Odjel za matematiku
Sveučilište Josipa Jurja Strossmayera u Osijeku

Department of Mathematics
J. J. Strossmayer University of Osijek

Serije: Rukopisi u pripremi

Series: Technical Reports

Analysis of the k-means algorithm in the case of data points occuring on the border of two or more clusters

Rudolf Scitovski
Kristian Sabo

Analysis of the k-means algorithm in the case of data points occurring on the border of two or more clusters

Rudolf Scitovski[1]
Department of Mathematics, University of Osijek
Trg Lj. Gaja 6, HR – 31 000 Osijek, Croatia
e-mail: scitowsk@mathos.hr

Kristian Sabo
Department of Mathematics, University of Osijek
Trg Lj. Gaja 6, HR – 31 000 Osijek, Croatia
e-mail: ksabo@mathos.hr

Abstract. In this paper, the well-known k-means algorithm for searching for a locally optimal partition of the set $\mathcal{A} \subset \mathbb{R}^n$ is analyzed in the case if some data points occur on the border of two or more clusters. For this special case, a useful strategy by implementation of the k-means algorithm is proposed.

Key words: k-means; clustering; data mining;

1 Introduction

Clustering or grouping a data set into conceptually meaningful clusters is a well-studied problem in recent literature, and it has practical importance in a wide variety of applications (Gan et al., 2007; Iyigun, 2007; Jain, 2010; Liao et al., 2012; Morales-Esteban et al., 2010; Mostafa, 2013; Pintér, 1996; Sabo et al., 2011; Scitovski and Scitovski, 2013).

A partition of the set $\mathcal{A} = \{a_i \in \mathbb{R}^n : i = 1, \ldots, m\} \subset \mathbb{R}^n$ into k disjoint subsets $\pi_1, \ldots, \pi_k$, $1 \le k \le m$, such that

$$\bigcup_{i=1}^{k} \pi_i = \mathcal{A}, \qquad \pi_r \cap \pi_s = \emptyset, \quad r \neq s, \qquad |\pi_j| \geq 1, \quad j = 1, \ldots, k, \tag{1}$$

will be denoted by $\Pi(\mathcal{A}) = \{\pi_1, \ldots, \pi_k\}$ and the set of all such partitions by $\mathcal{P}(\mathcal{A}, k)$. The elements $\pi_1, \ldots, \pi_k$ of the partition Π are called *clusters in* $\mathbb{R}^n$.

Suppose also that a weight $w_i > 0$ is associated to each data point. If $d\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$, $\mathbb{R}_+ = [0, +\infty\rangle$ is some distance-like function (see e.g. Kogan (2007); Teboulle (2007)), then to each cluster $\pi_j \in \Pi$ we can associate its center c_j defined by

$$c_j = c(\pi_j) := \operatorname*{argmin}_{x \in \operatorname{conv}(\pi_j)} \sum_{a_i \in \pi_j} w_i d(x, a_i). \tag{2}$$

[1]Corresponding author: Rudolf Scitovski, e-mail: scitowsk@mathos.hr, Telephone number: ++385-31-224-800, Fax number: ++385-224-801

where $\mathrm{conv}(\pi_j)$ denotes the convex hull of the cluster π_j. It is said that the partition $\Pi^\star \in \mathcal{P}(\mathcal{A}, k)$ is a globally optimal k-partition if

$$\Pi^\star = \operatorname*{argmin}_{\Pi\in\mathcal{P}(\mathcal{A},k)} F(\Pi), \qquad F(\Pi) = \sum_{j=1}^{k} \sum_{a_i\in\pi_j} w_i d(c_j, a_i), \tag{3}$$

where $F\colon \mathcal{P}(\mathcal{A}, k) \to \mathbb{R}_+$ is the objective function.

Conversely, for a given set of mutually different points $z_1, \ldots, z_k \in \mathbb{R}^n$, by applying the minimal distance condition, we can define the partition $\Pi = \{\pi_1, \ldots, \pi_k\}$ of the set $\mathcal{A}$, where one has to take care that every element of the set $\mathcal{A}$ occurs in one and only one cluster (Kogan, 2007; Späth, 1983). Therefore, the problem of finding an optimal partition of the set $\mathcal{A}$ can be reduced to the following global optimization problem (Sabo et al., 2013; Teboulle, 2007)

$$\operatorname*{argmin}_{z_1,\ldots,z_k\in\mathbb{R}^n} \Phi(z_1, \ldots, z_k), \qquad \Phi(z_1, \ldots, z_k) = \sum_{i=1}^{m} w_i \min_{1\le j\le k} d(z_j, a_i), \tag{4}$$

where $\Phi\colon \mathbb{R}^{kn} \to \mathbb{R}_+$. Thereby the objective function Φ can have a great number of independent variables (the number of clusters in the partition multiplied by the dimension of data points $(k \cdot n)$), it does not have to be either convex or differentiable and generally it may have several local minima. Therefore, this becomes a complex global optimization problem (Bagirov and Ugon, 2005; Bagirov, 2008; Floudas and Gounaris, 2009; Jain, 2010; Scitovski and Scitovski, 2013). The solution of (3) and (4) coincides (Späth, 1983). Since our objective function (4) is a Lipschitz continuous function (Pintér, 1996; Sabo et al., 2013), there are numerous methods for its minimization (Grbić et al., 2012; Pintér, 1996; Sergeyev and Kvasov, 2011).

The most popular algorithm for searching for a locally optimal partition is the *k-means algorithm*. By knowing a good initial approximation, this algorithm can provide acceptable solutions (Cao et al., 2009; Tasoulis and Vrahatis, 2007; Volkovich et al., 2007). In case we do not have a good initial approximation, what is usually recommended (Leisch, 2006) are multi-run algorithms with various random initializations.

In the sequel, a special *least square distance-like function* (LS-distance-like function) given by $d(x, y) = \|x - y\|_2^2$, $x, y \in \mathbb{R}^n$ will be used.

