The Permutable *k*-means for the Bi-partial Criterion

Sergey Dvoenko
Tula State University, 92 Lenin Ave., Tula, Russian Federation
E-mail: sergedv@yandex.ru

Jan Owsinski
Systems Research Institute, Polish Academy of Sciences, 6 Newelska, 01 447 Warsaw, Poland
E-mail: owsinski@ibspan.waw.pl, http://www.ibspan.waw.pl/glowna/en /

Keywords: distance, similarity, dissimilarity, cluster, *k*-means, objective function

Received: December 18, 2017

The bi-partial criterion for clustering problem consists of two parts, where the first one takes into account intra-cluster relations, and the second – inter-cluster ones. In the case of k-means algorithm, such bi-partial criterion combines intra-cluster dispersion with inter-cluster similarity, to be jointly minimized. The first part only of such objective function provides the "standard" quality of clustering based on distances between objects (the well-known classical k-means). To improve the clustering quality based on the bi-partial objective function, we develop the permutable version of k-means algorithm. This paper shows that the permutable k-means appears to be a new type of a clustering procedure.

Povzetek: Študija se ukvarja z gručenjem znotraj in med gručami, pri čemer izvirna metoda uporablja permutirano verzijo običajnega algoritma za gručenje.

1 Introduction and related works

1.1 Clustering by *k*-means

According to the basic idea of the classical *k*-means algorithm [1-5], a set $\Omega=\{\omega_1,\dots\omega_N\}$ of N elements is divided into clusters Ω_k, $k=1,\dots K$, represented in a feature space by their "representative" objects $\tilde{\mathbf{x}}_k$, and/or "mean" objects $\bar{\mathbf{x}}_k$ (centers), where $\mathbf{x}=(x_1,\dots x_n)^T$ is a vector in the $n-$dimensional space.

In this paper, we consider means as representatives and calculate new means as in the classical procedure.

The well-known respective clustering criterion minimizes average of squared distances to cluster centers

$$J(K)=\frac{1}{N}\sum_{k=1}^{K}N_k\sigma_k^2=\sum_{k=1}^{K}\frac{N_k}{N}\sigma_k^2\ , \qquad (1)$$

$$\sigma_k^2=\frac{1}{N_k}\sum_{i=1}^{N_k}\|\mathbf{x}_i-\bar{\mathbf{x}}_k\|^2=\frac{1}{N_k}\sum_{i=1}^{N_k}d^2(\mathbf{x}_i,\bar{\mathbf{x}}_k)\,,$$

where σ_k^2 is the dispersion of the cluster Ω_k having size N_k, and $d(\mathbf{x},\mathbf{y})$ is the Euclidean distance between vectors $\mathbf{x}$ and $\mathbf{y}$.

As it is well-known [6–10], cluster dispersions can be calculated without direct use of cluster means, based on pairwise distances between vectors

$$\sigma_k^2=\frac{1}{2N_k^2}\sum_{i=1}^{N_k}\sum_{j=1}^{N_k}\|\mathbf{x}_i-\mathbf{x}_j\|^2=\frac{1}{2N_k^2}\sum_{i=1}^{N_k}\sum_{j=1}^{N_k}d^2(\mathbf{x}_i,\mathbf{x}_j)\,. \quad (2)$$

Empirical data often appear in the form of a matrix of pairwise comparisons of elements of the set. Such comparisons can be nonnegative values of dissimilarity or similarity of objects from the set Ω [11].

This is important for our approach, since the permutable *k*-means, developed in this paper, uses only distance $D(N,N)$ or similarity $S(N,N)$ matrices. Therefore, cluster means are not presented in them, and we need to develop equivalent forms of (1) and (2).

The basis of our approach is the Torgerson's idea of the "gravity center", developed for multidimensional scaling problem [6] in the method of double centering for principal projections to get the appropriate feature space with the raw distance matrix, immersed in it.

Our goal is different: we do not want to restore a feature space itself, since it is sufficient to suppose that objects are immersed in some metric (more closely, Euclidean) space, as we show this later on.

Naturally, the two-component criteria, similar to the bi-partial one (Dunn, Calinski-Harabasz, Xie-Beni etc.), are used in cluster-analysis [9, 12]. They are mainly used to assess the proper number of clusters *K*. Such criteria are usually heuristic constructions, used to assess the results of some algorithms of quite different origins and properties.

Here we are interested in improving the results of the classical clustering problem with a predefined number of clusters *K*. Namely, we try to develop here the bi-partial objective function to build a homogeneous and strict metric criterion for standard *k*-means algorithm only for a predefined number *K*, and not to use any other idea of procedure than that of *k*-means.

1.2 The bi-partial criterion

In order to introduce here a general bi-partial objective function, we refer to an illustrative problem of dividing a unidimensional empirical distribution of real values into a set of categories to get the "best" set in a definite sense [13–15]. This case serves merely the purpose of illustration, and assumptions made on data do not apply to the general bi-partial approach.

Let a sequence of N positive real observations $x_i, i=1,\dots N$ be given in non-decreasing order, i.e. with $x_{i+1} \geq x_i$ for all of them. Any such sequence can be represented through a cumulative form, obtained via transformation $z_i = \sum_{p=1}^{i} x_p$, $i=1,\dots N$.

As a result, we deal with a convex non-decreasing sequence $z_i, i=1,\dots N$. This means that a straight line, connecting two observations, z_q and z_s, with $1 \leq q < s \leq N$ has all values not under the corresponding observations $z_i, i=q,\dots s$.

Obviously, for the sequence of constant values $x_1 = \dots = x_N = c$ the convex cumulative form is the line $z_i = ic, i=1,\dots N,$ with $z_1 = c, z_N = cN$, represented perfectly by the single linear piece.

Otherwise, for non-constant values, by increasing the number of linear segments from the single one (with $q=1, s=N$), we steadily decrease the error of approximation of the original distribution $\{z_i\}$ by the broken line, composed of such segments, down to zero, when the maximal number $N-1$ of linear segments, corresponded to the number of observations N, is used to represent the distribution.

Under these conditions, the problem of obtaining the optimal piece-wise linear approximation of the cumulative sequence with the number of linear segments also being optimized was investigated in [13-15].

According to [13-15], the respective bi-partial objective function J_{DS} penalizes, first, deviation C_D of linear segments from the respective distribution, and, second, penalizes similarity C_S of linear segments to each other, and was represented in the form

$$J_{DS}(K) = (1-\alpha)C_D(K) + \alpha C_S(K) \to \min , \qquad (3)$$

where $K \geq 1$ is the number of segments, $0 \leq \alpha \leq 1$ is the coefficient of linear combination of two parts of the criterion.

The criterion J_{DS}, investigated in [13, 14] for the above problem, is a particular case of the general bi-partial form, representing the fundamental "intra-cluster cohesion + inter-cluster separation" paradigm [15, 16].

