AGGREGATION AND SOFT CLUSTERING OF INFORMETRIC DATA

Anna Cena
Systems Research Institute,
Polish Academy of Sciences
ul. Newelska 6, 01-447 Warsaw, Poland
cena@ibspan.waw.pl

Marek Gagolewski
Systems Research Institute,
Polish Academy of Sciences
ul. Newelska 6, 01-447 Warsaw, Poland
Faculty of Mathematics and Information Science,
Warsaw University of Technology
ul. Koszykowa 75, 00-661 Warsaw, Poland

Summary

The aim of this contribution is to inspect possible applications of clustering techniques computed over a set consisting of nonincreasingly ordered vectors of possibly nonconforming lengths. Such data sets appear in the field of informetrics, where one may need to evaluate the quality of information items, e.g research papers, and their producers. In this paper we investigate the notion of cluster centers as an aggregated representation of all vectors from a given cluster and analyze them by means of aggregation operators.

Keywords: clustering, fuzzy clustering, c-means algorithm, distance, producers assessment problem

1 INTRODUCTION

The Producers Assessment Problem (PAP, see e.g. [7]) concerns the evaluation of a set of information resources producers according to both number and quality of their products (e.g. forum posts, research publications, etc.). More formally, this problem may be modeled by a set of vectors $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(l)}\}$, where $\mathbf{x}^{(i)} = (x_1^i, x_2^i, \dots, x_{n_i}^i)$, $x_1^i \geq x_2^i \geq \dots \geq x_{n_i}^i$ with possibly $n_i \neq n_j$ for some $i \neq j$. Please note that in this model $\mathbf{x}^{(i)}$ represents the state of the i-th producer and x_j^i denotes the quality assessment of his/her j-th top product. Moreover, in many real life applications, it is necessary to assume that $x_j^i \in \mathbb{I} = [-\infty, \infty]$, cf. [3] for discussion.

Usually, aggregation operators are most often used to summarize informetric data sets. However, also machine learning techniques may be applied for this very purpose. For example, in [13] some algorithms were applied on several indicators in order to obtain an automatic categorization of universities. Similarly, in [4] a data set on the 500 best world universities was divided into groups according to their various bibliometric performance indicators. Moreover, Costas, van Leeuwen, and Bordons in order to split a group of scientists into 3 clusters (top, medium, low class ones) used e.g. the h-index [9], number of publications, number of highly cited papers, median impact factor, etc, see [5].

Investigation carried out in this paper focus on clustering techniques. In our previous work we discussed problems and challenges one may encounter while dealing with clustering tasks on informetric data sets, see [3]. We also proposed modifications of the well-known metrics, so they can be calculated over vectors of nonconforming lengths. Obtained measures were then applied in a hierarchical clustering method. Moreover, the notion of such measures allows to adapt k-means algorithm for such task, see [2]. In this paper we are going to generalize the obtained results. Moreover, we focus on centroids of derived clusters, which can be conceived as an aggregated representation of the data set.

The structure of this contribution is as follows. In the next section the definition of a metric and dissimilarity measure for vectors of nonconforming lengths is recalled. In Sec. 3 the notion of the c-means algorithm is generalized so it can be computed over PAP data sets. Next, in Sec. 4, the performance of the obtained method is investigated. Finally, Sec. 5 concludes the paper and indicates future research directions.

2 METRICS

For any $n \in \mathbb{N}$, let $\mathcal{S}_n$ denote the set of nonincreasingly ordered real vectors of length n, i.e. $\mathcal{S}_n = \{(x_1, \dots, x_n) \in \mathbb{R}^n, x_1 \geq \dots \geq x_n\}$. Moreover, let $\mathcal{S}_{\leq n}$ be a set of nonincreasingly ordered vectors of length at most n, that is $\mathcal{S}_{\leq n} = \bigcup_{i=1}^{n} \mathcal{S}_i$. Assume that we are given l producers and $k = \max\{n_i : i = 1, 2, \ldots, l\}$. Obviously, such k is finite and well defined for each set of producers. Moreover, let $\tilde{\mathbf{x}}$ denote the vector of length k and equivalent to $\mathbf{x}$ padded with 0's, i.e. $\tilde{\mathbf{x}} = (x_1, x_2, \ldots, x_n, 0, \ldots, 0) \in \mathcal{S}_k$.

