objects, storing the distances between all objects (assuming double precision storage) requires 8 TB of memory, which is unavailable on a general purpose machine [7].

To make the FGKM algorithm more effective, a modified global $k$-means (MGKM) algorithm was proposed in [4]. The algorithm minimizes an auxiliary objective function to determine new cluster centers. Compared with the FGKM algorithm, the MGKM algorithm can obtain a slightly better result but with a longer computational time [18]. Bagirov et al. [5] suggested a new version of the MGKM algorithm to reduce the computational time of the clustering process and obtain an approximation result. Likas et al. [22] proposed the kd-tree to speed up the generation of new cluster centers in the FGKM algorithm. Unfortunately, the kd-tree-based algorithm is unsuitable for data sets with high dimensions, given that its computational complexity grows exponentially with the data dimensions [18, 10]. Lai and Huand [18] presented a fast search algorithm. Section 5 illustrates the effectiveness of the proposed algorithm on real data sets. Finally, Section 6 concludes the paper with some remarks.

The rest of this paper is organized as follows. Section 2 reviews the $k$-means and FGKM algorithms. Section 3 presents a more efficient FGKM algorithm (i.e., FGKM*+A). Section 4 analyzes the space and computational complexity of the proposed algorithm. Section 5 illustrates the effectiveness of the proposed algorithm on real data sets. Finally, Section 6 concludes the paper with some remarks.

2. The $k$-means and FGKM algorithms

Let $U = \{x_1, x_2, \ldots, x_n\}$ be a set of $n$ objects. Object $x_i = \{x_{i1}, x_{i2}, \ldots, x_{im}\}$ is characterized by a set of $m$ attributes (variables). The $k$-means algorithm [17] searches for a partition of $U$ into $k$ clusters that minimizes the objective function $F$ with unknown variables $W$ and $V$:

$$F(W, V) = \sum_{l=1}^{k} \sum_{i=1}^{n} w_{li} \|x_i - v_l\|^2,$$

subject to

$$w_{li} \in \{0, 1\}, \sum_{i=1}^{k} w_{li} = 1, 0 < \sum_{i=1}^{n} w_{li} < n, \quad 1 \leq l \leq k, \quad 1 \leq i \leq n. \quad (2)$$

where

- $W = [w_{ij}]$ is a $k$-by-$n$ $(0, 1)$ matrix, $w_{ij}$ is a binary variable, and indicates whether object $x_i$ belongs to the $l$th cluster, $w_{ij} = 1$ if $x_i$ belongs to the $l$th cluster and 0 otherwise;
- $V = [v_1, v_2, \ldots, v_k]$ and $v_l = [v_{l1}, v_{l2}, \ldots, v_{lm}]$ is the $l$th cluster center with $m$ attributes;
- $\|x_i - v_l\|^2 = \sum_{j=1}^{m} (x_{ij} - v_{lj})^2$ is Euclidean distance between the object $x_i$ and the $l$th cluster center $v_l$.

The minimization of $F$ in (1) with the constraints in (2) forms a class of constrained nonlinear optimization problems whose solutions are unknown. The usual method toward the optimization of $F$ in (1) is to use partial optimization for $V$ and $W$. In this method, we first fix $V$ and find necessary conditions on $W$ to minimize $F$. Thereafter, we fix $W$ and minimize $F$ with respect to $V$. The above optimization problem can be solved by iteratively solving the following two minimization problems.

- Problem $P_1$: Fix $V = \hat{V}$, solve the reduced problem $F(W, \hat{V})$;
- Problem $P_2$: Fix $W = \hat{W}$, solve the reduced problem $F(\hat{W}, V)$.

Problem $P_1$ is solved by

$$w_{li} = \begin{cases} 1, & \text{if } \|x_i - \hat{v}_l\|^2 \leq \|x_i - \hat{v}_h\|^2, \quad 1 \leq h \leq k, \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

for $1 \leq i \leq n, 1 \leq l \leq k$.