It should be noted that the parameter α in (3) need not appear at all, if two parts of the objective function are assumed to reflect correctly the respective inner and outer measures. Note that by solving with respect to (3) we get both the cluster (segment) content and the number of clusters (segments). We can also represent (3) in different forms to obtain different data analysis problems as particular cases. So, e.g., (3) can be transformed to the linear regression problem for $K=1,\ \alpha=0$.

In other interesting cases, the problem (3) can be considered for other kinds of parameters than α, say, K. Thus, we can treat $K \geq 1$ as a hyper-parameter and find the optimal linear combination of parts in $J_{DS}(K)$.

Thus, in the context of the illustrative problem quoted, we would fix the number of line segments K, and look with (3) for the optimum weight α, meaning the significance we attach to accuracy of the approximation vs. distinctiveness of the consecutive segments.

In this paper, we investigate the single-parametric reduced form of (3) to find the optimal α for the predefined hyper-parameter K based on the direct implementing of the well-known k-means algorithm.

2 Distance and similarity *k*-means

In this paper, we use the specially developed *k*-means algorithm only for the case of distances or similarities between objects [17, 18].

A positive definite similarity matrix can be obtained as a matrix of pairwise scalar products of object descriptions in some metric space with the dimensionality of not more than a set cardinal number. This matrix of scalar products can be transformed into a distance matrix and vice versa. As a result, the dissimilarity matrix can be used as the distance matrix in the same space.

In this case, the mean object $\omega(\bar{\mathbf{x}}_k)$ cannot be defined in Ω by the distance matrix $D(N,N)$ as a center of a cluster. Usually, the object minimizing the sum of distances to the others in the cluster can be used as the center $\bar{\omega}_k$. Therefore, if representatives and centers coincide each with other, $\tilde{\omega}_k = \bar{\omega}_k$ for all clusters, then we get an unbiased clustering.

Nevertheless, if we immerse the set Ω in some feature space, we obtain in general the biased clustering, since the center $\mathbf{x}(\bar{\omega}_k)$ may not be the mean object $\bar{\mathbf{x}}_k$ in the unknown feature space.

The classical *k*-means algorithm was developed for distances and similarities in [17, 18]. Centers $\bar{\omega}_k$ provide the unbiased clustering with cluster dispersions $\sigma_k^2 = (1/N_k)\sum_{i=1}^{N_k} d^2(\omega_i, \bar{\omega}_k)$ minimizing $J(K)$. If the set Ω is immersed in a feature space, then two criteria

$$J^X(K) = \min_{\bar{\mathbf{x}}_1,\dots \bar{\mathbf{x}}_K} J(K) ,\ J^D(K) = \min_{\bar{\omega}_1,\dots \bar{\omega}_K} J(K)$$

have not the same values, since $J^D(K) \geq J^X(K)$ in general. Yet, $J^X(K) = J^D(K)$, if objects $\mathbf{x}(\bar{\omega}_k)$ and $\bar{\mathbf{x}}_k$ are the same.

We would like to guarantee this condition. For some $\omega_l \in \Omega$, as a point of the origin and a pair ω_i, ω_j, the scalar product is $s_{ij} = (d_{li}^2 + d_{lj}^2 - d_{ij}^2)/2$, where distance is $d_{pq} = d(\omega_p, \omega_q)$ and $s_{ii} = d_{li}^2$ for $i = j$. Therefore, the main diagonal of the matrix $S_l(N,N)$ represents the squared distances from the origin $\omega_l \in \Omega$ to other objects.

According to [6], it is convenient to put the origin of the feature space in the center of all objects $\omega_i \in \Omega,\ i=1,\dots N$. Therefore, we put the origin of the feature space, cluster by cluster, in each center $\bar{\omega}_k$ to represent it by its distances to all other objects in the unknown feature space (N_k is the number of objects in Ω_k, $\omega_p, \omega_q \in \Omega_k$):

$$d^2(\omega_i, \bar{\omega}_k) = \frac{1}{N_k}\sum_{p=1}^{N_k} d_{ip}^2 - \frac{1}{2N_k^2}\sum_{p=1}^{N_k}\sum_{q=1}^{N_k} d_{pq}^2 , \qquad (4)$$

where, according to (1), (4), the cluster dispersion is

$$\sigma_k^2 = \frac{1}{N_k}\sum_{i=1}^{N_k}\left(\frac{1}{N_k}\sum_{p=1}^{N_k} d_{ip}^2 - \frac{1}{2N_k^2}\sum_{p=1}^{N_k}\sum_{q=1}^{N_k} d_{pq}^2\right) = \frac{1}{2N_k^2}\sum_{p=1}^{N_k}\sum_{q=1}^{N_k} d_{pq}^2 . \qquad (5)$$

Hence, we develop the distance *k*-means algorithm based on the classical principle of the “minimum distance to a cluster center”:

(a) **Step 0.** Determine in some way K centers $\bar{\omega}_k^1$ and put them as representatives $\tilde{\omega}_k^1 = \bar{\omega}_k^1,\ k = 1,\dots K;\ \ s = 1$.

Step *s*. Reallocate all objects between clusters:

1. $\omega_i \in \Omega_k^s$, if $d(\omega_i,\ \tilde{\omega}_k^s) \le d(\omega_i,\ \tilde{\omega}_j^s)$ for $\omega_i \in \Omega_{j\ne k}^s$, $j = 1,\dots K,\ i = 1,\dots N.$
2. Recalculate centers $\bar{\omega}_k^s,\ k = 1,\dots K$, represented by distances $d(\omega_i, \bar{\omega}_k^s),\ \ i = 1,\dots N.$
3. Stop, if $\tilde{\omega}_k^s = \bar{\omega}_k^s,\ k = 1,\dots K,$ else $\tilde{\omega}_k^{s+1} = \bar{\omega}_k^s,\ \bar{\omega}_k^{s+1} = \bar{\omega}_k^s,\ k = 1,\dots K;$ $s = s+1.$

Based on the direct recalculation of the criterion (1), the equivalent realization is:

(b) **Step 0.** Determine in some way K centers $\bar{\omega}_k^1$ and put them as representatives $\tilde{\omega}_k^1 = \bar{\omega}_k^1,\ k = 1,\dots K;$ calculate $J^1 = J^1(K)$ and put $\tilde{J}^1 = \tilde{J}^1(K) = J^1$ relative to representatives; $s = 1$.