Let us now recall the definition of a class of metrics over $\mathcal{S}_{\le k}$ (see [3] for more details and a proof). Please keep in mind, that *metric* is a function $d(\mathbf{x}, \mathbf{y})$ such that $(\forall \mathbf{x}, \mathbf{y})$ (a) $d(\mathbf{x}, \mathbf{y}) = 0 \Leftrightarrow \mathbf{x} = \mathbf{y}$ and (b) $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ and fulfills triangle inequality, i.e. $d(\mathbf{x}, \mathbf{y}) \le d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y})$. Moreover, in case when only conditions (b) and (c) hold, $d(\mathbf{x}, \mathbf{y})$ is a *pseudometrc.*

Theorem 1. *Let $d_M : \mathcal{S}_{\le k} \times \mathcal{S}_{\le k} \to [0, \infty)$ be such that $d_M(\mathbf{x}, \mathbf{y}) = \mu(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) + \nu(\mathbf{x}, \mathbf{y})$, where μ is a metric on $\mathbb{R}^k$ and ν is a pseudo-metric on $\mathcal{S}_{\le k}$. Then d_M is a metric on $\mathcal{S}_{\le k}$ if and only if for all $\mathbf{x}, \mathbf{y}$ such that $\tilde{\mathbf{x}} = \tilde{\mathbf{y}}$ it holds $\nu(\mathbf{x}, \mathbf{y}) = 0 \Longrightarrow n_x = n_y$.*

Because of computational reasons, in some clustering tasks it is more convenient to consider dissimilarity measure instead of metric, i.e. a function $d(\mathbf{x}, \mathbf{y})$ such that $(\forall \mathbf{x}, \mathbf{y})$ (a) $d(\mathbf{x}, \mathbf{y}) \ge 0$, (b) $d(\mathbf{x}, \mathbf{y}) = 0 \Leftrightarrow \mathbf{x} = \mathbf{y}$ and (c) $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$. Of course, each metric is a dissimilarity measure.

3 CLUSTERING

Cluster analysis is an machine learning technique which allows to partition the general population of objects into distinguishable – according to some criteria – groups (clusters), i.e. in the most desired partition entities within each group are similar and objects in distinct groups differ as much as possible from each other (see e.g. [8]).

One approach of cluster analysis is to assume that objects are divided into c nonempty pairwise disjoint sets $\mathcal{C} = \{C_1, C_2, \ldots, C_c\}$, $\bigcup_{i=1}^{c} C_i = \mathcal{X}$, which minimizes the within cluster dissimilarity – sum of dissimilarities between points in the same cluster, i.e.:

$$\mathcal{C} = \mathop{\arg\min}_{\text{partition } \mathcal{C} \text{ of } \mathcal{X}} \sum_{j=1}^{c} \sum_{\mathbf{x} \in C_j} d_{L_2}^2 \left(\mathbf{x}, \boldsymbol{\mu}^{(j)} \right), \qquad (1)$$

where $d_{L_2}^2$ denotes the squared Euclidean distance (a dissimilarity measure) and $\boldsymbol{\mu}^{(j)}$ is the j-th cluster centroid.

On the other hand, in fuzzy clustering, every point has a degree of belonging to each cluster, rather than belonging entirely to just one of them, see e.g. [10]. In such a case, given a set of observations $\mathcal{X} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(l)}\}$, where each $\mathbf{x}^{(i)} \in \mathbb{R}^n$, we aim to determine a fuzzy pseudopartition – a family of fuzzy subsets of $\mathcal{X}$: $\mathcal{W} = \{W_1, \ldots, W_c\}$, where $W_j = W_j(\mathbf{x}^{(i)})$, such that $\sum_{j=1}^{c} W_j(\mathbf{x}^{(i)}) = 1$, describes the degree of belonging of the i-th observation to the j-th cluster. Here, we aim to find the fuzzy pseudopartitioning which minimizes the weighted within cluster dissimilarity, i.e.

$$\mathop{\arg\min}_{\text{fuzzy partition } \mathcal{W} \text{ of } \mathcal{X}} \sum_{i=1}^{l} \sum_{j=1}^{c} W_j(\mathbf{x}^{(i)})^m d_{L_2}^2(\mathbf{x}^{(i)}, \boldsymbol{\mu}^{(j)}). \qquad (2)$$

where the $m \in R$, $m \ge 1$ is a fuzzifier, cf [1].