Problem $P_2$ is solved by
\[ v_j = \frac{\sum_{i=1}^{n} w_i x_{ij}}{\sum_{i=1}^{n} w_i}, \]

for \( 1 \leq l \leq k, 1 \leq j \leq m. \)

This process is formalized in the \( k \)-means algorithm [17]:

**The \( k \)-means algorithm**

**Step 1.** Choose an initial point set \( V(1) \in R^{mk} \). Determine \( W(1) \) such that \( F(W, V(1)) \) is minimized. Set \( t = 1 \).

**Step 2.** Determine \( V(t+1) \) such that \( F(W(t), V(t+1)) \) is minimized. If \( F(W(t), V(t+1)) = F(W(t), V(t)) \), then stop; otherwise goto Step 3.

**Step 3.** Determine \( W(t+1) \) such that \( F(W(t+1), V(t+1)) \) is minimized. If \( F(W(t+1), V(t+1)) = F(W(t), V(t+1)) \), then stop; otherwise set \( t = t + 1 \) and goto Step 2.

Since the time complexity of the algorithm is \( O(n knmt) \), it can efficiently cluster large data sets. However, the obtained clustering results guarantee local minimum solutions only. Thus, the performance of the algorithm heavily depends on initial cluster centers.

The global \( k \)-means clustering algorithm introduced by [22] constitutes a deterministic global optimization method that is independent of any initial parameter values and employs the \( k \)-means algorithm as a local search procedure. Instead of randomly selecting initial values for all cluster centers as is the case with most global clustering algorithms, the proposed technique proceeds in an incremental manner to add optimally a new cluster center at each stage.

According to (3), we can obtain \( W \) and minimize \( F(W, V) \) when \( V \) is given. Therefore, the objective function (1) can be rewritten as follows:

\[ F(V) = \min_{W} F(W, V) = \sum_{i=1}^{n} \min_{v_i \in V(i)} \| x_i - v_i \|^2. \]

The global \( k \)-means algorithm is briefly described as follows [22]:

**The global \( k \)-means algorithm**

**Step 1.** Compute \( V^*_1 = \{ v_1 \} \) from the data set \( U \), where \( v_1 = \frac{\sum_{i=1}^{n} x_i}{n} \) and \( n \) is the number of objects in \( U \). Set \( h = 1 \).

**Step 2.** Set \( h = h + 1 \). If \( h > k \), then stop.

**Step 3.** For each object \( x_i \in U \), apply the \( k \)-means algorithm with the initial set of cluster centers \( V_{h-1} \cup \{ x_i \} \) and obtain the resulting set of cluster centers \( V_h(i) = \{ v_1(i), v_2(i), \ldots, v_h(i) \} \).

**Step 4.** Set \( V_h = V_h(r) \) which satisfies

\[ F(V_h(r)) = \min_{i=1}^{n} F(V_h(i)), \]

and goto Step 2.

When handling large data sets, the global \( k \)-means algorithm is inefficient, since it has a time complexity of \( O(n^2 mk^2 t) \). Therefore, several modified algorithms have been proposed to reduce the computational load.

Likas et al. [22] proposed a FGKM algorithm, which is described as follows:

**The FGKM algorithm**

**Step 1.** Compute \( V^*_1 = \{ v_1 \} \) from the data set \( U \), where \( v_1 = \frac{\sum_{i=1}^{n} x_i}{n} \) and \( n \) is the number of objects in \( U \). Set \( h = 1 \).

**Step 2.** Set \( h = h + 1 \). If \( h > k \), then stop.

**Step 3.** For each object \( x_i \in U \), compute

\[ b_h^i = \sum_{j=1}^{n} \max \left( 0, d_{h-1}^{i,j} - \| x_i - x_j \|^2 \right) \]

where \( d_{h-1}^{i,j} = \min_{v_i \in V_{h-1}} \| v_i - x_j \|^2 \).

**Step 4.** Set \( V = V_{h-1} \cup \{ x_i \} \) which satisfies

\[ b_h^i = \max_{i=1}^{n} b_h^i. \]

**Step 5.** Apply the \( k \)-means algorithm with the initial set of cluster centers \( V \), save the resulting set of cluster centers into \( V_h \) and compute \( d_h^i \) for each \( x_j \in U \). Goto Step 2.