Step *s*. Reallocate all objects between clusters:

1. $\omega_i \in \Omega_k^s$, if $J_{ik}^s \le J_{ip}^s$ for $\omega_i \in \Omega_{p\ne k}^s$, $p = 1,\dots K,$ $i = 1,\dots N.$
2. Recalculate centers $\bar{\omega}_k^s,\ k = 1,\dots K$, represented by distances $d(\omega_i, \bar{\omega}_k^s),\ \ i = 1,\dots N;$ recalculate J^s.
3. Stop, if $\tilde{J}^s = J^s$, else $\tilde{J}^{s+1} = J^s,\ J^{s+1} = J^s;$ $s = s+1.$

A positive definite similarity matrix $S(N,N)$ with elements $s_{ij} = s(\omega_i, \omega_j) \ge 0$ can be obtained as a matrix of scalar products in the positive quadrant of the feature space. Relative to some point $\omega_k \in \Omega$ as the origin, with $s_{ij} = (d_{ki}^2 + d_{kj}^2 - d_{ij}^2)/2,\ \ s_{ii} = d_{ki}^2$, distances are defined as $d_{ij}^2 = s_{ii} + s_{jj} - 2s_{ij}$. The cluster center $\bar{\omega}_k$ is represented by its similarities with other objects

$$s(\omega_i, \bar{\omega}_k) = \frac{1}{N_k}\sum_{p=1}^{N_k} s_{ip}\ ,\ \ \omega_p \in \Omega_k, \omega_i \in \Omega,\ \ i = 1,\dots N\ . \ (6)$$

The cluster compactness is the mean similarity of the cluster center with respect to other objects (6):

$$\delta_k = \frac{1}{N_k}\sum_{i=1}^{N_k} s(\omega_i, \bar{\omega}_k) = \frac{1}{N_k^2}\sum_{i=1}^{N_k}\sum_{p=1}^{N_k} s_{ip}\ ;\ \omega_i, \omega_p \in \Omega_k\ .$$

The unbiased clustering minimizes the cluster dispersion σ_k^2 and maximizes the compactness δ_k according to (5):

$$\sigma_k^2 = \frac{1}{2N_k^2}\sum_{i=1}^{N_k}\sum_{j=1}^{N_k}(s_{ii} + s_{jj} - 2s_{ij}) = \frac{1}{N_k}\sum_{i=1}^{N_k} s_{ii} - \delta_k\ ,$$

and for all clusters:

$$J(K) = \sum_{k=1}^{K}\frac{N_k}{N}\sigma_k^2 = \frac{1}{N}\sum_{i=1}^{N} s_{ii} - \sum_{k=1}^{K}\frac{N_k}{N}\delta_k = C - I(K)\ .$$

For similarity clustering, we maximize compactness $I(K)$, with $I(K) = C - J(K)$. The similarity *k*-means algorithm is the analogue of algorithms (a) and (b) relative to $I(K)$.

3 The bi-partial criterion for clustering

In this paper, we develop the bi-partial objective function like (3) for the dissimilarity *k*-means

$$J_\delta(K) = (1-\alpha)J(K) + \alpha\delta(K)\ , \qquad (7)$$

so as to combine $J(K)$ for intra-cluster distances with the inter-cluster similarity $\delta(K)$. We define the inter-cluster similarity $\delta(K) = (1/K)\sum_{k=1}^{K} s(\bar{\omega}_k, \bar{\omega}_0)$ relative to the center of the whole set, being the object $\bar{\omega}_0$, represented by its similarities with respect to all other centers $s(\bar{\omega}_k, \bar{\omega}_0) = (1/K)\sum_{p=1}^{K} s(\bar{\omega}_k, \bar{\omega}_p)$; $\bar{\omega}_k, k = 1,\dots K$:

$$\delta(K) = \frac{1}{K^2}\sum_{k=1}^{K}\sum_{l=1}^{K} s(\bar{\omega}_k, \bar{\omega}_l)\ . \qquad (8)$$

Unfortunately, the bi-partial criterion $J_\delta(K)$, as defined here, does not work for the classical *k*-means (b), since (8), as the second part of $J_\delta(K)$ in (7), cannot be changed for constant centers while attempting to transfer objects in step *s*.

Therefore, for any $0 \le \alpha < 1$, the clustering results are the same as for the classical case with $\alpha = 0$. And the algorithm does not work properly with $\alpha = 1$.

We develop here the new “permutable” version of the classical *k*-means (b) without direct calculation of cluster centers. Here, the new permutable *k*-means is the meanless clustering for the classical *k*-means (b).

As we can see in (5), the cluster dispersion is half of the average of squared distances between objects in the cluster. This representation does not contain centers themselves, and we calculate the criterion (1) without centers in the form

$$J(K) = \sum_{k=1}^{K}\frac{N_k}{N}\sigma_k^2 = \frac{1}{2N}\sum_{k=1}^{K}\frac{1}{N_k}\sum_{p=1}^{N_k}\sum_{q=1}^{N_k} d_{pq}^2\ . \qquad (9)$$

Next, we would like to calculate the similarity $s(\bar{\omega}_k, \bar{\omega}_l)$ between cluster centers in (8). According to (6), the average similarity of the center $\bar{\omega}_k$ with the objects from the other cluster $\omega_i \in \Omega_l$ is

$$s(\Omega_l, \bar{\omega}_k) = \frac{1}{N_l}\sum_{i=1}^{N_l} s(\omega_i, \bar{\omega}_k) = \frac{1}{N_l N_k}\sum_{i=1}^{N_l}\sum_{p=1}^{N_k} s_{ip} \ , \ \omega_p \in \Omega_k .$$

It is evident that $s(\Omega_l, \bar{\omega}_k) = s(\Omega_k, \bar{\omega}_l)$, as $s_{ij} = s_{ji}$. Therefore, we can use the suitable notation $s(\Omega_l, \bar{\omega}_k) = s(\Omega_k, \bar{\omega}_l) = s(\Omega_l, \Omega_k) = s(\bar{\omega}_l, \bar{\omega}_k)$. Hence, (8) is converted into ($\omega_p \in \Omega_k$, $\omega_q \in \Omega_l$):

$$\delta(K) = \frac{1}{K^2}\sum_{k=1}^{K}\sum_{l=1}^{K}\frac{1}{N_k N_l}\sum_{p=1}^{N_k}\sum_{q=1}^{N_l} s_{pq}. \qquad (10)$$

The goal of $J_\delta(K)$ is to produce clusters with possibly low dispersion and possibly dissimilar centers. We note that (10) is in a way an inconsistent function, since for $k = l$ it contains the cluster compactness δ_k. Hence, we modify (10) to get the inter-cluster similarities only and take into account the symmetry

$$\delta(K) = \frac{1}{2K(K-1)}\sum_{k=1}^{K}\sum_{l=1, l\neq k}^{K}\frac{1}{N_k N_l}\sum_{p=1}^{N_k}\sum_{q=1}^{N_l} s_{pq}, \qquad (11)$$

for $\omega_p \in \Omega_k$, $\omega_q \in \Omega_l$.

We develop here the classical *k*-means (b) in the new form of the permutable *k*-means based on (9)–(11):

(c) **Step 0.** Determine in some way the sets $\Omega_l^1, l = 1, \dots K$; define α , calculate $J^1 = J_\delta^1(K)$; $s = 1$.

Step *s*. Reallocate all objects between clusters:

1. Remember, but do not move: $\omega_i \in \Omega_k^s$, if $J_{ik}^s \leq J_{ip}^s$ for $\omega_i \in \Omega_{p \neq k}^s$, $p = 1, \dots K, \ i = 1, \dots N$.
2. Reallocate all objects $\omega_i, i = 1, \dots N$ at once; calculate J^{s+1}.
3. If $J^{s+1} = J^s$ then stop;
   If $J^{s+1} > J^s$ then: cancel last reallocations, $J^{s+1} = J^s$, stop;
   If $J^{s+1} < J^s$ then: $J^{s+1} = J^s$, $s = s + 1$.

