OWA-based Linkage and the Genie Correction for Hierarchical Clustering

Anna Cena
Systems Research Institute,
Polish Academy of Sciences
ul. Newelska 6, 01-447 Warsaw, Poland
Email: 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, 00-662 Warsaw, Poland

***Abstract*—In this paper we thoroughly investigate various OWA-based linkages in hierarchical clustering on numerous benchmark data sets. The inspected setting generalizes the well-known single, complete, and average linkage schemes, among others. The incorporation of weights into the cluster merge procedure creates an opportunity to make use of experts' knowledge about a particular data domain so as to generate partitions of a given data set that better reflect the true underlying cluster structure. Moreover, we introduce a correction for the inequality of cluster size distribution – similar to the one proposed in our recently introduced Genie algorithm – which results in a significant performance boost in terms of clustering quality.**

I. Introduction

Hierarchical clustering algorithms provide a clear and simple way to perform data segmentation. By forming a whole hierarchy of nested partitions, they give a user a better insight into the underlying data structure. Let $\mathcal{C} = \{C_1, \dots, C_l\}$ be an l-partition of a data set $\mathcal{X} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(n)}\}$, i.e., a partition such that $C_u \cap C_v = \emptyset$ for $u \neq v$, $C_u \neq \emptyset$, and $\bigcup_{u=1}^{l} C_u = \mathcal{X}$. The main idea behind agglomerative hierarchical clustering methods is that in each step of the procedure, two clusters that are the "closest" to each other (with respect to some extension of a pointwise metric $\mathfrak{d} : \mathcal{X} \times \mathcal{X} \to [0, \infty]$) are joined. Initially each cluster consists of only one data point, i.e., $\mathcal{C}^{(0)} = \{C_1^{(0)}, \dots, C_n^{(0)}\}$, $C_i^{(0)} = \{\mathbf{x}^{(i)}\}$, $i = 1, \dots, n$. Proceeding from the $(j-1)$-th to the j-th step, we merge the clusters $C_u^{(j-1)}$ and $C_v^{(j-1)}$ with $u < v$ such that the distance between them is the smallest. In result we obtain $C_i^{(j)} = C_i^{(j-1)}$ for $u \neq i < v$, $C_u^{(j)} = C_u^{(j-1)} \cup C_v^{(j-1)}$, and $C_i^{(j)} = C_{i+1}^{(j-1)}$ for $i > v$.

Of course, there are many ways of extending $\mathfrak{d}$ so as to measuring the distance between two clusters is possible. A procedure which allows to quantify the "closeness" of two data groups is called a *linkage scheme*. Among the most commonly used ones, see, e.g., [1], we find:

- single linkage:

$$\mathfrak{d}^*_{\mathrm{MIN}}(C_u, C_v) = \min_{\mathbf{u} \in C_u, \mathbf{v} \in C_v} \mathfrak{d}(\mathbf{u}, \mathbf{v});$$

- complete linkage:

$$\mathfrak{d}^*_{\mathrm{MAX}}(C_u, C_v) = \max_{\mathbf{u} \in C_u, \mathbf{v} \in C_v} \mathfrak{d}(\mathbf{u}, \mathbf{v});$$

- average linkage:

$$\mathfrak{d}^*_{\mathrm{AMean}}(C_u, C_v) = \frac{1}{|C_u||C_v|} \sum_{\mathbf{u} \in C_u, \mathbf{v} \in C_v} \mathfrak{d}(\mathbf{u}, \mathbf{v}).$$

In [2] Nasıbov and Kandemır-Cavas stated a natural extension of the three above linkages based on the Ordered Weighted Averaging (OWA) operators [3], which are amongst the most well-known and extensively studied aggregation functions, see, e.g., [4], [5], [6]. The main motivation behind the introduction of OWA-based linkages was not only the possibility to take into account experts' knowledge, but also the fact that such a general criterion can easily interpolate between single, complete, and average linkages. However, the authors did not convey any systematic studies on the effects of choosing different OWA operator weights. Therefore, in order to fill this void, in this paper we are going to evaluate the effects of choosing different weighting vectors on the clustering quality based on numerous benchmark data sets.

What is more, very recently we have introduced a new, fast, and outlier resistant *Genie* algorithm [7]. The routine takes into account the correction for the inequality of cluster size distribution and aims to boost the performance of the single linkage scheme. Here we shall verify whether such a modification works fabulously in a more general setting too.