Fuzzy c-means algorithm. Clustering tasks can be solved using various heuristics, that differ significantly in their structure. Moreover, even when an algorithm converges, the obtained minimum may only be a local minimum. Also the initial choice of weights can have a great impact on the results. Investigation carried out here concerns the fuzzy c-means algorithm (cf. [10, 1]), which may be viewed as a weighted generalization of the k-means procedure.

Let us now recall basic steps of the c-means method.

1. Set a number of clusters and randomly assign for each observation the degree of memberships.
2. Until the convergence condition is met, i.e. the coefficients' change between two iterations is not greater than the given sensitivity threshold ε, repeat:
   (a) Compute the centroid for each cluster $\boldsymbol{\mu}^{(j)}$ with respect to the weighted distance.
   (b) For each point $\mathbf{x}^{(i)}$, compute its coefficients (weights, membership) w_{ij} of being in the clusters $j = 1, \ldots, c$ as $w_{ij} = \left(\sum_{u=1}^{c} \left(\frac{d_{L_2}^2(\mathbf{x}^{(i)} - \boldsymbol{\mu}^{(j)})}{d_{L_2}^2(\mathbf{x}^{(i)} - \boldsymbol{\mu}^{(u)})} \right)^{1/(m-1)} \right)^{-1}$

Special attention should be paid when implementing the above algorithm. The main problems here may be caused by numerical errors, which can occur during the computation of weights. Please note that if $d_{L_2}^2(\mathbf{x}^{(i)} - \boldsymbol{\mu}^{(u)}) = 0$ for some $u = 1, \ldots, c$, according to the formula given above there is division by 0. In such a case, one may choose as a weight an arbitrary real number (keeping in mind that the weights must add up to 1, see e.g [10]).

3.1 DETERMINING THE WEIGHTED CENTROID

Let us focus on a squared version of an Euclidean-like metric $d_{D;pq}^2 : \mathcal{S}_{\le k} \times \mathcal{S}_{\le k}$, defined as $d_{D;pq}^2(\mathbf{x}, \mathbf{y}) = d_{L_2}^2(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) + p|n_x^r - n_y^r|$, where d_{L_2} denotes the Euclidean metric in $\mathbb{R}^k$.

Please note that this approach bases on the idea of padding input vectors with zeros. It is because in the standard arithmetic of real numbers we have $|a-0| = |0-a| = |a|$. Also note that by the fact that 0 is a distinguished value in the set of reals, the introduced metrics can be rewritten as: the distance of $\min\{n_x, n_y\}$ observations plus a norm of the remaining observations in the longer vector (which is the same as the distance to $\mathbf{0}$) plus some penalty for the difference in vectors' lengths. This provides an appealing interpretation of the proposed solution.

Let $\mathbf{w} = (w_1, w_2, \dots, w_l)$ denote the degree of membership of the observations to a given cluster. Our task is to find the vector $\boldsymbol{\mu}$ which minimizes

$$\arg\min_{\boldsymbol{\mu}\in\mathcal{S}} \sum_{i=1}^{l} w_i^m d^2_{D;pq}(\mathbf{x}^i, \boldsymbol{\mu}),$$

i.e. the vector which minimizes the objective function given by $F_{\mathbf{w}}(\boldsymbol{\mu}) = \sum_{i=1}^{l} w_i^m d^2_{D;pq}(\mathbf{x}^i, \boldsymbol{\mu})$.

It is easily seen that such a task is a generalization of the results presented in [2], where determining the $d^2_{D;pq}$-centroid was derived for the k-means procedure with $w_i = 1$ or $w_i = 0$ for all i.