As we can see, in the step (*s*.1) we recalculate the criterion J^s in order to get its modified value J_{ip}^s. Let $\omega_i \in \Omega_j^s$. When trying to move ω_i from Ω_j^s to some other Ω_p^s , we try to change the respective sets to $\Omega_j^s \setminus \omega_i$ and to $\Omega_p^s \bigcup \omega_i$. Changes in the sets result in implicit changes of their centers, even though we do not calculate them. Consequently, this action differs from the same one in algorithms (a) and (b) for constant centers.

Algorithm (c) appears to be a new type of clustering procedures, since its result differs, in general, from those of the classical (a) and (b) procedures, both for the classical $(\alpha = 0)$ and the proper bi-partial $(\alpha > 0)$ cases. In addition, we can use some optimal initial clusters to enhance the quality of results, and optimal recalculations to improve performance of permutations.

As we can see, the algorithm (c) is the same as the classical ones (a) and (b) for the standard criterion $J(K)$ and differs (sometimes subtly and finely) from them for the bi-partial criterion $J_\delta(K)$.

It is clear that the new algorithm gives the classical result for non-intersecting clusters. Nevertheless, its result can be improved for intersecting clusters, since by means of the criterion $J_\delta(K)$ a cluster center can be shifted in some vicinity without changing the cluster itself. Such possibility depends on the gaps between real points in continuous feature space and the discrete cluster structure superimposed.

4 Redistribution of data dispersion by the bi-partial criterion

Here, we explain why by means of the criterion (7) it is possible to improve the classical clustering of *k*-means.

Consider the classical case. Let the set of size *N* be divided into *K* subsets (clusters). In our perspective, we consider balancing of total dispersion between its intra- and inter- parts. We know [19, 20] that $S_T = S_W + S_B$, where S_T is the total scatter matrix, S_W is the intra-cluster and S_B is the inter-cluster scatter matrices. Therefore, $trS_T = trS_W + trS_B$ for diagonal elements only. Since $trS_T = N\sigma_T^2$, $trS_W = N\sigma_W^2$, and $trS_B = N\sigma_B^2$, then finally $\sigma_T^2 = \sigma_W^2 + \sigma_B^2$.

Let the set $\Omega = \{\omega_1, \dots \omega_N\}$ be immersed in some metric space and represented by the distance matrix $D(N, N)$ only with elements $d_{ij} = d(\omega_i, \omega_j) \geq 0$. Let Ω be split into groups $\Omega_k, k = 1, \dots K$. Based on the Torgerson's formula, we define the following:

for single group dispersions

$$\sigma_k^2 = \frac{1}{2N_k^2}\sum_{p=1}^{N_k}\sum_{q=1}^{N_k} d^2(\omega_p, \omega_q), \ k = 1, \dots K \ ;$$

for the intra-group dispersion

$$\sigma_W^2 = \sum_{k=1}^{K}\frac{N_k}{N}\sigma_k^2 = \sum_{k=1}^{K}\frac{N_k}{N}\frac{1}{2N_k^2}\sum_{p=1}^{N_k}\sum_{q=1}^{N_k} d^2(\omega_p, \omega_q) = \frac{1}{2N}\sum_{k=1}^{K}\frac{1}{N_k}\sum_{p=1}^{N_k}\sum_{q=1}^{N_k} d^2(\omega_p, \omega_q) ;$$

for the total dispersion

$$\sigma_T^2 = \frac{1}{2N^2}\sum_{p=1}^{N}\sum_{q=1}^{N} d^2(\omega_p, \omega_q) = \frac{1}{2N^2}\sum_{k=1}^{K}\sum_{l=1}^{K}\sum_{p=1}^{N_k}\sum_{q=1}^{N_l} d^2(\omega_p, \omega_q) ;$$

for the inter-center dispersion

$$\sigma_{IC}^2 = \frac{1}{K}\sum_{k=1}^{K} d^2(\bar{\omega}_k, \bar{\omega}_0) = \frac{1}{2K^2}\sum_{p=1}^{K}\sum_{q=1}^{K} d^2(\bar{\omega}_p, \bar{\omega}_q) ,$$

where the center $\bar{\omega}_0$ of the set Ω is represented by its distances to other centers $\bar{\omega}_k$ through

$$d^2(\bar{\omega}_k, \bar{\omega}_0) = \frac{1}{K}\sum_{p=1}^{K} d^2(\bar{\omega}_k, \bar{\omega}_p) - \sigma_{IC}^2 .$$

We remark that the classical inter-group dispersion is not given by the Torgerson's formula

$$\sigma_B^2 = \sum_{k=1}^{K} \frac{N_k}{N} d^2(\bar{\omega}_k, \bar{\omega}_0) .$$

Therefore, the classical inter-group dispersion is

$$\sigma_B^2 = \sum_{k=1}^{K} \frac{N_k}{N} \left(\frac{1}{K} \sum_{p=1}^{K} d^2(\bar{\omega}_k, \bar{\omega}_p) - \sigma_{IC}^2 \right) =$$

$$\frac{1}{K} \sum_{k=1}^{K} \frac{N_k}{N} \sum_{p=1}^{K} d^2(\bar{\omega}_k, \bar{\omega}_p) - \sigma_{IC}^2 \sum_{p=1}^{K} \frac{N_k}{N} =$$

$$\frac{1}{K} \sum_{k=1}^{K} \frac{N_k}{N} \sum_{p=1}^{K} d^2(\bar{\omega}_k, \bar{\omega}_p) - \sigma_{IC}^2 .$$

As shown above, we minimize the classical criterion $J(K)$ based on the distance matrix $D(N,N)$, and maximize the criterion in the dual form $I(K) = C - J(K)$ based on the similarity matrix $S(N,N)$.

Hence, in the dual form of the bi-partial criterion we try to maximize the classical part $I(K)$ and the new second part for the inter-center dispersion σ_{IC}^2, as based on the Torgerson's formula. Since the classical inter-group dispersion σ_B^2 is not based on the Torgerson's formula, we calculate it with distances $d^2(\bar{\omega}_k, \bar{\omega}_0)$. Such distances refer to distances between sets, not being a topic here.

Hence, in the dual form by maximizing $I(K)$, we minimize strictly equivalent classical $J(K)$ and maximize the inter-center dispersion σ_{IC}^2. Since $\sigma_T^2 = \sigma_W^2 + \sigma_B^2$, we have the decomposition

$$\sigma_T^2 = \sigma_W^2 + \frac{1}{K} \sum_{k=1}^{K} \frac{N_k}{N} \sum_{p=1}^{K} d^2(\bar{\omega}_k, \bar{\omega}_p) - \sigma_{IC}^2 .$$

Let us denote $\sigma_{B \cup IC}^2 = \frac{1}{K} \sum_{k=1}^{K} \frac{N_k}{N} \sum_{p=1}^{K} d^2(\bar{\omega}_k, \bar{\omega}_p)$ and represent the classical inter-group dispersion in the form $\sigma_B^2 = \sigma_{B \cup IC}^2 - \sigma_{IC}^2$ without the contribution of the inter-center dispersion, where $\sigma_T^2 + \sigma_{IC}^2 = \sigma_W^2 + \sigma_{B \cup IC}^2$.