Compared with the global \( k \)-means clustering algorithm, the FGKM algorithm does not execute the \( k \)-means algorithm for each data object in Step 3. Instead, the FGKM computes an upper bound \( F(V_h(i)) \leq F(V_{h-1}) - b_h^i \) making the time complexity \( O(n^2 mk + nmk^2 t) \). The numerical experiment results in [4] have shown that the FGKM algorithm can determine a global or near global minimizer of the objective function.

Please cite this article in press as: L. Bai et al., Fast global \( k \)-means clustering based on local geometrical information, Inform. Sci. (2013), http://dx.doi.org/10.1016/j.ins.2013.05.023
3. The FGKM+A algorithm

In the FGKM algorithm, computing \( b^k_i \) for each \( x_i \) in \( U \) is an important step, given that \( b^k_i \) is necessary to determine which object in \( U \) will be the initial center of a new cluster. However, the process is extremely time consuming. Here, \( b^k_i = \sum_{x_j \in P_b(x_i)} (d_{ij}^k - ||x_j - x_i||^2) \), where \( P_b(x_i) = \{j ||x_j - x_i||^2 \leq d_{ij-1}^k, x_j \in U\} \), for \( 1 \leq i \leq n \). The FGKM algorithm needs to identify a set \( P_b(x_i) \) from \( U \) by calculating the distance between each object \( x_i \) in \( U \) and \( x_i \). The entire process needs \( n^2 \) distance calculations. In this section, we propose an acceleration mechanism to reduce the computing cost of the process.

The mechanism will enhance the efficiency of the FGKM algorithm in the following two aspects:

- Instead of directly computing \( ||x_j - x_i||^2 \) for each object \( x_j \in U \), we will compute its estimated value and determine whether \( x_j \) belongs to \( P_b(x_i) \).
- Given that the FGKM algorithm requires only the determination of the object \( x_i \) with the maximum \( b^k_i \), computing the exact \( b^k_i \) value for each \( x_i, 1 \leq i \leq n \) is unnecessary. We will first compute an upper bound \( \bar{b}^k_{\text{max}} \) of \( b^k_i \). If \( \bar{b}^k_{\text{max}} < b^k_i \), it is concluded that \( x_i \) is impossible to be the new initial center; otherwise, we will compute the exact \( b^k_i \).

We will introduce how to take advantage of the local geometrical information of objects to build the acceleration mechanism. We first pre-process the given data set and tends to partition the data set into clusters with relatively uniform sizes [8,26], we suggest applying the algorithm to obtain relatively uniform clusters. Since the k-means algorithm has an approximate linear time complexity for the number of objects and tends to partition the data set into clusters with relatively uniform sizes [8,26], we suggest applying the algorithm with \( k' \) randomly selected initial centers to quickly produce \( k' \) small clusters.

After obtaining the partition \( S \), we can rewrite \( b^k_i = \sum_{x_j \in S_{\text{close}}} \psi_{S_{\text{close}}}(x_i) \) where \( \psi_{S_{\text{close}}}(x_i) = \sum_{x_j \in S_{\text{close}}} \max\{0, d_{ij-1}^k - ||x_j - x_i||^2\} \), and \( P_b(x_i) = \bigcup_{j=1}^{k'} B^j(x_i) \) where \( B^j(x_i) = \{x_j ||x_j - x_i||^2 \leq d_{ij-1}^k, x_j \in S_{\text{close}}\} \) (Fig. 2). When determining whether the objects in \( S_{\text{close}} \) belong to \( B^j(x_i) \), we do not directly compute the distance between \( x_i \) and each \( x_j \) in \( S_{\text{close}} \). Considering

\[
||x_i - c_i||^2 - ||x_j - c_i||^2 \leq ||x_i - x_j||^2 \leq ||x_i - c_i||^2 + ||x_j - c_i||^2
\]

for each \( x_i \) in \( S_{\text{close}} \), we use \( ||x_i - c_i||^2 \) and \( ||x_j - c_i||^2 \) to estimate \( ||x_i - x_j||^2 \) and obtain the following inequality rules to reduce several unnecessary operations when constructing \( P_b(x_i) \):

(a) If \( ||x_i - c_i||^2 - ||x_j - c_i||^2 \geq d_{ij-1}^k \) and \( x_j \in S_{\text{close}} \), the object \( x_j \) does not belong to \( B^j(x_i) \).
(b) If \( ||x_i - c_i||^2 + ||x_j - c_i||^2 \leq d_{ij-1}^k \) and \( x_j \in S_{\text{close}} \), the object \( x_j \) belongs to \( B^j(x_i) \).