In this paper, we especially consider the problem of the occurrence of some data point on the border of two or more clusters during the execution of the k-means algorithm. Explicit criteria which clearly define locally optimal behavior in this case are proposed and proved.

The paper is organized as follows. In the next section, some auxiliary results are given. In Section 3, the optimal behavior strategy during the k-means algorithm in the case of the occurrence of some data points on the border of two clusters and the case when such data points occur on the border of several clusters are considered, because different behavior is observed in these cases. Finally, some conclusions are given in Section 4.

2 Preliminaries

Here are a few auxiliary results which will be used in the following sections. The following lemma (see e.g. Kogan (2007)) shows the relationship between the weighted sum of squares of distances from the data points to the centroid and from the data points to any point from $\mathbb{R}^n$.

Lemma 1. *Let $\mathcal{A} = \{a_i \in \mathbb{R}^n : i = 1, \dots, m\}$ be a set of data points with a corresponding set of weights $\mathcal{W} = \{w_i > 0 : i = 1, \dots, m\}$. Then for each $x \in \mathbb{R}^n$ there holds*

$$\sum_{i=1}^{m} w_i \|a_i - x\|^2 = \sum_{i=1}^{m} w_i \|a_i - c\|^2 + \sigma \|x - c\|^2, \tag{5}$$

where $\sigma = \sum_{i=1}^{m} w_i$, and $c = \frac{1}{\sigma} \sum_{i=1}^{m} w_i a_i$ is a centroid of the data.

For two disjoint sets of data with the corresponding weights, the following lemma gives explicit formulas for the centroid and the objective function value of the union of these two sets.

Lemma 2. *Let $\mathcal{A} = \{a_i \in \mathbb{R}^n : i = 1, \dots, p\}$, $\mathcal{B} = \{b_i \in \mathbb{R}^n : i = 1, \dots, q\}$ be disjoint sets with corresponding sets of weights $\mathcal{W}_A = \{\alpha_i > 0 : i = 1, \dots, p\}$, $\mathcal{W}_B = \{\beta_i > 0 : i = 1, \dots, q\}$. Then the following holds*

$$F(\{(\mathcal{W}_A; \mathcal{A}), (\mathcal{W}_B; \mathcal{B})\}) = F(\mathcal{W}_A; \mathcal{A}) + F(\mathcal{W}_B; \mathcal{B}) + \sigma_A \|c - c_A\|^2 + \sigma_B \|c - c_B\|^2, \tag{6}$$

where

$$\begin{aligned} &c_A = \tfrac{1}{\sigma_A} \sum_{i=1}^{p} \alpha_i a_i, \qquad c_B = \tfrac{1}{\sigma_B} \sum_{i=1}^{q} \beta_i b_i, \qquad \sigma_A = \sum_{i=1}^{p} \alpha_i, \qquad \sigma_B = \sum_{i=1}^{q} \beta_i, \\ &c = \tfrac{\sigma_A}{\sigma_A + \sigma_B} c_A + \tfrac{\sigma_B}{\sigma_A + \sigma_B} c_B. \end{aligned} \tag{7}$$

Proof. Formula (7) is obtained by direct checking. Furthermore, because of

$$F(\{(\mathcal{W}_A; \mathcal{A}), (\mathcal{W}_B; \mathcal{B})\}) = \sum_{i=1}^{p} \alpha_i \|a_i - c_A\|^2 + \sum_{i=1}^{q} \beta_i \|b_i - c_B\|^2,$$

if (5) is applied to each right-hand side of the sum, then we obtain (6). □

For the given set of data points with the corresponding weights the following lemma shows how the objective function value changes if the weight of some data increases or if the weight of this data vanishes.

Lemma 3. *Let $\mathcal{A} = \{a_i \in \mathbb{R}^n : i = 1, \dots, p\}$ be a data set with the corresponding set of weights $\mathcal{W} = \{w_i > 0 : i = 1, \dots, p\}$. Denote $\sigma := \sum_{i=1}^{p} w_i$ and $c_w = \frac{1}{\sigma} \sum_{i=1}^{p} w_i a_i$. Then the following holds:*

(i) If the weight w_{i_0} of the i_0-th data is increased for $\delta > 0$ and if the new set of weights is denoted by $\mathcal{W}^+$, then there holds

$$F(\mathcal{W}^+;\mathcal{A}) = F(\mathcal{W};\mathcal{A}) + \tfrac{\sigma\delta}{\sigma+\delta}\|c_w - a_{i_0}\|^2. \tag{8}$$

(ii) If the weight w_{i_0} of the i_0-th data vanishes and if the new set of weights is denoted by $\mathcal{W}^-$, then there holds

$$F(\mathcal{W}^-;\mathcal{A}) = F(\mathcal{W};\mathcal{A}) - \tfrac{w_{i_0}\sigma}{\sigma-w_{i_0}}\|c_w - a_{i_0}\|^2. \tag{9}$$

Proof. (i) Suppose that the weight w_{i_0} of the i_0-th data point is increased for δ. Formally, we can consider that a new data (δ, a_{i_0}) is added to the data set $(\mathcal{W};\mathcal{A})$. According to (7), the centroid c_{w^+} of the new data set $(\mathcal{W}^+;\mathcal{A})$ is given by

$$c_{w^+} = \tfrac{\sigma}{\sigma+\delta}c_w + \tfrac{\delta}{\sigma+\delta}a_{i_0}. \tag{10}$$

Since $F(w_{i_0};a_{i_0}) = 0$, by using (6) and (10) we obtain

$$\begin{aligned}
F(\mathcal{W}^+;\mathcal{A}) =& F(\mathcal{W};\mathcal{A}) + \sigma\|c_{w^+} - c_w\|^2 + \delta\|c_{w^+} - a_{i_0}\|^2\\
=& F(\mathcal{W};\mathcal{A}) + \sigma\|\tfrac{-\delta}{\sigma+\delta}c_w + \tfrac{\delta}{\sigma+\delta}a_{i_0}\|^2 + \delta\|\tfrac{\sigma}{\sigma+\delta}c_w - \tfrac{\sigma}{\sigma+\delta}a_{i_0}\|^2\\
=& F(\mathcal{W};\mathcal{A}) + \tfrac{\sigma\delta}{(\sigma+\delta)}\|c_w - a_{i_0}\|^2.
\end{aligned}$$