As we can see, the permutable k-means is targeted to minimize $J(K) = \sigma_W^2$. Since the total dispersion $\sigma_T^2 = const$, at the same time the classical inter-group dispersion $\sigma_B^2 = \sigma_{B \cup IC}^2 - \sigma_{IC}^2$ is maximized. Therefore, the balance $\sigma_T^2 = \sigma_W^2 + \sigma_B^2$ remains true. The decomposition $\sigma_T^2 + \sigma_{IC}^2 = \sigma_W^2 + \sigma_{B \cup IC}^2$ shows that the balance of two parts is maintained, while we increase both of them.

In this case, the bi-partial criterion influences σ_{IC}^2 only. Hence, by means of the bi-partial criterion we manipulate to maximize the inter-center dispersion σ_{IC}^2 with the other part $\sigma_{B \cup IC}^2$ being maximized "as is".

5 Experiments

5.1 Experimental setup

Experimental data are the original Fisher's *Iris data* [21]. We chose this data set as a simple illustration for the basic properties of the approach developed. Such data consist of 150 measurements of 50 plants, belonging to three varieties: *Iris setosa*, *Iris versicolor*, and *Iris virginica*. Four flower measurements are made: petal length and width, and sepal length and width.

It is known that the 1st class (*Iris setosa*) is well separated from other two classes (2nd class, *Iris versicolor*, and 3rd class, *Iris virginica*). The 2nd and 3rd classes intersect each other. Another peculiarity of *Iris data* is the coincidence of objects 102 and 143 from the 3rd class. *Iris data* are also included in Matlab.

There are also other available variants of the *Iris data*, differing from the classic set of [21]. Such differences usually concern corrections in some measurements.

Since the classification of data has been defined, we show that the bi-partial objective function $J_\delta(K)$, developed above, allows us to improve the classical clustering result. According to it, we separate as usual 1st class correctly from two others, and decrease the errors in separation of the 2nd and the 3rd class.

According to the formulation above, we investigate the problem

$$\alpha^* = \underset{0 \le \alpha \le 1}{\arg\min}\, J_\delta(K) = \underset{0 \le \alpha \le 1}{\arg\min} \left((1-\alpha) J(K) + \alpha \delta(K) \right) .$$

As we can see, this formulation implies balancing of two parts of the criterion. Therefore, it would be good to measure $J(K)$ and $\delta(K)$ on the same scale.

The dispersion of standardized data is *n*, i.e. the number of features (*n*=4 for *Iris data*), and usually more than *n* for original (non-standardized) data. The clustering results for original and standardized data can differ.

In order to get rid of the potential scale bias, we normalize inter-cluster similarities $s'_{kl} = s_{kl} / \sqrt{s_{kk} s_{ll}}$ to get $s'_{kk} = 1,\ 0 \le s'_{kl} \le 1;\ k, l = 1, \dots K$.

The last technical remark regarding the correctness of the criterion $J_\delta(K)$ is that in the case of usual standard multidimensional data, we need to move the origin out of the convex cover of the set relative to its center and provide positive scalar products as similarities between objects. This problem was discussed in [22].

Indeed, as it is mentioned above, all similarities in (6), (8), (10), (11) must be nonnegative for correct $I(K)$ and $\delta(K)$. According to (4), the origin is placed in the center of the data set in the feature space.

Unfortunately, it this case we cannot use scalar products $s_{ij} = (d_{ki}^2 + d_{kj}^2 - d_{ij}^2)/2$ in $J_\delta(K)$, since they can have negative values. Nevertheless, scalar products change to nonnegative values with respect to the origin placed out of the convex cover of the set, since all of them appear to be in the positive quadrant of the feature space. Hence, it does not matter at all for distances (they have been calculated and not changed for any place of the origin), but it is correct to represent nonnegative similarities by scalar products.

It is known that the *k*-means algorithm is the locally optimal procedure with results dependent on initial decisions (partition or choice of centers).

For all classes, we test three initial partitions: 50/50/50 (plant varieties as classes), 50/70/30 (20 plants from the 3rd class are wrongly placed in the 2nd class), 50/30/70 (20 plants from the 2nd class are wrongly placed in the 3rd class).

For just two intersecting classes (2nd and 3rd) we test also three initial partitions: 50/50 (plant varieties as classes), 70/30 (20 plants from the 3rd class are wrongly placed in the 2nd class), 30/70 (20 plants from the 2nd class are wrongly placed in the 3rd class).

In yet another case we investigate two classes of the entire set, organized as the small one (1st class) and the big one (2nd and 3rd classes). We test three initial partitions: 50/100 (plants from the 1st class versus all plants from the 2nd and 3rd classes together), 100/50 (all plants from the 1st and the 2nd classes together versus plants from the 3rd class), 30/120 (only first 30 plants from the 1st class versus all others).

In all experiments, we first get the classical result with $\alpha = 0$, starting from the predefined initial partitions as above. Second, starting, as well, from the predefined initial partitions characterized above, we vary the parameter $0 < \alpha \le 1$ with increment 0.01 to find the optimal α^* among the tested 100 points.

5.2 Results and discussion

In the first experiment with original *Iris data* for all initial partitions for three classes, we correctly separate the 1st class and decrease errors in separation of intersecting 2nd and 3rd classes (Table 1, Fig.1). For two intersecting classes only, we decrease errors in the separation of the 2nd and 3rd classes, too (Table 1, Fig. 2, 3). It can be seen that the optimal intervals for α^* depend on the number of clusters (Table 1), hence on data dispersion, and can slightly differ for different initial partitions. Error diagrams are not monotonic functions (Fig. 1–3).

As we can see, original *Iris data* are some sort of “well structured” data, since for different initial partitions we get the same 16 misclassified objects for the classical $(\alpha = 0)$ criterion and the same 15 misclassified objects for the bi-partial (α^*) criterion (Table 1). For the classical criterion, misclassified objects are generally from the 3rd class (Table 2). The object 135 is well classified and shown here, since it is misclassified for the bi-partial criterion.

Misclassified objects for the bi-partial criterion are from the 3rd class, too (Table 3). Here, objects 53 and 78 are well classified, and the object 135 is misclassified.

We repeat this experiment for standardized data (Table 4). Such data are more complicated. As we know, *Iris* classes are not so spherical ones in the original feature space, and that is why the *k*-means type of approach is not the best suited for them.

After data standardization, classes appear to be more spherical and contain more “mixed” objects from intersecting classes, usually giving more misclassifications in the classical case (Table 4).