Step 1: for $n = 1, 2, \dots, k$

$$\boldsymbol{\mu}^{(n)} = \arg\min_{\boldsymbol{\mu}\in\mathcal{S}_n} F_{\mathbf{w}}(\boldsymbol{\mu}) \tag{3}$$

Step 2:

$$\boldsymbol{\mu} = \arg\min_{n=1,\dots,k} F_{\mathbf{w}}(\boldsymbol{\mu}^{(n)}). \tag{4}$$

Let $n \in [k] := \{1, 2, \dots, k\}$ be fixed and let $\mathcal{P} \subseteq 2^{[n]}$ denote the partition of a set $[n]$, such that for each $P, P' \in \mathcal{P}$ we have $P \cap P' = \emptyset$, $|P| > 0$, $\bigcup_{P\in\mathcal{P}} = [n]$ and $\{i, j\} \in P$ with $i \le j$ implies that $i{+}1, i{+}2, \dots, j{-}1 \in P$. Moreover, $P_{\{i\}}$ stands for such $P \in \mathcal{P}$, that $\{i\} \in P$. Let $\mathcal{CP}[n]$ denote the whole class of such partitions.

Theorem 2. *For some $\mathcal{P} \in \mathcal{CP}[n]$ the vector $\mathbf{y} \in \mathbb{R}^n$ given by*

$$y_i = \frac{\sum_{f=1}^{l} \left(w_f^m \sum_{j\in P_i} \tilde{x}_j^{(f)} \right)}{|P_{\{i\}}| \sum_{f=1}^{l} w_f^m} \qquad \text{for } i = 1, \dots, n,$$

is a solution to Eq. (3) if $y_1 \ge y_2 \ge \dots \ge y_n$ and for all $i \in [n]$ with $i \in (P_{\{i\}} \setminus \{\max P_{\{i\}}\})$ we have

$$\frac{i - \min P_{\{i\}} + 1}{|P_{\{i\}}|} \sum_{f=1}^{l} w_f^m \Big\{ \sum_{j\in P_{\{i\}}} \tilde{x}_j^{(f)} \Big\} - \sum_{f=1}^{l} w_f^m \Big\{ \sum_{j\in P_{\{i\}}, j\le i} \tilde{x}_j^{(f)} \Big\} > 0,$$

Proof. The task is to find

$$\min_{\mathbf{y}\in\mathbb{R}^n} F_{\mathbf{w}}(\mathbf{y})$$

with respect to $n-1$ constraints of the form:

$$g_i(\mathbf{y}) \;=\; y_{i+1} - y_i \le 0 \quad \text{for } i = 1, \dots, n-1.$$

By means of the Karush-Kuhn-Tucker (KKT) theorem (cf. [12]), we need to find $\mathbf{y}$ and $\lambda_1, \dots, \lambda_{n-1}$ such that

$$\nabla F_{\mathbf{w}}(\mathbf{y}) + \sum_{i=1}^{n-1} \lambda_i \nabla g_i(\mathbf{y}) = 0,$$

with $\lambda_i g_i(\mathbf{y}) = 0$ and $\lambda_i \ge 0$ for $i \in [n-1]$. Note that for $h \in [n]$ we have:

$$\frac{\partial F_{\mathbf{w}}}{\partial y_h}(\mathbf{y}) = 2 \sum_{i=1}^{l} w_i^m (y_h - \tilde{x}_h^i).$$

For brevity of notation, let us assume that $\lambda_0 := 0$ and $\lambda_n := 0$. Thus, our task reduces to solving the following system of linear equations:

$$\left\{ \begin{array}{lllll}
0 & = & 2\sum_{i=1}^{l} w_i^m (y_1 - \tilde{x}_1^i) & +\lambda_0 & -\lambda_1 \\
0 & = & 2\sum_{i=1}^{l} w_i^m (y_2 - \tilde{x}_2^i) & +\lambda_1 & -\lambda_2 \\
\vdots & & & & \\
0 & = & 2\sum_{i=1}^{l} w_i^m (y_{n-1} - \tilde{x}_{n-1}^i) & +\lambda_{n-2} & -\lambda_{n-1} \\
0 & = & 2\sum_{i=1}^{l} w_i^m (y_n - \tilde{x}_n^i) & +\lambda_{n-1} & -\lambda_n \\
0 & = & \lambda_1 (y_2 - y_1) & & \\
\vdots & & & & \\
0 & = & \lambda_{n-1} (y_n - y_{n-1}) & &
\end{array} \right.$$

under constraints $\lambda_1 \ge 0, \dots, \lambda_{n-1} \ge 0$ and $y_1 \ge y_2 \ge \dots \ge y_n$.