(ii) Suppose that the weight w_{i_0} of the i_0-th data point vanishes. Formally, we can consider that the data (w_{i_0}, a_{i_0}) is deleted from the data set $(\mathcal{W};\mathcal{A})$. The centroid c_{w^-} of the new data set $(\mathcal{W}^-;\mathcal{A})$ is given by

$$c_{w^-} = \tfrac{\sigma}{\sigma-w_{i_0}}c_w - \tfrac{w_{i_0}}{\sigma-w_{i_0}}a_{i_0}. \tag{11}$$

Namely,

$$c_{w^-} = \tfrac{1}{\sigma-w_{i_0}}\sum_{i=2}^{p} w_i a_i = \tfrac{1}{\sigma-w_{i_0}}\left(\sigma\tfrac{1}{\sigma}\sum_{i=1}^{p} w_i a_i - w_{i_0}a_{i_0}\right) = \tfrac{1}{\sigma-w_{i_0}}\left(\sigma c_w - w_{i_0}a_{i_0}\right).$$

Furthermore, since $F(w_{i_0};a_{i_0}) = 0$, by using (6) and (11) we obtain

$$\begin{aligned}
F(\mathcal{W};\mathcal{A}) =& F(\mathcal{W}^-;\mathcal{A}) + (\sigma - w_{i_0})\|c_w - c_{w^-}\|^2 + w_{i_0}\|c_w - a_{i_0}\|^2\\
=& F(\mathcal{W}^-;\mathcal{A}) + (\sigma - w_{i_0})\|\tfrac{-w_{i_0}}{\sigma-w_{i_0}}c_w + \tfrac{w_{i_0}}{\sigma-w_{i_0}}a_{i_0}\|^2 + w_{i_0}\|c_w - a_{i_0}\|^2\\
=& F(\mathcal{W}^-;\mathcal{A}) + \tfrac{\sigma w_{i_0}}{\sigma-w_{i_0}}\|c_w - a_{i_0}\|^2,
\end{aligned}$$

from which (9) follows immediately.

□

3 Analysis of the k-means algorithm in the case if some data points occur on border of two or more clusters

Let $\mathcal{A} = \{a_i \in \mathbb{R}^n : i = 1, \dots, m\}$ be a given data set with the corresponding weights $w_i > 0$. The set $\mathcal{A}$ should be divided into k ($2 \le k \le m$) disjoint unempty clusters. The most known algorithm for searching for a locally optimal partition is the k-means algorithm (Dhillon et al., 2004; Durak, 2011; Leisch, 2006; Ng, 2000; Steinley and Brusco, 2007; Su and Kogan, 2008; Tasoulis and Vrahatis, 2007), which can be described by two steps which are iteratively repeated.

Step 1 For each set of mutually different assignment points $z_1, \dots, z_k \in \mathbb{R}^n$ the set $\mathcal{A}$ should be divided into k disjoint unempty clusters $\pi_1, \dots, \pi_k$ by using the minimal distance principle

$$\pi_j = \{a \in \mathcal{A} : \|z_j - a\| \le \|z_s - a\|, \forall s = 1, \dots, k\}; \tag{12}$$

Step 2 Given a partition $\Pi = \{\pi_1, \dots, \pi_k\}$ of the set $\mathcal{A}$, one can define the corresponding centroids by

$$c_j = \operatorname*{argmin}_{x \in \operatorname{conv} \pi_j} \sum_{a_i \in \pi_j} w_i \|x - a_i\|^2 = \tfrac{1}{W_j} \sum_{a_i \in \pi_j} w_i a_i, \quad W_j = \sum_{a_i \in \pi_j} w_i, \quad j = 1, \dots, k; \tag{13}$$

Suppose that in Step 1 some data point might occur on the border of two or several clusters. An example of such situation in applications is a uniform distribution of the number of voters of some country in several constituencies. Thereby a requirement to divide the voters of some city into two or several constituencies (clusters) appears almost always (Sabo et al., 2012; Ricca et al., 2011). Such situations in fuzzy clustering are also considered (see e.g. Peters (2006)). A decision on alignment of this data point to some cluster can significantly determine a further flow of the iterative process. Thereby, it will be shown that there exists an essential difference in the case when this data point lies on the border of two clusters and in the case when this data point lies on the border of several clusters. Therefore, we will carry out a separate analysis for these two cases, whereby the following lemma will play an important role.

Lemma 4. *Let $m_1, \dots, m_\kappa \ge 2$ be $\kappa \ge 2$ integers. Then for each $r = 1, \dots, \kappa$ there holds*

$$\delta_r := \tfrac{(\kappa-1)(1+\kappa(m_r-1))}{\kappa^2 m_r} - \tfrac{1}{\kappa^2} \sum_{s \ne r} \tfrac{1+\kappa(m_s-1)}{m_s-1} < 0, \tag{14}$$

and

$$\Delta_{rt} := \delta_r - \delta_t = \tfrac{(1+\kappa(m_r-1))^2}{\kappa^2 m_r(m_r-1)} - \tfrac{(1+\kappa(m_t-1))^2}{\kappa^2 m_t(m_t-1)}, \quad \forall r, t \in \{1, \dots, \kappa\}. \tag{15}$$