Hence, for the classical criterion $(\alpha = 0)$ for standardized data, 25 misclassified objects are from two intersecting classes, 2nd and 3rd, with well classified all objects from the 1st class (Table 5). Objects 104, 109, 112, 126, 129 are well classified and shown too, since they are misclassified for the bi-partial criterion.

Misclassified objects for the bi-partial criterion are mainly from the 3rd class again (Table 6). Here, objects 52, 57, 66, 71, 76, 86, 87 are well classified and objects 104, 109, 112, 128, 129 are misclassified.

Table 1: Clustering results of original *Iris data.*

Initial partitions	Errors $(\alpha = 0)$	α^*	Errors (α^*)	Diagrams
50/50/50	16	0.6 ÷ 0.75	15	Fig. 1
50/70/30	16	0.6 ÷ 0.75	15	
50/30/70	16	0.6 ÷ 0.75	15	
50/50	16	0.81 ÷ 0.92	15	Fig. 2
70/30	16	0.81 ÷ 0.92	15	
30/70	16	0.81 ÷ 0.91	15	Fig. 3

Table 2: Classical 16 misclassifications of original *Iris data.*

$\alpha = 0$ 50/50/50 50/50 50/70/30 70/30 50/30/70 30/70	2nd cluster	3rd cluster
Iris versicolor 2nd class (51-100)		53 78

Iris virginica
3rd class (101-150)
	102 120 128 147
	107 122 134 150
	114 124 139
	115 127 143
		Correct: 135

Table 3: Bi-partial 15 misclassifications of original *Iris data.*

α^* 50/50/50 50/50 50/70/30 70/30 50/30/70 30/70	2nd cluster	3rd cluster
Iris versicolor 2nd class (51-100)	**53** **78**	
Iris virginica 3rd class (101-150)	102 120 128 147 107 122 134 150 114 124 139 115 127 143 **135**	

Table 4: Clustering results of standardized *Iris data.*

Initial partitions	Errors $(\alpha = 0)$	α^*	Errors (α^*)	Diagrams
50/50/50	25	0.85	22	Fig. 4
50/70/30	25	0.85	22	
50/30/70	25	0.85	22	
50/50	17	0.94 ÷ 0.97	15	Fig. 5
70/30	17	0.92 ÷ 0.97	15	Fig. 6
30/70	17	0.83 ÷ 0.95	14	Fig. 7

Table 5: Classical 25 misclassifications of standardized *Iris data.*

$\alpha = 0$ 50/50/50 50/70/30 50/30/70	2nd cluster	3rd cluster

Iris versicolor
2nd class (51-100)
		51 57 76 86
		52 66 77 87
		53 71 78

Iris virginica
3rd class (101-150)
	102 120 134 147
	107 122 135 150
	114 124 139
	115 127 143
		Correct:
		104 129
		109
		112
		128

Table 6: Bi-partial 22 misclassifications of standardized *Iris data.*

$\alpha *$ 50/50/50 50/70/30 50/30/70	2nd cluster	3rd cluster
Iris versicolor 2nd class (51-100)	**52 71 86** **57 76 87** **66 77**	51 53 78
Iris virginica 3rd class (101-150)	102 122 139 **104** 107 124 143 **109** 114 127 147 **112** 115 134 150 **128** 120 135 **129**	

Table 7: Classical 17 misclassifications of standardized *Iris data.*

$\alpha = 0$		2nd cluster	3rd cluster
50/50	*Iris versicolor* 2nd class (51-100)		51 53 78
	Iris virginica 3rd class (101-150)	102 122 134 147 107 124 135 150 114 127 139 120 128 143	Correct: 112
70/30	*Iris versicolor* 2nd class (51-100)		51 53 78
	Iris virginica 3rd class (101-150)	102 122 134 147 107 124 135 150 114 127 139 120 128 143	Correct: 112
30/70	*Iris versicolor* 2nd class (51-100)	Correct: 52 71 87 57 77 66 86	51 53 78
	Iris virginica 3rd class (101-150)	102 122 134 147 107 124 135 150 114 127 139 120 128 143	

Table 8: Bi-partial misclassifications of standardized *Iris data.*

$\alpha *$		2nd cluster	3rd cluster
50/50	*Iris versicolor* 2nd class (51-100)	**51** **53** **78**	
	Iris virginica 3rd class (101-150)	102 122 134 147 107 124 135 150 114 127 139 120 128 143 **112**	
70/30	*Iris versicolor* 2nd class (51-100)	**51** **53**	78
	Iris virginica 3rd class (101-150)	102 122 134 147 107 124 135 150 114 127 139 120 128 143	
30/70	*Iris versicolor* 2nd class (51-100)	52 71 87 57 77 66 86	51 **52 71 87** 53 **57 77** 78 **66 86**
	Iris virginica 3rd class (101-150)	107 114 120 135	102 128 147 122 134 150 124 139 127 143

For two intersecting classes of standardized data and for the classical $(\alpha = 0)$ criterion (Table 7), we get the same 3 misclassified objects from the 2nd class and 14 misclassified objects from the 3rd class.

We get different misclassified objects (Table 8) for different initial partitions in the bi-partial case (15 objects for the 50/50 and 70/30 initial partitions, 14 objects for the 30/70 initial partition).

For standardized data, we usually get different results for three and two classes relative to original data. As we can see, the best result with the minimum of 14 errors for the initial partition 30/70 differs in terms of objects from the results for other initial partitions (Table 8).

Even though standardization is a usual step in data processing, we can see that the clustering results for standardized *Iris data* are not so "natural" as for the original ones. This is the well known and unwanted effect of standardization.

Clustering results for *Iris data* by both classical and by bi-partial criteria are more "natural" for original data than for standardized data.

In the second experiment, we investigate the already mentioned general defect of the criterion (1). As it is well known, the classical *k*-means clustering tries to get clusters, which are approximately equal by size.

In case of classes that differ as to their sizes, the new permutable algorithm decreases usually the size of the bigger class (2nd and 3rd together) and increases the size of the smaller class (1st).

This is the classical result for $\alpha = 0$ with three errors for original *Iris data* (objects 58, 94, 99 were misclassified to the 1st class).

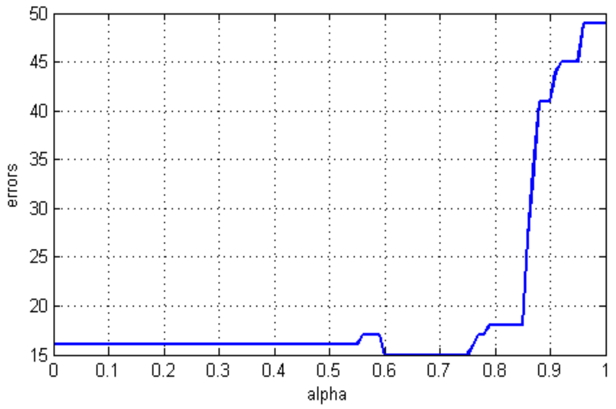

Figure 1. Clustering errors of original *Iris data* for Setosa/Versicolor/Virginica varieties (50/50/50, 50/70/30, 50/30/70) with 15 misclassified objects.