![Fig. 1. Pre-processing of a data set.](image-url)
If 

\[ k x_i \] is the new cluster center, \( x_i \) can represent when \( x_i \) is the new cluster center. (b) \( d_{k-1}, r_i, ||x - c||^2, ||x - c||^2 \) and \( ||x - x_j||^2 \) when giving \( S_i, x_i, \) and \( x_j. \) (c) \( B_l(x) \) in \( S_i. \)

(c) If \( ||x - c||^2 - r_i \geq \max_{x \in S_i} d_{k-1} \) all the objects in \( S_i \) do not belong to \( B_l(x). \)

(d) If \( ||x + c||^2 + r_i \leq \min_{x \in S_i} d_{k-1} \) all the objects in \( S_i \) belong to \( B_l(x). \)

Similar to rough set theory [21,24,25], we use these rules to build the upper and lower approximations for each \( B_l(x) \), namely, \( \overline{B}_l(x) \) and \( \underline{B}_l(x) \), which are described as follows:

\[
\overline{B}_l(x) = \{ x_j | ||x - c||^2 - ||x_j - c||^2 \leq d_{k-1}, x_j \in S_i \} \tag{7}
\]

and

\[
\underline{B}_l(x) = \{ x_j | ||x - c||^2 + ||x_j - c||^2 \leq d_{k-1}, x_j \in S_i \}. \tag{8}
\]

We use the two sets to describe approximately \( B_l(x) \) (Fig. 3). \( \overline{B}_l(x) \) denotes a set including the objects that belong to \( B_l(x). \) \( \overline{B}_l(x) \) denotes a set including the objects that may belong to \( B_l(x). \) \( S_i - \overline{B}_l(x) \) denotes a set including the objects that do not belong to \( B_l(x). \) These sets have the following relation:

\[
\overline{B}_l(x) \subseteq \underline{B}_l(x) \subseteq \overline{B}_l(x). \tag{9}
\]
The boundary of $B_n^l(\mathbf{x}_i)$ is given as follows:

$$\text{B}_n^l(\mathbf{x}_i) = \overline{B}_n^l(\mathbf{x}_i) - B_n^l(\mathbf{x}_i).$$

For each $\psi_n^l(\mathbf{x}_i)$, we obtain the following relation:

$$\psi_n^l(\mathbf{x}_i) = \sum_{\mathbf{x}_k \in \tilde{B}^l_n(\mathbf{x}_i)} \max \left\{ 0, d_{h-1} - \| \mathbf{x}_i - \mathbf{x}_k \|^2 \right\}$$

$$= \sum_{\mathbf{x}_k \in \tilde{B}^l_n(\mathbf{x}_i)} \left( d_{h-1} - \| \mathbf{x}_i - \mathbf{x}_k \|^2 \right) + \sum_{\mathbf{x}_k \neq \mathbf{x}_j \in \tilde{B}^l_n(\mathbf{x}_i)} \max \left\{ 0, d_{h-1} - \| \mathbf{x}_i - \mathbf{x}_k \|^2 \right\}$$

$$= \sum_{\mathbf{x}_k \in \tilde{B}^l_n(\mathbf{x}_i)} d_{h-1} - \sum_{\mathbf{x}_k \in \tilde{B}^l_n(\mathbf{x}_i)} \| \mathbf{x}_i - \mathbf{x}_k \|^2 + \sum_{\mathbf{x}_k \neq \mathbf{x}_j \in \tilde{B}^l_n(\mathbf{x}_i)} \max \left\{ 0, d_{h-1} - \| \mathbf{x}_i - \mathbf{x}_k \|^2 \right\}$$