Thereby, if $\kappa = 2$, then

$$\Delta_{rt} < 0 \quad \Leftrightarrow \quad m_r > m_t; \tag{16}$$

and if $\kappa \geq 3$, then

$$\Delta_{rt} < 0 \quad \Leftrightarrow \quad m_r < m_t. \tag{17}$$

Proof. Inequality (14) follows from

$$\begin{aligned}
\tfrac{(\kappa-1)(1+\kappa(m_r-1))}{\kappa^2 m_r} - \tfrac{1}{\kappa^2}\sum_{s\neq r}\tfrac{1+\kappa(m_s-1)}{m_s-1} &= \tfrac{1}{m_r}\left(1-\tfrac{1}{\kappa}\right)\left(\tfrac{1}{\kappa}+m_r-1\right) - \tfrac{1}{\kappa^2}\sum_{s\neq r}\tfrac{1}{m_s-1} - \tfrac{1}{\kappa^2}(\kappa-1)\kappa \\
&= -\tfrac{1}{m_r}\left(1-\tfrac{1}{\kappa}\right)^2 + 1 - \tfrac{1}{\kappa} - \tfrac{1}{\kappa^2}\sum_{s\neq r}\tfrac{1}{m_s-1} - 1 + \tfrac{1}{\kappa} \\
&= -\tfrac{1}{m_r}\left(1-\tfrac{1}{\kappa}\right)^2 - \tfrac{1}{\kappa^2}\sum_{s\neq r}\tfrac{1}{m_s-1} < 0.
\end{aligned}$$

Equality (15) follows from

$$\begin{aligned}
\Delta_{rt} &= \tfrac{(\kappa-1)(1+\kappa(m_r-1))}{\kappa^2 m_r} - \tfrac{1}{\kappa^2}\sum_{s\neq r}\tfrac{1+\kappa(m_s-1)}{m_s-1} - \tfrac{(\kappa-1)(1+\kappa(m_t-1))}{\kappa^2 m_t} + \tfrac{1}{\kappa^2}\sum_{s\neq t}\tfrac{1+\kappa(m_s-1)}{m_s-1} \\
&= \tfrac{(\kappa-1)(1+\kappa(m_r-1))}{\kappa^2 m_r} + \tfrac{1}{\kappa^2}\tfrac{1+\kappa(m_r-1)}{m_r-1} - \left(\tfrac{(\kappa-1)(1+\kappa(m_t-1))}{\kappa^2 m_t} + \tfrac{(\kappa-1)(1+\kappa(m_t-1))}{\kappa^2 m_t}\right) \\
&= \tfrac{(1+\kappa(m_r-1))^2}{\kappa^2 m_r(m_r-1)} - \tfrac{(1+\kappa(m_t-1))^2}{\kappa^2 m_t(m_t-1)}.
\end{aligned}$$

In order to prove equivalences (16) and (17), we define an auxiliary function $f\colon [2,+\infty\rangle \times [1,+\infty\rangle \to \mathbb{R}$,

$$f(\kappa,x) := \tfrac{(1+\kappa(x-1))^2}{\kappa^2 x(x-1)}, \tag{18}$$

whose partial derivative with respect to the variable x can be written as

$$\tfrac{\partial f(\kappa,x)}{\partial x} = \tfrac{((\kappa-1)^2-1)x^2-2(\kappa-1)^2x+(\kappa-1)^2}{\kappa^2x^2(x-1)^2}. \tag{19}$$

If $\kappa = 2$, from (19) it can be seen that function (18) is decreasing according to x, and therefore (16) holds. If $\kappa > 2$, from (19) it can be seen that function (18) is increasing according to x, and therefore (17) holds. $\square$

Remark 1. Note that specially for $\kappa = 2$ and $m_1 = p \geq 2$, $m_2 = q \geq 2$ inequality (14) becomes

$$\tfrac{2p-1}{4p} - \tfrac{2q-1}{4(q-1)} < 0 \quad \text{and} \quad \tfrac{2q-1}{4q} - \tfrac{2p-1}{4(p-1)} < 0,$$

that is equivalent to a simple inequality

$$p + q > 1, \tag{20}$$

which is fulfilled in this case.

Furthermore, (15) becomes

$$\Delta := \tfrac{(1-2p)^2}{4p(p-1)} - \tfrac{(1-2q)^2}{4q(q-1)},$$

thereby

$$\Delta < 0 \quad \Leftrightarrow \quad p > q. \tag{21}$$

3.1 Some data point can occur on the border of two clusters

Let $\mathcal{A} = \{a_i \in \mathbb{R}^n \colon i = 1, \dots, m\}$ be a set of data points, which should be divided into two unempty disjoint clusters by applying the k-means algorithm (12)-(13). We start from Step 1, choose two different assignment points z_1, z_2, and by applying the minimal distance principle define clusters

$$\pi_1 = \{a \in \mathcal{A} \colon \|z_1 - a\| < \|z_2 - a\|\}, \quad |\pi_1| = p - 1, \tag{22}$$

$$\pi_2 = \{a \in \mathcal{A} \colon \|z_2 - a\| < \|z_1 - a\|\}, \quad |\pi_2| = q - 1. \tag{23}$$

Suppose that thereby a data point $a_0 \in \mathcal{A}$ occurs such that

$$\|z_1 - a_0\| = \|z_2 - a_0\|. \tag{24}$$

This would mean that $\pi_1 \cup \pi_2 \neq \mathcal{A}$. Note that $m = p + q - 1$. In order to fulfill condition (1) it is usually recommended in literature (Kogan, 2007; Steinley and Brusco, 2007) that the data point a_0 is assigned either to the cluster π_1 or to the cluster π_2.