The paper is organized as follows. In Sect. II we recall the OWA-based linkage scheme as well as review different generators of OWA weights. Next, in Sect. III, we propose a generalization of the Genie algorithm and discuss its possible computer implementation. In Sect. IV we evaluate its performance on a comprehensive set of benchmark data. Finally, in Sect. V we conclude the paper and sketch some interesting future research directions.

II. OWA-based Linkage

A z-ary *Ordered Weighting Average* (OWA) operator $\mathrm{OWA}_{\mathbf{w}} : [0, \infty]^z \to [0, \infty]$ is given by:

$$\mathrm{OWA}_{\mathbf{w}}(d_1, d_2, \dots, d_z) = \sum_{i=1}^{z} w_i d_{(i)},$$

where $\mathbf{w} \in [0,1]^z$ is a *weighting vector* such that $\sum_{i=1}^{z} w_i = 1$ and $d_{(i)}$ denotes the i-th greatest value in $\mathbf{d} = (d_1, d_2, \dots, d_z)$, i.e.:

$$d_{(1)} \geq d_{(2)} \geq \dots \geq d_{(z)}.$$

For the sake of our discussion, we shall be interested in conceiving OWA operators as extended aggregation functions i.e., defined for any number of arguments. Here we shall follow the convention from [8], see also [9]: given a *weighting triangle* $\triangle = (w_{i,z} \in [0,1],\ i = 1, \dots, z,\ z \in \{1, 2, \dots\})$ such that $\sum_{i=1}^{z} w_{i,z} = 1$ for all z, the corresponding $\mathrm{OWA}_\triangle : \bigcup_{z=1}^{\infty} [0,\infty]^z \to [0,\infty]$ is defined as:

$$\mathrm{OWA}_\triangle(d_1, d_2, \dots, d_z) = \sum_{i=1}^{z} w_{i,z} d_{(i)}.$$

Now the $\mathrm{OWA}_\triangle$-based linkage is given by:

$$\mathfrak{d}^*_{\mathrm{OWA}_\triangle}(C_u, C_v) = \mathrm{OWA}_\triangle(d_1, d_2, \dots, d_z),$$

where $C_u = \{\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(|C_u|)}\}$ and $C_v = \{\mathbf{v}^{(1)}, \dots, \mathbf{v}^{(|C_v|)}\}$ are some point sets, $z = |C_u||C_v|$, and:

$$d_{i+|C_u|(j-1)} = \mathfrak{d}(\mathbf{u}^{(i)}, \mathbf{v}^{(j)}),$$

for all $i = 1, \dots, |C_u|$ and $j = 1, \dots, |C_v|$.

A. Generation of Weights

Please note that the choice of a weighting triangle can provide us with a very wide range of possible ways to perform data clustering. It is clear to see that for weights like $w_{i,z} = \frac{1}{z}$ we obtain the average linkage, $w_{z,z} = 1$ and $w_{i,z} = 0$ for $i < z$ gives us the single linkage scheme, and $w_{1,z} = 1$, $w_{i,z} = 0$, $i > 1$ – the complete linkage.

Let us denote with $\varphi(\cdot; \mu_z, \sigma_z)$ the probability density function of the normal distribution $\mathrm{N}(\mu_z, \sigma_z)$:

$$\varphi(i; \mu_z, \sigma_z) = \frac{1}{\sqrt{2\pi\sigma_z^2}} \exp\left(-\frac{(i-\mu_z)^2}{2\sigma_z^2}\right). \qquad (1)$$

Please note that in [2] only one additional weighting triangle $\triangle$ was actually considered, namely:

$$w_{i,z} = \frac{\varphi(i; \mu_z, \sigma_z)}{\sum_{j=1}^{z} \varphi(j; \mu_z, \sigma_z)},$$

where $\mu_z = \frac{z+1}{2}$ and $\sigma_z = \sqrt{\frac{1}{z}\sum_{i=1}^{z}(i-\mu_z)^2}$, see [10].

Convenient ways to generate weighting triangles include settings like:

1) $w_{i,z} = \frac{c_i}{\sum_{j=1}^{z} c_j}$, where a sequence $(c_1, c_2, \dots)$ is such that $c_i \geq 0$ for all $i = 1, 2, \dots$ and $c_1 > 0$, see, e.g., [11];
2) $w_{i,z} = \mathsf{w}\left(\frac{i}{z}\right) - \mathsf{w}\left(\frac{i-1}{z}\right)$, where $\mathsf{w} : [0,1] \to [0,1]$ is a nondecreasing function with $\mathsf{w}(0) = 0$ and $\mathsf{w}(1) = 1$, see, e.g., [3].