$$= \sum_{\mathbf{x}_k \in \tilde{B}^l_n(\mathbf{x}_i)} d_{h-1} - \sum_{\mathbf{x}_k \in \tilde{B}^l_n(\mathbf{x}_i)} \| \mathbf{B}_n^l(\mathbf{x}_i) \|_2 \| \mathbf{x}_k \|^2 - \sum_{\mathbf{x}_k \neq \mathbf{x}_j \in \tilde{B}^l_n(\mathbf{x}_i)} \| \mathbf{x}_k \|^2 + 2 \mathbf{x}_i \sum_{\mathbf{x}_k \neq \mathbf{x}_j \in \tilde{B}^l_n(\mathbf{x}_i)} \mathbf{x}_j$$

$$\sum_{\mathbf{x}_k \neq \mathbf{x}_j \in \tilde{B}^l_n(\mathbf{x}_i)} \max \left\{ 0, d_{h-1} - \| \mathbf{x}_i - \mathbf{x}_k \|^2 \right\}$$

$$= \sum_{\mathbf{x}_k \neq \mathbf{x}_j \in \tilde{B}^l_n(\mathbf{x}_i)} \| \mathbf{B}_n^l(\mathbf{x}_i) \|_2 \| \mathbf{x}_k \|^2 - \sum_{\mathbf{x}_k \neq \mathbf{x}_j \in \tilde{B}^l_n(\mathbf{x}_i)} \| \mathbf{x}_k \|^2$$

$$+ \sum_{\mathbf{x}_k \neq \mathbf{x}_j \in \tilde{B}^l_n(\mathbf{x}_i)} \max \left\{ 0, d_{h-1} - \| \mathbf{x}_i - \mathbf{x}_k \|^2 \right\}$$

where $\mathbf{x}_i = \frac{\sum_{\mathbf{x}_k \in \tilde{B}^l_n(\mathbf{x}_i)} \mathbf{x}_k}{|\tilde{B}^l_n(\mathbf{x}_i)|}$ is the mean of the objects in $\tilde{B}^l_n(\mathbf{x}_i)$. By computing $\psi_n^l(\mathbf{x}_i)$ according to (11), we can derive the following observations:

1. If rule (c) is satisfied, then $\overline{B}_n^l(\mathbf{x}_i) = \emptyset$. In this case, we can directly reject all the objects in $S_l$ and set $\psi_n^l(\mathbf{x}_i) = 0$.
2. If rule (d) is satisfied, then $\overline{B}_n^l(\mathbf{x}_i) = S_l$. In this case, we can directly compute

$$\psi_n^l(\mathbf{x}_i) = \sum_{\mathbf{x}_k \in \tilde{B}^l_n(\mathbf{x}_i)} d_{h-1} - |S_l| \left( \| \mathbf{x}_i - \mathbf{c}_l \|^2 - \| \mathbf{c}_l \|^2 \right) - \sum_{\mathbf{x}_k \neq \mathbf{x}_j \in \tilde{B}^l_n(\mathbf{x}_i)} \| \mathbf{x}_k \|^2.$$

On the basis of the above analyses, we can reduce lots of unnecessary distance calculations by $\overline{B}_n^l(\mathbf{x}_i)$ and $\overline{B}_n^l(\mathbf{x}_i)$ for $1 \leq l \leq k'$. We will further reduce the computational complexity. According to Step 4 of the FGKM algorithm, only the object $\mathbf{x}_i$ with the maximum $b_i^l$ should be identified, indicating that computing the exact $b_i^l$ value for each $\mathbf{x}_i$, $1 \leq i \leq n$, is unnecessary. Therefore, we will first use the approximate description of $B_n^l(\mathbf{x}_i)$ by $\overline{B}_n^l(\mathbf{x}_i)$ and $\tilde{B}_n^l(\mathbf{x}_i)$ to calculate the upper bound $b_i^l$ of $b_i^l$ for $1 \leq l \leq n$, which is defined as follows:
\[
\hat{b}_h^l = \sum_{l=1}^{k'} \psi_h^l(x_i),
\]

where

\[
\psi_h^l(x_i) = \begin{cases} 
0, & \text{if } \bar{B}_h(x_i) = \emptyset, \\
|S_l| (||x_i - c||^2 - ||c||^2) + \sum_{x \in S_l} ||x||^2, & \text{if } \bar{B}_h(x_i) = S_l, \\
\sum_{x \in \bar{B}_h(x_i)} (d_h(x) - ||x_i - c||^2 - ||x_i - c||^2), & \text{otherwise}.
\end{cases}
\]

Given that \(\psi_h^l(x_i) \leq \psi_h^l(x_i)\) for \(1 \leq l \leq k'\), we have \(b_h^l \leq b_h^l\).