Alternatively, we can introduce weights of the data such that weight 1 is associated to all data, except the data point a_0, and the data point a_0 with weight $\frac{1}{2}$ is associated to both the cluster π_1 and the cluster π_2 (as if the data point a_0 were halved). Centroids and the objective function value of clusters obtained in that way are given by

$$c_1 = \tfrac{1}{p-\frac{1}{2}} \Big(\sum_{a_i \in \pi_1} a_i + \tfrac{1}{2} a_0 \Big), \qquad c_2 = \tfrac{1}{q-\frac{1}{2}} \Big(\sum_{a_i \in \pi_2} a_i + \tfrac{1}{2} a_0 \Big), \tag{25}$$

$$F_0 = \sum_{a_i \in \pi_1} \|c_1 - a_i\|^2 + \sum_{a_i \in \pi_2} \|c_2 - a_i\|^2 + \tfrac{1}{2}\|c_1 - a_0\|^2 + \tfrac{1}{2}\|c_2 - a_0\|^2. \tag{26}$$

If the whole data point a_0 is assigned to the cluster π_1, we obtain a new centroid of the cluster π_1 given by (10) and a new centroid of the cluster π_2 given by (11), and by using (8) and (9) we get a new objective function value

$$F_1 := F_0 + \tfrac{2p-1}{4p}\|c_1 - a_0\|^2 - \tfrac{2q-1}{4(q-1)}\|c_2 - a_0\|^2. \tag{27}$$

If the whole data point a_0 is assigned to the cluster π_2, we obtain a new centroid of the cluster π_2 given by (10) and a new centroid of the cluster π_1 given by (11), and by using (8) and (9) we get a new corresponding objective function value

$$F_2 := F_0 + \tfrac{2q-1}{4q}\|c_2 - a_0\|^2 - \tfrac{2p-1}{4(p-1)}\|c_1 - a_0\|^2. \tag{28}$$

One also gets

$$\Delta := F_1 - F_2 = \tfrac{(1-2p)^2}{4p(p-1)}\|c_1 - a_0\|^2 - \tfrac{(1-2q)^2}{4q(q-1)}\|c_2 - a_0\|^2. \tag{29}$$

Similar formulas appear in the Incremental k-means algorithm (Kogan2007, Steinley2007).

In the following theorem we summarize the obtained results and show the manner of optimal behavior in the case when some data point $a_0 \in \mathcal{A}$ occurs on the border of two clusters.

Theorem 1. *Let $\mathcal{A} = \{a_i \in \mathbb{R}^n : i = 1, \ldots, m\}$ be a set of data points, let z_1, z_2 be two different assignment points by which clusters* (22)-(23) *are defined, and let there exist a data point $a_0 \in \mathcal{A}$, such that $\|z_1 - a_0\| = \|z_2 - a_0\|$. Then*

(i) If the data point a_0 is uniformly divided on both clusters π_1 and π_2, centroids c_1, c_2 of clusters are given by (25)*, and the corresponding objective function value F_0 is given by* (26)*;*

(ii) If the data point a_0 is assigned to the cluster π_1 completely, the new objective function value F_1 is given by (27)*;*

(iii) If the data point a_0 is assigned to the cluster π_2 completely, the new objective function value F_2 is given by (28)*;*

(iv) If the data point a_0 is assigned either to the cluster π_1 or to the cluster π_2 completely, a reduction in the objective function value is attained, i.e. $\min\{F_1, F_2\} < F_0$*;*

(v) If $\lambda_1 = \frac{(1-2p)^2}{4p(p-1)}\|c_1 - a_0\|^2$, $\lambda_2 = \frac{(1-2q)^2}{4q(q-1)}\|c_2 - a_0\|^2$, then a reduction in the objective function value is attained by assigning the data point a_0 completely to the cluster π_1 (i.e. to the cluster π_2) if and only if $\lambda_1 \leq \lambda_2$ (i.e. $\lambda_2 \leq \lambda_1$).

Proof. (iv) If $\|c_1 - a_0\| \leq \|c_2 - a_0\|$, then according to Remark 1, from (27) there follows

$$F_1 \leq F_0 + \left(\tfrac{2p-1}{4p} - \tfrac{2q-1}{4(q-1)}\right)\|c_2 - a_0\|^2 < F_0,$$

which means that in this case a reduction in the objective function value can be attained by assigning the data point a_0 completely to the cluster π_1. Analogously, if $\|c_2 - a_0\| \leq \|c_1 - a_0\|$, then according to Remark 1, from (28) there follows

$$F_2 = F_0 + \left(\tfrac{2q-1}{4q} - \tfrac{2p-1}{4(p-1)}\right)\|c_1 - a_0\|^2 < F_0,$$

which means that in this case a reduction in the objective function value can be attained by assigning the data point a_0 completely to the cluster π_2. So, a reduction in the objective function value can always be attained by assigning the data point a_0 either to the cluster π_1 or to the cluster π_2 completely.

(v) follows immediately from (29). □

Corollary 1. *Let the data be given as in Theorem 1. If $\|c_1 - a_0\| \leq \|c_2 - a_0\|$, the lower objective function value is attained by assigning the data point a_0 completely to the cluster with more data.*

Proof. If $\|c_1 - a_0\| \leq \|c_2 - a_0\|$, the difference (29) becomes

$$\Delta \leq \left(\tfrac{(1-2p)^2}{4p(p-1)} - \tfrac{(1-2q)^2}{4q(q-1)}\right)\|c_1 - a_0\|^2,$$

and the assertion follows immediately by applying Lemma 4 and Remark 1. □

Example 1. *As an illustration of Theorem 1 we consider the data set* $A \subset \mathbb{R}$*, which consists of the subset* $\pi_0 = \{1, 2\}$*, the data point* $a_0 = 6$ *and alternatively the data point* $b \in \{9, 11.4, 12\}$*. Various situations that may occur are shown in Table 1 for the same assignment points* $z_1 = 5$*,* $z_2 = 7$*. The lowest objective function value attained in particular cases is especially assigned.*

	$\{\{\pi_0, (\frac{1}{2}, 6)\}, \{(\frac{1}{2}, 6), b\}\}$							$\{\{\pi_0, 6\}, \{b\}\}$			$\{\{\pi_0\}, \{6, b\}\}$		
b	c_1	c_2	λ_1	λ_2	$d(c_1, a_0)$	$d(c_2, a_0)$	F_0	c_1	c_2	F_1	c_1	c_2	F_2
9	2.4	8	13.5	4.5	3.6	2	14	3	9	14	1.5	7.5	$\boxed{5}$
11.4	2.4	9.6	13.5	14.58	3.6	3.6	18.32	3	11.4	$\boxed{14}$	1.5	8.7	15.08
12	2.4	10	13.5	18	3.6	4	20.6	3	12	$\boxed{14}$	1.5	9	18.5

Table 1: Displacement of data points from the border of two clusters

Remark 2. If in the k-means algorithm (12)-(13) we started from Step 2 and determined centroids c_1, c_2 for given clusters, then the problem considered above might occur in the next step (Step 1): if there exists $a_0 \in \mathcal{A}$ such that $\|c_1 - a_0\| = \|c_2 - a_0\|$, a decision has to be made again referring to which cluster it should be assigned to.