In the investigation carried out in this paper we shall take into account weights that interpolate around the single and complete linkage, sample quartiles, various means and some mixture of the above. The complete list of the weighting triangles considered is given in Tab. I. The settings for the weighting triangle generation were as follows: σ_z in each use of φ was set to $z/3$ as well as $z/9$; p in the trimmed and Winsorized means was set to 0.25; (a, b) in the Yager step function was set to $(0, 0.5)$, $(0.5, 1)$, as well as $(0.3, 0.7)$.

TABLE I: Exemplary weighting triangles

Alias	**Weighting triangle**
AMean (average)	$w_{i,z} = \frac{1}{z}$
MIN (single)	$w_{z,z} = 1$, $w_{i,z} = 0$ for $i < z$
MAX (complete)	$w_{1,z} = 1$, $w_{i,z} = 0$ for $i > 1$
Q2 (median)	$w_{(z+1)/2,z} = 1$ for $z = 2k+1$ $w_{z/2,z} = w_{z/2+1,z} = 0.5$ for $z = 2k$
Q1 (first quartile)	Median taken over the upper half of the vector sorted nonincreasingly (without median)
Q3 (third quartile)	Median taken over the lower half of the vector sorted nonincreasingly (without median)
Norm [10]	$w_{i,z} = \frac{\varphi(i;\mu_z,\sigma_z)}{\sum_{j=1}^{z}\varphi(j;\mu_z,\sigma_z)}$ $\mu_z = \frac{z+1}{2}$ $\sigma_z = \sqrt{\frac{1}{z}\sum_{i=1}^{z}(i-\mu_z)^2}$
fuzzyMIN_{σ_z}	$w_{i,z} = \frac{\varphi(i;z,\sigma_z)}{\sum_{j=1}^{z}\varphi(j;z,\sigma_z)}$
fuzzyMAX_{σ_z}	$w_{i,z} = \frac{\varphi(i;1,\sigma_z)}{\sum_{j=1}^{z}\varphi(j;1,\sigma_z)}$
fuzzy$\mathrm{MIN\&MAX}_{\sigma_z}$	$c_i = \max\left\{\varphi(i;1,\sigma_z), \varphi(i;z,\sigma_z)\right\}$ $w_{i,z} = \frac{c_i}{\sum_{j=1}^{z} c_j}$
fuzzy$\mathrm{Q3}_{\sigma_z}$	$w_{i,z} = \frac{\varphi(i;\frac{1}{4}z,\sigma_z)}{\sum_{j=1}^{z}\varphi(j;\frac{1}{4}z,\sigma_z)}$
fuzzy$\mathrm{Q1}_{\sigma_z}$	$w_{i,z} = \frac{\varphi(i;\frac{3}{4}z,\sigma_z)}{\sum_{j=1}^{z}\varphi(j;\frac{3}{4}z,\sigma_z)}$
fuzzy$\mathrm{Q1\&Q3}_{\sigma_z}$	$c_i = \max\left\{\varphi(i;\frac{3z}{4},\sigma_z), \varphi(i;\frac{z}{4},\sigma_z)\right\}$ $w_{i,z} = \frac{c_i}{\sum_{i=1}^{z} c_i}$
$\mathrm{TriMean}_p$	$w_{i,z} = \frac{1}{z-2k}$ for $i = k+1, \dots, z-k$ $w_{i,z} = 0$ otherwise, $k = \lfloor pz \rfloor$
$\mathrm{WinMean}_p$	$w_{i,z} = \frac{1}{z}$ for $i = k+2, \dots, z-k-1$ $w_{i,z} = \frac{(k+1)}{z}$ for $i = k+1, z-k$ $w_{i,z} = 0$ otherwise, $k = \lfloor pz \rfloor$
$\text{Yager-step}_{(a,b)}$ [3]	$w_{i,z} = Q\left(\frac{i}{z}; a, b\right) - Q\left(\frac{i-1}{z}; a, b\right)$ $Q(x;a,b) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & x > b \end{cases}$ $0 \leq a < b \leq 1$

```
$\mathrm{NNI}[1..n] = -1$                                ▷ Nearest Neighbor Index
$\mathrm{NND}[1..n] = \infty$                            ▷ Nearest Neighbor Distance
$\mathrm{L}[1..n, 1..n] = -1$                            ▷ Auxiliary Matrix (Distance Cache)
$\mathrm{ds} = \mathrm{DisjointSets}\left(\{\mathbf{x}^{(1)}\}, \dots, \{\mathbf{x}^{(n)}\}\right)$    ▷ Union-Find
for all $u = 1, \dots, n-1; v = u+1, \dots, n$ do
    $w = \mathfrak{d}(\mathbf{x}^{(u)}, \mathbf{x}^{(v)})$
    if $\mathrm{NND}[u] > w$ then
        $\mathrm{NND}[u] = w$
        $\mathrm{NNI}[u] = v$
    end if
    if $\mathrm{NND}[v] > w$ then
        $\mathrm{NND}[v] = w$
        $\mathrm{NNI}[v] = u$
    end if
end for
for $j = 1, \dots, n-1$ do                                ▷ Main loop
    Merge and update...                                   ▷ See algorithm in Fig. 2
end for
```