Thus, let us consider a solution (not necessarily feasible) that fulfills $\boldsymbol{\lambda} \ge 0$. First of all, let us take u such that $\lambda_{u-1} = \lambda_u = 0$. It immediately implies that:

$$y_u = \frac{\sum_{f=1}^{l} w_f^m \tilde{x}_u^{(f)}}{\sum_{f=1}^{l} w_f^m}.$$

On the other hand, for each u and $p \ge 2$ such that $\lambda_{u-1} = 0$, $\lambda_u > 0$, $\lambda_{u+1} > 0$, …,$\lambda_{u+p-2} > 0$, $\lambda_{u+p-1} = 0$ we get that $y_u = \dots = y_{u+p-1}$. More specifically, we have:

$$y_i = \frac{\sum_{f=1}^{l} w_f^m \sum_{j=u}^{u+p-1} \tilde{x}_j^{(f)}}{p \sum_{f=1}^{l} w_f^m} \qquad \text{for } i = u, \dots, u+p-1,$$

In such a case, we have that for $i = u, \dots, u+p-2$:

$$\lambda_i = 2\frac{i-u+1}{p} \sum_{f=1}^{l} w_f^m \Big(\sum_{j=u}^{u+p-1} \tilde{x}_j^{(f)} \Big) - 2 \sum_{f=1}^{l} w_f^m \Big(\sum_{j=u}^{i} \tilde{x}_j^{(f)} \Big) > 0,$$

is a solution to Eq. 3. □

The procedure to compute the cluster centroid with respect to d_D^2 dissimilarity measure is given by Algorithm 1. Please note that this approach is a simple generalization of the algorithm included in [2], where the proof of its correctness was included. The main modification concern the form of input data. In the scenario considered in this paper, the algorithm is applied to a list of vectors of the form $\mathbf{x}^j = w_j^m \mathbf{x}^j / \sum_{i=1}^l w_j^m$. It is clear to see that when weights are either 0 or 1, both procedures return the same result.

```
**Data**: A set of $l$ vectors $\mathcal{X} \subset \mathcal{S}$ and $n \in \mathbb{N}$.
**Result**: $\boldsymbol{\mu}^{(n)} = \arg\min_{\boldsymbol{\mu} \in \mathcal{S}_n} F_{\mathbf{w}}(\boldsymbol{\mu})$.
Let $\tilde{\mathbf{x}}$ be such that $\bar{x}_i = \sum_{j=1}^l w_j^m \tilde{x}_i^j / \sum_{i=1}^l w_i^m$, for all $i \in [n]$;
Let $\mathcal{P} = \emptyset$;
Let $\mathbf{y} \in \mathbb{R}^n$;
for $k = 1, 2, \ldots, n$ do
    $y_k = \bar{x}_k$;
    Let $\mathcal{P} := \mathcal{P} \cup \{\{k\}\}$;    (we have $\mathcal{P} \in \mathcal{CP}([k])$)
    while $|\mathcal{P}| > 1$ and $y_{\min P^{(|\mathcal{P}|)}} > y_{\max P^{(|\mathcal{P}|-1)}}$ do
        $\mathcal{P} := \Big((\mathcal{P} \setminus \{P^{(|\mathcal{P}|)}\}) \setminus \{P^{(|\mathcal{P}|-1)}\}\Big) \cup \{P^{(|\mathcal{P}|-1)} \cup P^{(|\mathcal{P}|)}\}$;    (merge $P^{(|\mathcal{P}|-1)}, P^{(|\mathcal{P}|)}$)
        for $i \in P^{(|\mathcal{P}|)}$ do
            Set $y_i := \frac{1}{|P^{(|\mathcal{P}|)}|} \sum_{j \in P^{(|\mathcal{P}|)}} \bar{x}_j$;
        end
    end
end
return $\mathbf{y}$;
```

Algorithm 1: An algorithm to determine the weighted d_D^2-centroid.