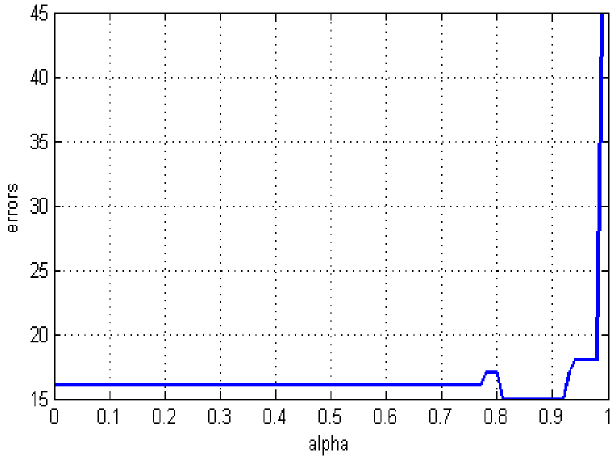

Figure 2. Clustering errors of original *Iris data* for Versicolor/Virginica varieties (50/50, 70/30) with 15 misclassified objects.

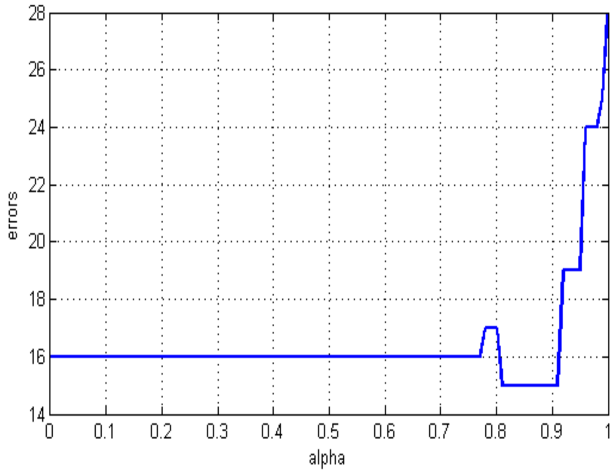

Figure 3. Clustering errors of original *Iris data* for Versicolor/Virginica varieties (30/70) with 15 misclassified objects.

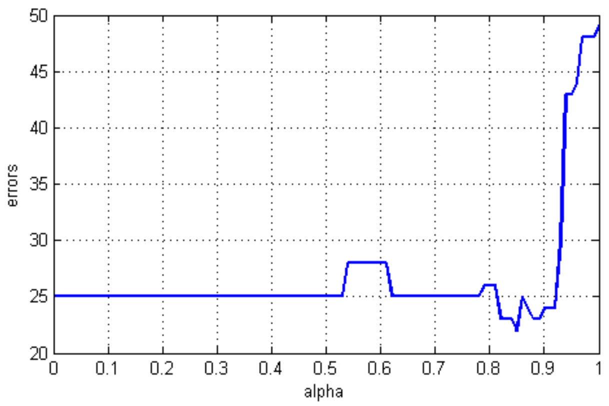

Figure 4. Clustering errors of standardized *Iris data* for Setosa/Versicolor/Virginica varieties (50/50/50, 50/70/30, 50/30/70) with 22 misclassified objects.

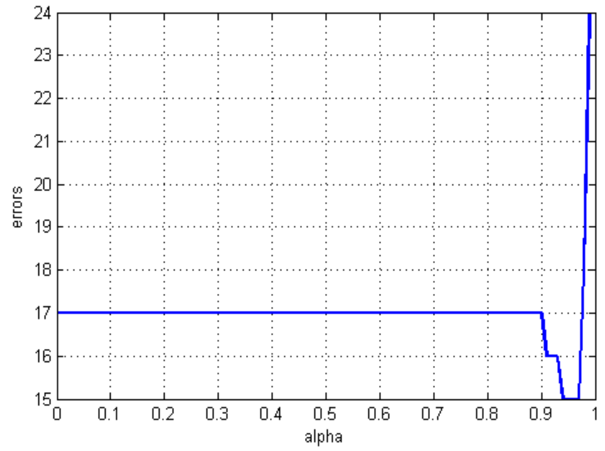

Figure 5. Clustering errors of standardized *Iris data* for Versicolor/Virginica varieties (50/50) with 15 misclassified objects.

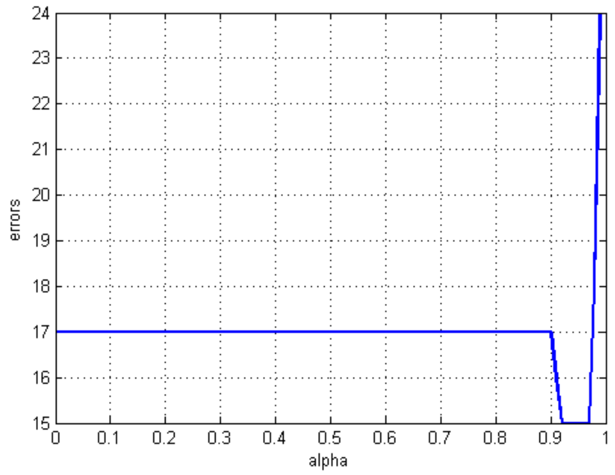

Figure 6. Clustering errors of standardized *Iris data* for Versicolor/Virginica varieties (70/30) with 15 misclassified objects.

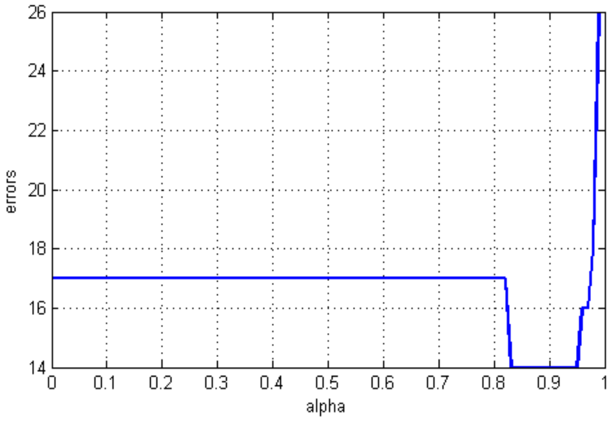

Figure 7. Clustering errors of standardized *Iris data* for Versicolor/Virginica varieties (30/70) with 14 misclassified objects.

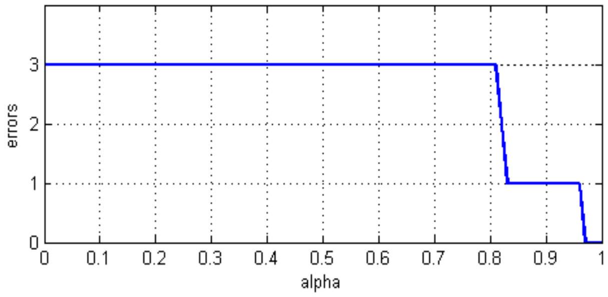

Figure 8: Clustering errors of original *Iris data* for Setosa versus Versicolor/Virginica varieties (50/100, 100/50, 30/120).