3.2 Some data point can occur on the border of several clusters

Let $\mathcal{A} = \{a_i \in \mathbb{R}^n : i = 1, \ldots, m\}$ be a set of data points, which should be divided into k unempty disjoint clusters by applying the k-means algorithm (12)-(13).

We start from Step 1, choose k mutually different assignment points $z_1, \ldots, z_k \in \mathbb{R}^n$ and by applying the minimal distance principle define clusters

$$\pi_j = \{a \in \mathcal{A} \colon \|z_j - a\| < \|z_s - a\|,\, s \neq j\}, \quad |\pi_j| = m_j - 1, \quad j = 1, \ldots, k. \tag{30}$$

Suppose that thereby a data point $a_0 \in \mathcal{A}$ lies on the common border of $3 \leq \kappa \leq k$ clusters

$$\|z_1 - a_0\| = \cdots = \|z_\kappa - a_0\|. \tag{31}$$

This would mean that $\bigcup_{j=1}^{k} \pi_j \neq \mathcal{A}$. In order to fulfill condition (1) it is usually recommended in literature (Kogan, 2007; Steinley and Brusco, 2007) that the data point a_0 is assigned to some of clusters $\pi_1, \ldots, \pi_k$.

For the purposes of further analysis of such situation, without loss of generality, we furthermore suppose that $\kappa = k$, and introduce the notation $J = \{1, \ldots, \kappa\}$.

Alternatively, we can introduce weights of the data such that weight 1 is associated to all data, except the data point a_0, and the data point a_0 with weight $\frac{1}{\kappa}$ is associated to all clusters (as if the data point a_0 were uniformly divided into all κ clusters). Centroids and the objective function value of clusters obtained in that way are given by

$$c_j = \frac{1}{m_j+\frac{1}{\kappa}-1}\left(\sum_{a_i\in\pi_j} a_i + \tfrac{1}{\kappa}a_0\right), \qquad j = 1\ldots,\kappa, \tag{32}$$

$$F_0 = \sum_{j=1}^{\kappa}\sum_{a_i\in\pi_j}\|c_j - a_i\|^2 + \tfrac{1}{\kappa}\sum_{j=1}^{\kappa}\|c_j - a_0\|^2. \tag{33}$$

If the whole data point a_0 is assigned to the cluster π_r, $r \in J$, we obtain a new centroid of the cluster π_r given by (10) and new centroids of other clusters given by (11), and by using (8) and (9) we get a new corresponding objective function value

$$F_r := F_0 + \tfrac{(\kappa-1)(1+\kappa(m_r-1))}{\kappa^2 m_r}\|c_r - a_0\|^2 - \tfrac{1}{\kappa^2}\sum_{s\neq r}\tfrac{1+\kappa(m_s-1)}{m_s-1}\|c_s - a_0\|^2. \tag{34}$$

Also $\forall r, t \in J$ one gets

$$\Delta_{rt} :=F_r - F_t = \tfrac{(1+\kappa(m_r-1))^2}{\kappa^2 m_r(m_r-1)}\|c_r - a_0\|^2 - \tfrac{(1+\kappa(m_t-1))^2}{\kappa^2 m_t(m_t-1)}\|c_t - a_0\|^2. \tag{35}$$

In the following theorem we summarize the obtained results and show the manner of optimal behavior in the case when some data point $a_0 \in \mathcal{A}$ occurs on the border of several clusters.

Theorem 2. *Let $\mathcal{A} = \{a_i \in \mathbb{R}^n \colon i = 1,\ldots,m\}$ be a set of data points, let $z_1,\ldots,z_\kappa$ be mutually different assignment points by which clusters* (30) *are defined, and let there exist $a_0 \in \mathcal{A}$, such that $\|z_1 - a_0\| = \cdots = \|z_\kappa - a_0\|$. Then*

(i) If the data point a_0 is uniformly divided into all clusters, centroids $c_1,\ldots,c_\kappa$ of clusters are given by (32)*, and the corresponding objective function value F_0 is given by* (33)*;*

(ii) If the data point a_0 is assigned to the cluster π_r, $r \in J$ completely, the new objective function value F_r is given by (34)*;*

(iii) There exists $r \in J$, such that assigning the data point a_0 completely to the cluster π_r provides a reduction in the objective function value, i.e. $F_r = \min\limits_{s\in J} F_s < F_0$;

(iv) If $\lambda_j = \frac{(1+\kappa(m_j-1))^2}{\kappa^2 m_j(m_j-1)}\|c_j - a_0\|^2$, $j \in J$, then the lowest objective function value is attained by assigning the data point a_0 completely to the cluster π_{j_0}, $j_0 \in J$ if and only if $\lambda_{j_0} = \min\limits_{j\in J}\lambda_j$.

Proof. If $\|c_r - a_0\| = \min\limits_{s\in J}\|c_s - a_0\|$, then according to Lemma 4, from (34) it follows

$$F_r \leq F_0 + \left(\tfrac{(\kappa-1)(1+\kappa(m_r-1))}{\kappa^2 m_r} - \tfrac{1}{\kappa^2}\sum_{s\neq r}\tfrac{1+\kappa(m_s-1)}{m_s-1}\right)\|c_r - a_0\|^2 < F_0,$$

which means that a reduction in the objective function value can be attained by assigning the data point a_0 completely to the cluster π_r, whose centroid is nearest to the data point a_0.