Fig. 1: The hierarchical clustering algorithm incorporating the OWA-based linkage and the Genie correction.

III. Generalized Genie Algorithm

In [7] we have recently introduced the *Genie* algorithm, which incorporates the notion of an economic inequality index (see, e.g., [12] and references therein) into the single linkage procedure in order to prevent the formation of a highly uneven cluster structure. It is worth noting that the performance of the Genie algorithm was significantly better than not only that of the standard hierarchical clustering routines, but also that of the BIRCH or the l-means algorithm. Moreover, in [13] we studied the effects of applying the Genie correction on the generalized centroid (penalty-based aggregation) linkage scheme; the results were highly encouraging.

We shall now propose a generalized hierarchical clustering algorithm bracketing both the single-linkage-based Genie routine and the classic OWA linkage. Fix a weighting triangle $\triangle$, a threshold $g \in (0, 1]$, and let G be a chosen inequity index (e.g., the Gini-index, see below). Proceeding from the $(j-1)$-th to the j-th step of the clustering procedure, $j = 1, \dots, n-1$, merge clusters $C_u^{(j-1)}$, $C_v^{(j-1)}$ such that:

1) if $\mathrm{G}(c_1, \dots, c_{n-j+1}) \le g$, where $c_i = |C_i^{(j-1)}|$, apply the standard OWA-based linkage criterion:

$$(u, v) = \quad \mathop{\arg\min}_{u<v} \quad \mathfrak{d}^*_{\mathrm{OWA}_\triangle}\left(C_u^{(j-1)}, C_v^{(j-1)}\right);$$

2) otherwise, force merging of a cluster of the smallest size:

$$(u, v) = \mathop{\arg\min}_{\substack{u<v;\\ c_u = c_{(n-j+1)} \text{ or}\\ c_v = c_{(n-j+1)}}} \mathfrak{d}^*_{\mathrm{OWA}_\triangle}\left(C_u^{(j-1)}, C_v^{(j-1)}\right).$$

```
$\mathrm{bestX} = \mathrm{bestY} = -1$
$\mathrm{bestD} = \infty$
$m = \mathrm{ds.getMinClusterSize}()$                     ▷ $c_{(n-j+1)}$
for all $u = \mathrm{ds.iterateOverClusterIDs}()$ do
    if $\mathrm{NND}[u] < \mathrm{bestD}$ and
        $\big(\mathrm{ds.getGiniIndex}() \le g$ or        ▷ $\mathrm{G}(c_1, \dots, c_{n-j+1})$
        $\mathrm{ds.getClusterSize}(u) = m$ or
        $\mathrm{ds.getClusterSize}(\mathrm{NNI}[u]) = m\big)$ then
        $\mathrm{bestD} = \mathrm{NND}[u]$
        $\mathrm{bestX} = u$
        $\mathrm{bestY} = \mathrm{NNI}[u]$
    end if
end for
$\mathrm{bestX} = \mathrm{ds.union}(\mathrm{bestX}, \mathrm{bestY})$     ▷ Merge
$\mathrm{NND}[\mathrm{bestX}] = \infty$
for $u = 1, \dots, n$ do
    $\mathrm{L}[\mathrm{bestX}, u] = \mathrm{L}[u, \mathrm{bestX}] = -1.0$
    $\mathrm{L}[\mathrm{bestY}, u] = \mathrm{L}[u, \mathrm{bestY}] = -1.0$
end for
for all $\mathrm{bestX} \ne u = \mathrm{ds.iterateOverClusterIDs}()$ do
    $w = \mathfrak{d}^*_{\mathrm{OWA}_\triangle}(\mathrm{ds.getCluster}(\mathrm{bestX}), \mathrm{ds.getCluster}(u))$
    $\mathrm{L}[\mathrm{bestX}, u] = \mathrm{L}[u, \mathrm{bestX}] = w$
    if $\mathrm{NND}[\mathrm{bestX}] > w$ then
        $\mathrm{NND}[\mathrm{bestX}] = w$
        $\mathrm{NNI}[\mathrm{bestX}] = u$;
    end if
    if $\mathrm{NNI}[u] = \mathrm{bestX}$ or $\mathrm{NNI}[u] = \mathrm{bestY}$ then
        $\mathrm{NND}[u] = w$
        $\mathrm{NNI}[u] = \mathrm{bestX}$
        for all $u \ne v = \mathrm{ds.iterateOverClusterIDs}()$ do
            $w' = \mathrm{L}[u, v]$
            if $w' < 0$ then
                $w' = \mathfrak{d}^*_{\mathrm{OWA}_\triangle}(\mathrm{ds.getCluster}(u),$
                        $\mathrm{ds.getCluster}(v))$
                $\mathrm{L}[u, v] = \mathrm{L}[v, u] = w'$
            end if
            if $\mathrm{NND}[u] > w'$ then
                $\mathrm{NND}[u] = w'$
                $\mathrm{NNI}[u] = v$
            end if
        end for
    else if $\mathrm{NND}[u] > w$ then
        $\mathrm{NND}[u] = w$
        $\mathrm{NNI}[u] = \mathrm{bestX}$
    end if
end for
```