After obtaining the upper bound \(b_h^l\), if \(b_h^l < \varepsilon \leq \max_{h=1}^{k'} b_h^l\), we can easily conclude that \(x_i\) cannot be the new initial cluster center. This conclusion can further reduce the amount of distance calculations. Here, \(\varepsilon\) is a parameter set to the maximum value of all obtained exact \(b_h^l\) if they exist; otherwise, is set to zero.

The new acceleration mechanism is shown in Table 1, which describes how to use the above approximate description to obtain rapidly the \(h\)th initial cluster center, for \(1 < h < k\). We use the mechanism to expedite the clustering procedure of the FGKM algorithm. The accelerated FGKM algorithm is called FGKM+A which is described as follows:

**Table 1**

An acceleration mechanism for producing the \(h\)th initial cluster center.

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Set (\varepsilon = 0)</td>
</tr>
<tr>
<td>2</td>
<td>for each data object (x_i) in (U)</td>
</tr>
<tr>
<td>3</td>
<td>for each subset (S_i) in (S)</td>
</tr>
<tr>
<td>4</td>
<td>set (\bar{B}_0 = b_0^i = \emptyset)</td>
</tr>
<tr>
<td>5</td>
<td>if rule (c) is satisfied</td>
</tr>
<tr>
<td>6</td>
<td>set (\psi_h^l(x_i) = \psi_h^l(x_i) = 0)</td>
</tr>
<tr>
<td>7</td>
<td>continue;</td>
</tr>
<tr>
<td>8</td>
<td>end if</td>
</tr>
<tr>
<td>9</td>
<td>if rule (d) is satisfied</td>
</tr>
<tr>
<td>10</td>
<td>compute (\psi_h^l(x_i)), according to Eq. (12)</td>
</tr>
<tr>
<td>11</td>
<td>set (\psi_h^l(x_i) = \psi_h^l(x_i)) and (\bar{B}_h = \bar{B}_h = S_i)</td>
</tr>
<tr>
<td>12</td>
<td>else</td>
</tr>
<tr>
<td>13</td>
<td>for each data object (x_i) in (S_i)</td>
</tr>
<tr>
<td>14</td>
<td>if (</td>
</tr>
<tr>
<td>15</td>
<td>(\bar{B}_h = \bar{B}_h \cup {x_i})</td>
</tr>
<tr>
<td>16</td>
<td>end if</td>
</tr>
<tr>
<td>17</td>
<td>if (</td>
</tr>
<tr>
<td>18</td>
<td>(\bar{B}_h = \bar{B}_h \cup {x_i})</td>
</tr>
<tr>
<td>19</td>
<td>end if</td>
</tr>
<tr>
<td>20</td>
<td>end for</td>
</tr>
<tr>
<td>21</td>
<td>compute (\psi_h^l(x_i)), according to Eq. (14)</td>
</tr>
<tr>
<td>22</td>
<td>end if</td>
</tr>
<tr>
<td>23</td>
<td>end for</td>
</tr>
<tr>
<td>24</td>
<td>compute (\hat{b}<em>h^l = \sum</em>{i=1}^{k'} \psi_h^l(x_i))</td>
</tr>
<tr>
<td>25</td>
<td>if (\hat{b}_h^l \leq \varepsilon)</td>
</tr>
<tr>
<td>26</td>
<td>set (b_h^l = 0)</td>
</tr>
<tr>
<td>27</td>
<td>continue;</td>
</tr>
<tr>
<td>28</td>
<td>else</td>
</tr>
<tr>
<td>29</td>
<td>for (l = 1) to (k)</td>
</tr>
<tr>
<td>30</td>
<td>if (\bar{B}_h - \hat{b}_h^l \neq 0)</td>
</tr>
<tr>
<td>31</td>
<td>compute (\psi_h^l(x_i)), according to Eq. (11)</td>
</tr>
<tr>
<td>32</td>
<td>end if</td>
</tr>
<tr>
<td>33</td>
<td>end for</td>
</tr>
<tr>
<td>34</td>
<td>compute (\hat{b}<em>h^l = \sum</em>{i=1}^{k'} \psi_h^l(x_i))</td>
</tr>
<tr>
<td>35</td>
<td>if (\varepsilon &lt; \hat{b}_h^l)</td>
</tr>
<tr>
<td>36</td>
<td>set (\varepsilon = \hat{b}_h^l) and (q = i)</td>
</tr>
<tr>
<td>37</td>
<td>end if</td>
</tr>
<tr>
<td>38</td>
<td>end for</td>
</tr>
<tr>
<td>39</td>
<td>output the object (x_i);</td>
</tr>
</tbody>
</table>