According to (35), the lowest objective function value is attained by assigning the data point a_0 completely to the cluster π_{j_0} if and only if for each $s \in \{1, \dots, \kappa\}$ there holds

$$\Delta_{j_0 s} = \tfrac{(1+\kappa(m_{j_0}-1))^2}{\kappa^2 m_{j_0}(m_{j_0}-1)} \|c_{j_0} - a_0\|^2 - \tfrac{(1+\kappa(m_s-1))^2}{\kappa^2 m_s(m_s-1)} \|c_s - a_0\|^2 \leq 0,$$

i.e. if and only if $\lambda_{j_0} = \min\limits_{j \in J} \lambda_j$. □

Corollary 2. *Let the data be given as in Theorem 2. If* $\|c_1 - a_0\| = \cdots = \|c_\kappa - a_0\|$*, the lowest objective function value is attained by assigning the data point* a_0 *completely to the cluster with the least data.*

Furthermore, if π_{j_0} *is the cluster with the least data and if* $\|c_{j_0} - a_0\| \leq \|c_j - a_0\|$ *for each* $j \in J$*, then the lowest objective function value is attained by assigning the data point* a_0 *completely to the cluster* π_{j_0}*.*

Proof. If $\|c_1 - a_0\| = \cdots = \|c_\kappa - a_0\|$, equality (35) becomes

$$\Delta_{rt} = \left(\tfrac{(1+\kappa(m_r-1))^2}{\kappa^2 m_r(m_r-1)} - \tfrac{(1+\kappa(m_t-1))^2}{\kappa^2 m_t(m_t-1)} \right) \|c_1 - a_0\|^2,$$

and the assertion follows immediately by applying Lemma 4 .

The second part of the assertion follows from the fact that $\Delta_{j_0 t} < 0 \Leftrightarrow m_{j0} < m_t$ for each $t \in J \setminus \{j_0\}$. □

Example 2. *The data set is defined in the following way. First, the data point* $a_0 \in \mathbb{R}^2$ *and five assignment points* $z_1, \dots, z_5 \in \mathbb{R}^2$ *randomly chosen on the circumference with the origin at the point* a_0 *are determined. In the neighborhood of each point* z_j *have generated random points such that coordinates of points* z_j *are contaminated with binormal random additive errors with mean vector* $\mathbf{0} \in \mathbb{R}^2$ *and the identity covariance matrix. In this way we obtained a data set* $\mathcal{A}$*. According to the minimal distance principle, clusters* $\pi_j = \pi(z_j)$, $j = 1, \dots, 5$ *are defined by assignment points* $z_1, \dots, z_5$*. Thereby, the point* a_0 *lies on the common border of all five clusters (Fig. 1a).*

Table 2 gives values of parameters λ_j from Theorem 2, distances from the centroids of clusters to the data point a_0, and the objective function values obtained by assigning the data point a_0 completely to some cluster.

Fig. 1b and Fig. 1c show the partition with the lowest objective function value obtained by assigning the data point a_0 completely to the cluster π_4 and the partition with the highest objective function value obtained by assigning the data point a_0 completely to the cluster π_5, respectively.

	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$
λ_j	1.0966	1.2136	1.5409	1.3053	1.0933
$\|c_j - a_0\|$	1.0632	1.1278	1.2687	1.1591	1.0726
$F(a_0 \dashrightarrow \pi_j)$	24.9058	25.0228	25.3500	25.1144	24.9024

Table 2: Choosing the optimal position of the data point a_0

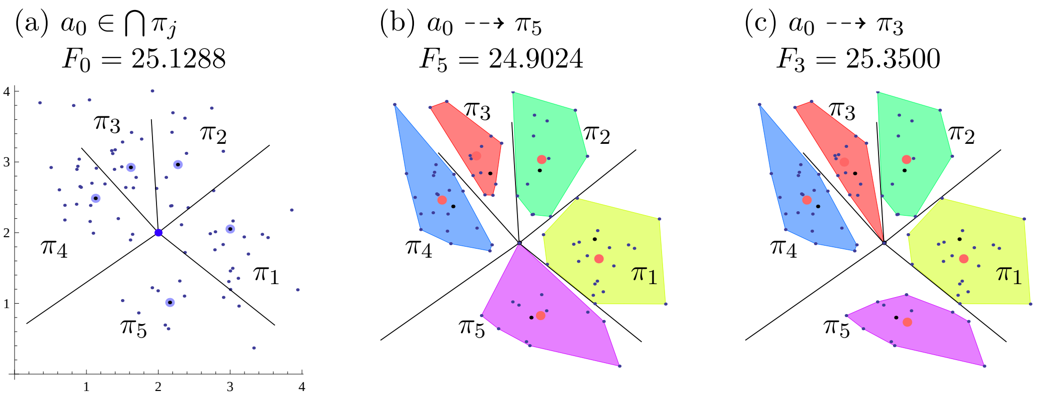

Figure 1: Choosing the optimal position of the data point a_0

4 Conclusions

The k-means algorithm is the most popular method for searching for the locally optimal partition of some data set $\mathcal{A} \subset \mathbb{R}^n$. If during the iterative process some data points occur on the border of two or more clusters, the known literature does not clearly indicate what to do. In this paper, explicit criteria which clearly define optimal behavior in this case are proposed and proved.

The position of some data point in the immediate neighborhood of the border of two or more clusters (Peters, 2006) or applications in fuzzy clustering can also be the subject of further research. This research could lead to an important improvement of the well-known k-means algorithm.

Acknowledgments

This work is supported by the Ministry of Science, Education and Sports, Republic of Croatia, through research grants `235-2352818-1034` and `165-0361621-2000`.