We reduce errors to zero (Fig. 8) and correctly separate the smaller 1st class from the bigger one (2nd and 3rd) in the optimal interval $0.97 \le \alpha^* \le 1$ for all initial partitions, i.e. 50/100, 100/50, 30/120. For standardized data, the result contains no errors at all for the whole interval $0 \le \alpha \le 1$ for all initial partitions.

6 Conclusion

The *k*-means procedure is very popular in machine learning and data mining fields. This procedure is very natural and understanding its principles and results is easy. Additionally, this procedure is deeply connected with other ideas, like the EM-algorithm, SOMs, etc.

On the other hand, the use of the bi-partial criterion can improve the classical clustering result. The bi-partial objective function consists of two parts, the first one supporting the best approximation of individual categories, and the second one supporting the appropriate separation among the categories. In the case of the *k*-means algorithm, the bi-partial objective function combines intra-cluster dispersions with the inter-cluster similarity, to be jointly minimized. In dual form, the bi-partial objective function combines cluster concentrations with the inter-cluster dispersion, to be maximized.

In this paper, we investigate the direct form of the bi-partial criterion function. The first part of this criterion provides the classical quality measure of *k*-means clustering, based on distances between objects.

As it is shown in this paper, the bi-partial criterion does not work directly through the standard procedure of the classical *k*-means, since the second part of the criterion cannot be changed within the classical procedure.

Therefore, to improve the clustering quality based on the bi-partial criterion, we develop here the new permutable version of the classical *k*-means algorithm.

As it is shown in this paper, the permutable *k*-means appears to be a new type of clustering procedures.

The permutable *k*-means uses distances and similarities only. Therefore, it does not need to use the feature-based representation of experimental data. To reduce the computational complexity of permutations we can use in further work the optimal iterative techniques.

It is easy to show that in the dual form the bi-partial objective function combines cluster concentrations with the inter-cluster dispersion, to be jointly maximized. The first part of both bi-partial objective functions provides the “standard” quality of clustering based on distances between objects (the classical *k*-means) or similarities between them in dual form (the similarity *k*-means).

As a result, what the algorithm have we built? It is clear, that we have merely shown the principle of developing a class of criteria and corresponding algorithms. As we can see in Figs. 1–7, error lines are not convex functions of α in general. The future study should, then, be oriented at defining conditions for convexity, on the one hand, and developing effective algorithms of extrema finding of the similar functions, on the other.

Acknowledgements

The work of the first author was supported by Russian Foundation for Basic Research under grant 17-07-00319.

References

[1] H. Steinhaus (1956). Sur la division des corps matériels en parties. *Bulletin de l'Academie Polonaise des Sciences* IV (C1.III), 801-804 (in French).

[2] M.I. Shlezinger (1965). Spontaneous discrimination of patterns. In: *Reading Automata*. Naukova Dumka, Kiev (in Russian).

[3] M.I. Shlezinger (1968). The interaction of learning and self-organization in pattern recognition. In: Kibernetika, 4(2), 81-88. http://irtc.org.ua/image/Files/Schles/non-supervised.pdf

[4] A.V. Milen'kii (1975). *Classification of signals in conditions of uncertainty*. Moscow, Soviet Radio (in Russian).

[5] E. Diday et al. (1979). *Optimisation en classification automatique*. INRIA, Domaine de Voluceau, Rocquencourt B.P. 105, 78150 Le Chesnay (in French).

[6] W.S. Torgerson (1958). *Theory and Methods of Scaling*. N.Y., Wiley.

[7] H. P. Friedman and J. Rubin (1967). On Some Invariant Criteria for Grouping Data. In: *J. of the American Statistical Association*, 62(320):1159-1178. https://doi.org/10.1080/01621459.1967.10500923

[8] H. Späth (1983). *Cluster-formation und -analyse: Theorie, FORTRAN-Programme und Beispiele*. R. Oldenbourg-Verlag, München — Wien.

[9] S.A. Aivazyan, et al. (1989). *Applied Statistics. Classification and reduction of dimensionality (Ch. 5. Basic concepts and definitions used in classification without training. 5.4. Classification quality functionals and extremal approach to cluster analysis problems)*. Finansy i statistika, Moscow (in Russian)

[10] H.-J. Mucha, U. Simon, R. Brüggemann (2002). *Model-based Cluster Analysis Applied to Flow Cytometry Data of Phytoplankton*. Tech. Report, Berlin. http://www.wias-berlin.de/techreport/5/ wias_technicalreports_5.pdf

[11] E. Pekalska, R.P.W. Duin (2005). *The Dissimilarity Representation for Pattern Recognition. Foundations and Applications*. W.S. Singapore.

[12] A.W.F. Edwards, L.L. Cavalli-Sforza (1965). A Method for Cluster Analysis. In: *Biometrics, 21*, 362–375. https://www.jstor.org/stable/ 2528096 ?seq=1#page_scan_tab_contents

[13] Jan W. Owsinski (2012). On the optimal division of an empirical distribution (and some related problems). In: *Przegląd Statystyczny*, *special issue*, 1, 109-122.

[14] Jan W. Owsinski (2013). On dividing an empirical distribution into optimal segments. http://new.sis-statistica.org/wp-content/uploads/ 2013/09/RS12-On-Dividing-an-Empirical-Distribution-into.pdf

[15] Jan W. Owsinski (2011). The bi-partial approach in clustering and ordering: the model and the algorithms. In: *Statistica & Applicazioni*. *Special Issue*, 43–59.

[16] Jan W. Owsinski (1990). On a new naturally indexed quick clustering method with a global objective function. In: *Applied Stochastic Models and Data Analysis*, 6(3), 157-171. https://doi.org/10.1002/asm.3150060303

[17] S.D. Dvoenko (2009). Clustering and separating of a set of members in terms of mutual distances and similarities. In: *Transactions on MLDM*. IBaI Publishing 2, 2 (Oct. 2009), 80-99.

[18] S. Dvoenko (2014). Meanless *k*-means as *k*-meanless clustering with the bi-partial approach. In: *Proc. of 12th Int. Conf. on Pattern Recognition and Image Processing (PRIP'2014).* UIIP NASB, Minsk, Belarus, 50-54.

[19] R.O. Duda, P.E. Hart (1973). *Pattern Classification and Scene Analysis*. N.Y., Wiley.

[20] R.O. Duda, P.E. Hart, D.G.Stork (2000). *Pattern Classification*. Wiley-Interscience New York, NY.

[21] R.A. Fisher (1936). The use of multiple measurements in taxonomic problems. In *Ann. Eugenics*. 7, 2 (Sept. 1936), 179-188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

[22] S. D. Dvoenko, D.O. Pshenichny (2016). A recovering of violated metric in machine learning. In: *Proceedings of 8th Int. Symposium on Information and Communication Technology (SoICT'2016)*. ACM NY, 15-21. https://doi.org/10.1145/3011077.3011084