Please cite this article in press as: L. Bai et al., Fast global k-means clustering based on local geometrical information, Inform. Sci. (2013), http://dx.doi.org/10.1016/j.ins.2013.05.023
The FGKM+A algorithm

Table 2
Data sets from UCI.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Objects</th>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Handwritten digits</td>
<td>5620</td>
<td>64</td>
</tr>
<tr>
<td>Statlog</td>
<td>6435</td>
<td>36</td>
</tr>
<tr>
<td>Musk</td>
<td>6598</td>
<td>168</td>
</tr>
<tr>
<td>Isolet</td>
<td>7797</td>
<td>617</td>
</tr>
<tr>
<td>Coil</td>
<td>9000</td>
<td>86</td>
</tr>
<tr>
<td>Letters</td>
<td>20,000</td>
<td>16</td>
</tr>
<tr>
<td>Shuttle</td>
<td>58,000</td>
<td>9</td>
</tr>
<tr>
<td>Corel image</td>
<td>68,040</td>
<td>89</td>
</tr>
</tbody>
</table>

Fig. 4. (a) Computational times for different numbers of clusters on the handwritten digits data. (b) Numbers of distance calculations for different numbers of clusters on the handwritten digits data.

Fig. 5. (a) Computational times for different numbers of clusters on the statlog data. (b) Numbers of distance calculations for different numbers of clusters on the statlog data.

Please cite this article in press as: L. Bai et al., Fast global k-means clustering based on local geometrical information, Inform. Sci. (2013), http://dx.doi.org/10.1016/j.ins.2013.05.023
Step 1. Randomly select $k_0$ initial cluster centers and apply the $k$-means algorithm to partition the data set into $k_0$ subsets, i.e., $S = \{S_1, S_2, \ldots, S_{k_0}\}$. Furthermore, save the center of each subset and the distances between each object and these centers.

Step 2. Compute $V_1 = \{v_1\}$ from the data set $X$, where $v_1 = \sum_{i=1}^{n} x_i / n$ and $n$ is the number of objects in $U$. Set $h = 1$.

Step 3. Set $h = h + 1$. If $h > k$, then stop.

Step 4. Use the acceleration mechanism in Table 1 to select the object $x_q$ from $U$ as the $h$th initial cluster center. Set $V = V_{h-1} \cup \{x_q\}$.

Step 5. Apply the $k$-means algorithm with the initial set of cluster centers $V$, save the resulting set of cluster centers into $V_h$ and compute $d_j^h$ for each $x_j \in U$. Goto Step 2.

4. Space and time complexity

4.1. Space complexity

In the FGKM+A algorithm, we need to save a partition vector $Pc = [p_1, p_2, \ldots, p_n]$, where $p_i = l$ if the object $x_i$ belongs to $S_l$; the distance matrix $D = [D_{il}]$ which is a $n \times k'$ real matrix; $D_{il} = ||x_i - c_l||^2$ for $1 \leq i \leq n, 1 \leq l \leq k'$ and $||x_i||^2$ for each object $x_i$, $1 \leq i \leq n$. The above procedure requires $O(n(k' + 2))$ space. Given that $k' \ll n$, $n(k' + 2) \ll n^2$.

Fig. 6. (a) Computational times for different numbers of clusters on the musk data. (b) Numbers of distance calculations for different numbers of clusters on the musk data.

Fig. 7. (a) Computational times for different numbers of clusters on the isolet data. (b) Numbers of distance calculations for different numbers of clusters on the isolet data.