References

Bagirov, A.M., 2008. Modified global k-means algorithm for minimum sum-of-squares clustering problems. Pattern Recognition 41, 3192–3199.

Bagirov, A.M., Ugon, J., 2005. An algorithm for minimizing clustering functions. Optimization 54, 351–368.

Cao, F., Liang, J., Jiang, G., 2009. An initialization method for the k-means algorithm using neighborhood model. Computers and Mathematics with Applications 58, 474–483.

Dhillon, I.S., Guan, Y., Kulis, B., 2004. Kernel k-means, spectral clustering and normalized cuts, in: Proceedings of the 10-th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), August 22–25, 2004, Seattle, Washington, USA, pp. 551–556.

Durak, B., 2011. A Classification Algorithm Using Mahalanobis Distances Clustering of Data with Applications on Biomedical Data Set. Ph.D. thesis. The Graduate School of Natural and Applied Sciences of Middle East Technical University.

Floudas, C.A., Gounaris, C.E., 2009. A review of recent advances in global optimization. Journal of Global Optimization 45, 3 – 38.

Gan, G., Ma, C., Wu, J., 2007. Data Clustering: Theory, Algorithms, and Applications. SIAM, Philadelphia.

Grbić, R., Nyarko, E.K., Scitovski, R., 2012. A modification of the DIRECT method for Lipschitz global optimization for a symmetric function. Journal of Global Optimization Published online: 23 December 2012.

Hand, D.J., Krzanowski, W.J., 2005. Optimising k-means clustering results with standard software packages. Computational Statistics & Data Analysis 49, 969–973.

Hansen, P., Mladenović, N., 2001. J-means: a new local search heuristic for minimum sum of squares clustering. Pattern Recognition 34, 405–413.

Iyigun, C., 2007. Probabilistic Distance Clustering. Ph.D. thesis. Graduate School – New Brunswick, Rutgers.

Jain, A.K., 2010. Data clustering: 50 years beyond k-means. Pattern Recognition Letters 31, 651–666.

Kaufman, L., Rousseeuw, P.J., 2005. Finding groups in data: An introduction to cluster analysis. Jonh Wiley & Sons, Hoboken.

Kogan, J., 2007. Introduction to Clustering Large and High-dimensional Data. Cambridge University Press.

Leisch, F., 2006. A toolbox for k-centroids cluster analysis. Computational Statistics & Data Analysis 51, 526 – 544.

Liao, S.H., Chu, P.H., Hsiao, P.Y., 2012. Data mining techniques and applications – a decade review from 2000 to 2011. Expert Systems with Applications 39, 11303–11311.

Morales-Esteban, A., Martínez-Álvarez, F., Troncoso, A., Justo, J., Rubio-Escudero, C., 2010. Pattern recognition to forecast seismic time series. Expert Systems with Applications 37, 8333–8342.

Mostafa, M.M., 2013. More than words: Social networks' text mining for consumer brand sentiments. Expert Systems with Applications 40, 4241–4251.

Ng, M., 2000. A note on constrained k-means algorithms. Pattern Recognition 33, 525–519.

Peters, G., 2006. Some refinements of rough k-means clustering. Pattern Recognition 39, 1481 – 1491.

Pintér, J.D., 1996. Global Optimization in Action (Continuous and Lipschitz Optimization: Algorithms, Implementations and Applications). Kluwer Academic Publishers, Dordrecht.

Ricca, F., Scozzari, A., Simeone, B., 2011. The give-up problem for blocked regional lists with multi-winners. Mathematical Social Sciences 62, 14–24.

Sabo, K., Scitovski, R., Taler, P., 2012. Uniform distribution of the number of voters per constituency on the basis of a mathematical model (in croatian). Hrvatska i komparativna javna uprava 14, 229 – 249.

Sabo, K., Scitovski, R., Vazler, I., 2013. One-dimensional center-based l_1-clustering method. Optimization Letters 7, 5 – 22.

Sabo, K., Scitovski, R., Vazler, I., Zekić-Sušac, M., 2011. Mathematical models of natural gas consumption. Energy Conversion and Management 52, 1721 – 1727.

San, O.M., Huynh, V.N., Nakamori, Y., 2004. An alternative extension of the k-means algorithm for clustering categorical data. International Journal of Applied Mathematics and Computer Science 14, 241–247.

Scitovski, R., Scitovski, S., 2013. A fast partitioning algorithm and its application to earthquake investigation. Computers and Geosciences 59, 124 – 131.

Sergeyev, Y.D., Kvasov, D.E., 2011. Lipschitz global optimization, in: Cochran, J. (Ed.), Wiley Encyclopedia of Operations Research and Management Science. Wiley, New York. volume 4, pp. 2812–2828.

Späth, H., 1983. Cluster-Formation und Analyse. R. Oldenburg Verlag, München.

Steinley, D., Brusco, M.J., 2007. Initializing k-means batch clustering: a critical evaluation of several techniques. Journal of Classification 24, 99–121.

Su, Z., Kogan, J., 2008. Second order conditions for k-means clustering: Partitions vs. centroids, in: Text Mininig 2008 Workshop (held in conjuction with the 8th Siam International Conference on Data Mininig), Apr 26, 2008, Atlanta, Ga.

Su, Z., Kogan, J., Nicholas, C., 2010. Constrained clustering with k-means type algorithms, in: Berry, M.W., Kogan, J. (Eds.), Text Mining: Applications and Theory. Wiley, Chichester, p. 81–103.

Tasoulis, D., Vrahatis, M., 2007. Generalizing the k-windows clustering algorithm in metric spaces. Mathematical and Computer Modelling 46, 268–277.

Teboulle, M., 2007. A unified continuous optimization framework for center-based clustering methods. Journal of Machine Learning Research 8, 65 – 102.

Volkovich, V., Kogan, J., Nicholas, C., 2007. Building initial partitions through sampling techniques. European Journal of Operational Research 183, 1097 – 1105.