Remark 1. *Even though the $d_{D;11}^2$-centroid can be viewed as an aggregated representation of a set of vectors, in general the procedure given by Theorem 2 is not a $\trianglelefteq$-monotonic fusion function, where $\trianglelefteq$ is the partial ordering described in [7]. For example, let us consider $\mathcal{X} = \{\mathbf{x}^{(1)} = (10, 2, 1, 0, 0), \mathbf{x}^{(2)} = (-11), \mathbf{x}^{(3)} = (-5, -6, -10)\}$ and $\mathcal{Y} = \{\mathbf{y}^{(1)} = (10, 2, 1, 0, 0), \mathbf{y}^{(2)} = (10, -100), \mathbf{y}^{(3)} = (-5, -6, -10)\}$. It is clear to see that for each $i = 1, 2, 3$ we have $\mathbf{x}^{(i)} \trianglelefteq \mathbf{y}^{(i)}$. However, for the corresponding centroids we have $(-1.67, -1.67, -3) \ntrianglelefteq (5, -34.67)$. On the other hand, if all the input elements are non-negative, then $\trianglelefteq$-monotonicity always holds.*

4 EMPIRICAL ANALYSIS

Let us now consider a data set consisting of citations received by 5000 scientists[1]. The data were gathered from Elsevier's Scopus (see [6] for details).

Table 1 contains basic sample statistics of the vectors. Please note that 78% of them are only of length 1 and among them, 32% are equal to 0.

Table 1: Basic summary statistics of vectors' lengths (n), maximal value (max) and sum of all elements (sum) (Scopus data set).

	Min.	Median	Mean	Max.
n	1	1	1.62	98
max	0	3	8.96	791
sum	0	3	12.63	1211

The fuzzy c-means and k-means algorithms were applied in order to determine 6 clusters (groups). The closest crisp clustering based on weights form the fuzzy c-means algorithm, was obtained by assigning each vector to the cluster with the maximal weight. The number of common vectors in clusters obtained via the k-means and fuzzy c-means algorithm are presented in Table 2.

Table 2: Number of common vectors in clusters obtained via the k-means (columns) and the c-means (rows) algorithm.

Cluster no.	1	2	3	4	5	6
1	3472	0	0	0	0	0
2	5	1052	6	0	0	0
3	0	2	308	6	0	0
4	0	0	10	100	0	0
5	0	0	0	6	27	0
6	0	0	0	0	0	6

Figure 2 presents the step functions of citation vectors in each cluster. Corresponding centroids are marked by black and red for the k-means and the fuzzy c-means procedure, respectively. The agreement between these two partitioning schemes, being equal to 99%, was calculated via the Rand Index, i.e. $A/(A + D)$, where A denotes the number of all pairs of data points assigned by both partitions into the same cluster or into different clusters (both partitionings agree for all pairs A) and D denotes the number of all pairs assigned differently by both partitions (the partitions disagree for all pairs D), cf. [11]. Figure 1 presents the distribution of maximal weights per each vector. Please note that there are 188 vectors for which any weight is no greater than 0.5.

Table 3 presents the weighted sum of dissimilarities between bibliometric indices for the Scopus data set, i.e. $\sum_{j=1}^c \sum_{k,i=1}^l w_{ij} w_{kj} (x_i - x_k)^2$. We analyzed the weights determined by the fuzzy c-means algorithm, the uniformly distributed membership degrees to all clusters and weights of the form $w_{ij} = 1$ for $\mathbf{x}^i$ if the k-means algorithm assigns i-th vector to j-th cluster and

[1] The data set is available at http://cena.rexamine.com/research/

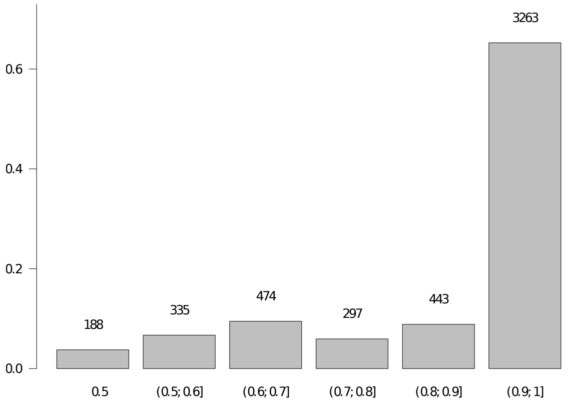


Figure 1: The distribution of cluster membership weights.

$w_{ij} = 0$ otherwise. Here, we consider the Hirsch index (H), the G-index (G), length (n), maximum (max) and the arithmetic mean ($mean$). Please note that the weights returned by the fuzzy c-means algorithm minimizes such sum for all the considered indexes.