Please cite this article in press as: L. Bai et al., Fast global $k$-means clustering based on local geometrical information, Inform. Sci. (2013), http://dx.doi.org/10.1016/j.ins.2013.05.023
4.2. Time complexity

In Step 1, we apply the $k$-means algorithm to partition the data set into $k'$ subsets, which needs $O(nk't)$ distance calculations, where $t$ is the number of iterations. In the procedure, the center $c_i$ of each subset $S_l(1 \leq l \leq k')$ and the distances between each object $x_i(1 \leq i \leq n)$ and all the centers can be obtained. Furthermore, we need to calculate $|x_i|^2$ for each object $x_i(1 \leq i \leq n)$, which is $O(n)$ operations. In Step 4, we need $O(n_1n_2)$ distance calculations to obtain the $l$th initial cluster center, where $n_1(<n)$ and $n_2(<n)$ are the numbers of objects and distance calculations required to obtain the exact $b_i$, respectively. To generate $k$ cluster centers, the proposed algorithm needs $O(nk't + n + n_1n_2k + nk^2t)$ distance calculations. Given that the computational complexity of FGKM is $O(n^2k + nk^2t)$ in terms of the number of distance calculations, we may conclude that the proposed algorithm has less computational complexity.

5. Experimental results

To verify the efficiency of the proposed algorithm, experiments with eight standard data sets are conducted on an Intel Q9400 computer with 2G RAM. These data sets (Table 2) are downloaded from the UCI Machine Learning Repository [28].
We compare the FGKM+A algorithm with the FGKM algorithm proposed by Likas et al. [22] and the MFGKM algorithm proposed by Lai and Huang [18] in terms of computing time and number of distance calculations. In the following experiments, we set $k_0 = \sqrt{n}$ for FGKM+A.

Figs. 4–11 show the total execution time and distance calculations of these algorithms on the eight data sets with the different numbers of clusters, respectively. These data sets include the handwritten digits, statlog, musk, isolet, coil, letters, shuttle and corel image data sets which have different sizes. On each of these provided data sets, the FGKM+A algorithm outperforms the FGKM and MFGKM algorithms in terms of computing time and distance calculations. When the number of clusters $k$ increases, the efficiency of the proposed algorithm becomes more remarkable than the FGKM and MFGKM algorithms.

Furthermore, we test the scalability with the different numbers of dimensions on the two data sets, namely, the musk and isolet data sets. We fix the numbers of clusters $k$ to be as 20. Figs. 12 and 13 show that the FGKM+A algorithm exhibits better scalability with increasing dimensions, compared with the FGKM and MFGKM algorithms.

Fig. 10. (a) Computational times for different numbers of clusters on the shuttle data. (b) Numbers of distance calculations for different numbers of clusters on the shuttle data.

Fig. 11. (a) Computational times for different numbers of clusters on the corel image data. (b) Numbers of distance calculations for different numbers of clusters on the corel image data.
6. Conclusions

To improve the efficiency of the FGKM clustering algorithm, an acceleration mechanism has been developed in this paper by using the local geometrical information of data objects. In the development, an approximate description of an object set has been proposed to help users reduce the computational complexity of determining new cluster centers. Compared with the FGKM and MFGKM algorithms, the accelerated FGKM algorithm, i.e., FGKM+A, requires less computing time and fewer distance calculations while retaining the same clustering results. The performance of the proposed algorithm is more remarkable as the number of dimensions or clusters of a data set increases.

Acknowledgements

The authors are very grateful to the editors and reviewers for their valuable comments and suggestions. This work was supported by the National Natural Science Foundation of China (No. 71031006), the Foundation of Doctoral Program Research of Ministry of Education of China (No. 20101401110002), the National Key Basic Research and Development Program of China (973) (No. 2013CB329404).

Please cite this article in press as: L. Bai et al., Fast global k-means clustering based on local geometrical information, Inform. Sci. (2013), http://dx.doi.org/10.1016/j.ins.2013.05.023
References


Please cite this article in press as: L. Bai et al., Fast global k-means clustering based on local geometrical information, Inform. Sci. (2013), http://dx.doi.org/10.1016/j.ins.2013.05.023