Table 3: Weighted sum of dissimilarities between the Hirsch index (H), the G-index (G), length (n), Max (max) and arithmetic mean (mean) calculated for the fuzzy c-means output, uniform membership and k-means partitioning.

	c-means	uniform membership	k-means
G	951.01	2404321.83	1404.00
H	420.33	1763367.67	564.00
n	1347.73	2337926.83	1404.00
max	15865.30	27334939.66	20394.00
mean	10148.87	20259504.16	20207.30

Moreover, Table 4 presents the Hirsch index (H), G-index (G), length (n), Max operator (max) and the arithmetic mean ($mean$) computed for clusters centroids determined by the fuzzy c-means algorithm. We may see that the first centroid is characterized by a low H- and G-index and also a small number of publications. On the other hand, centroids corresponding to clusters 2, 3 and 4, represent researchers with a rather small number of publications (between 2 and 4) and a low H- and G-index, however, the number of citation given to their most cited paper (max) increase (12.55, 32.74, 68.77). Finally, researchers from clusters 5 and 6, represented by corresponding centroids are characterized with a high number of citations (175.51 and 907.50), and larger number of publications (10 and 24).

Table 4: The Hirsch index (H), G-index (G), length (n), maximum (max) and the arithmetic mean ($mean$) computed for cluster centroids determined by the fuzzy c-means algorithm.

Cl.no.	1	2	3	4	5	6
G	1.00	2.00	3.00	4.00	10.00	24.00
H	1.00	1.00	2.00	3.00	4.00	12.00
n	1.00	2.00	3.00	4.00	10.00	24.00
max	1.64	12.55	32.74	68.77	127.21	398.21
mean	1.64	7.00	12.67	20.05	17.55	37.81
sum	1.64	13.99	38.00	80.19	175.51	907.50

5 CONCLUSIONS

In this paper applications of the fuzzy c-means clustering algorithm to sets of vectors of possibly nonconforming lengths were investigated. First of all, a generalization of the procedure to compute the centroids of such sets, with respect to some dissimilarity measure tailored for vectors of unequal lengths, was provided. Moreover, the presented approach was verified by an empirical analysis on a bibliometric data set. Special attention was paid to the investigation of the clusters' centroids as an aggregated representation of a considered data set. By means of various bibliometric indexes, we evaluated the degree in which they reflect the structure of the whole data set.

Acknowledgements

The contribution of A. Cena and M. Gagolewski was partially supported by the National Science Center, Poland, research project 2014/13/D/HS4/01700.

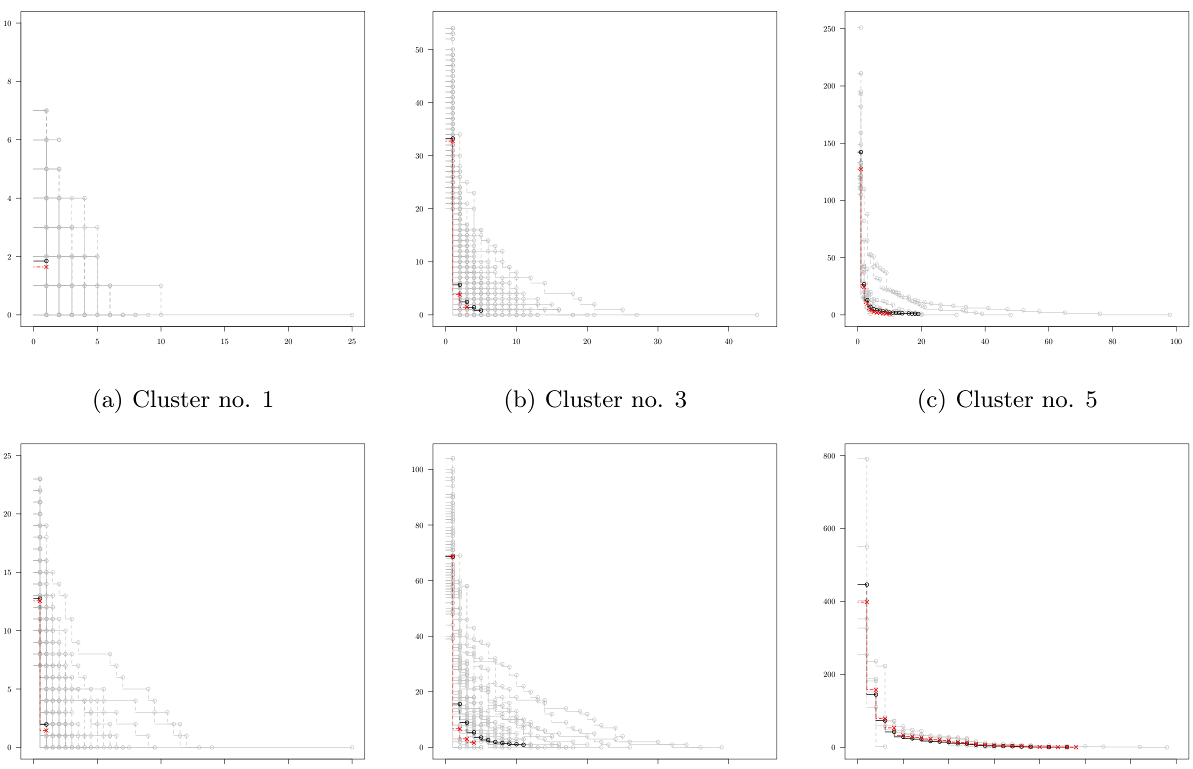
Anna Cena would like to acknowledge the support by the European Union from resources of the European Social Fund, Project PO KL “Information technologies: Research and their interdisciplinary applications”, agreement UDA-POKL.04.01.01-00-051/10-00 via the Interdisciplinary PhD Studies Program.

References

[1] J. C. Bezdek, R. Ehrlich, and W. Full (1984). FCM: The fuzzy c-means clustering algorithm. *Computer & Geosciences* 10(2–3), pp. 191–203.

[2] A. Cena and M. Gagolewski. A k-means-like algorithm for informetric data clustering, 2015. Submitted paper.

[3] A. Cena, M. Gagolewski, and R. Mesiar (2015). Problems and challenges of information resources producers' clustering. *Journal of Informetrics* 9(2), pp. 273–284.



(a) Cluster no. 1

(b) Cluster no. 3

(c) Cluster no. 5

(d) Cluster no. 2

(e) Cluster no. 4

(f) Cluster no. 6

Figure 2: Step functions depicting vectors in each cluster (Scopus) and their centroids (bold black – k-means, bold red – fuzzy c-means).

[4] Y. Cheng and N. C. Liu (2006). A first approach to the classification of the top 500 world universities by their disciplinary characteristics using scientometrics. *Scientometrics* 68(1), pp. 135–150.

[5] R. Costas, T. van Leeuwen, and M. Bordons (2001). A bibliometric classificatory approach for the study and assessment of research performance at the individual level: The effects of age on productivity and impact. *Journal of the American Society for Information Science and Technology* 61, pp. 1564–1581.

[6] M. Gagolewski (2011). Bibliometric impact assessment with R and the CITAN package. *Journal of Informetrics* 5(4), pp. 678–692.

[7] M. Gagolewski and P. Grzegorzewski (2011). Possibilistic analysis of arity-monotonic aggregation operators and its relation to bibliometric impact assessment of individuals. *International Journal of Approximate Reasoning* 52(9), pp. 1312–1324.

[8] T. Hastie, R. Tibshirani, and J. Friedman (2013). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer-Verlag.

[9] J. E. Hirsch (2005). An index to quantify individual's scientific research output. *Proceedings of the National Academy of Sciences* 102(46), pp. 16569–16572.

[10] G. J. Klir and B. Yuan (1995). *Fuzzy Sets and Fuzzy Logic: Theory and Applications.* Prentice Hall PTR.

[11] H. Lawrence and A. Phipps (1985). Comparing partitions. *Journal of Classification* 2, pp. 193–218.

[12] J. Nocedal and S. Wright (2006). *Numerical Optimization.* Springer-Verlag, New York.

[13] J. L. Ortega, E. López-Romero, and I. Fernández (2011). Multivariate approach to classify research institutes according to their outputs: The case of the CSIC's institutes. *Journal of Informetrics* 5, pp. 323–332.