Fig. 2: A subroutine for merging and updating clusters.

As far as computational issues are concerned, one of the advantages of the classic Genie approach [7] is that it can be implemented based on a minimum spanning tree. For instance, we may rely on a parallelizable version of the Prim algorithm, see [14], which requires $\Theta(n^2)$ time and exactly $(n^2-n)/2$ evaluations of the $\mathfrak{d}$ function. Unfortunately, its generalized version will not be as efficient as its predecessor, but we of course hope for its better performance in terms of clustering quality. Figures 1 and 2 give a pseudocode of the procedure we used for performing the empirical analysis described in the next section.

IV. Empirical Analysis

A. Benchmark Data Sets

For the sake of comparison of the clustering performance, we considered 29 benchmark data sets in $\mathbb{R}^d$ for different d. They consist of balanced and unbalanced data of various shapes. Twenty-one data sets (available at www.gagolewski.com/resources/data/clustering/) were already used in our previous contributions [7], [13], namely:

1) a1 ($n = 3000$, $d = 2$, $l = 20$);
2) a2 ($n = 5250$, $d = 2$, $l = 35$);
3) a3 ($n = 7500$, $d = 2$, $l = 50$);
4) g2-16-100 ($n = 2048$, $d = 16$, $l = 2$);
5) g2-2-100 ($n = 2048$, $d = 2$, $l = 2$);
6) g2-64-100 ($n = 2048$, $d = 64$, $l = 2$);
7) iris ($n = 150$, $d = 4$, $l = 3$);
8) iris5 ($n = 105$, $d = 4$, $l = 3$);
9) s1 ($n = 5000$, $d = 2$, $l = 15$);
10) s2 ($n = 5000$, $d = 2$, $l = 15$);
11) s3 ($n = 5000$, $d = 2$, $l = 15$);
12) s4 ($n = 5000$, $d = 2$, $l = 15$);
13) Aggregation ($n = 788$, $d = 2$, $l = 7$);
14) Compound ($n = 399$, $d = 2$, $l = 6$);
15) D31 ($n = 3100$, $d = 2$, $l = 31$);
16) flame ($n = 240$, $d = 2$, $l = 2$);
17) jain ($n = 373$, $d = 2$, $l = 2$);
18) pathbased ($n = 300$, $d = 2$, $l = 3$);
19) R15 ($n = 600$, $d = 2$, $l = 15$);
20) spiral ($n = 312$, $d = 2$, $l = 3$);
21) unbalance ($n = 6500$, $d = 2$, $l = 8$).

We decided to consider 8 additional data sets which are part of the Fundamental Clustering Problem Suite (see www.uni-marburg.de/fb12/arbeitsgruppen/datenbionik/data/ and [15]):

22) Atom ($n = 800$, $d = 3$, $l = 2$);
23) Chainlink ($n = 1000$, $d = 3$, $l = 2$);
24) EngyTime ($n = 4096$, $d = 2$, $l = 2$);
25) Lsun ($n = 400$, $d = 2$, $l = 3$);
26) Target ($n = 770$, $d = 2$, $l = 6$);
27) Tetra ($n = 400$, $d = 3$, $l = 4$);
28) TwoDiamonds ($n = 800$, $d = 2$, $l = 2$);
29) WingNut ($n = 1016$, $d = 2$, $l = 2$).

What is more, we set $\mathfrak{d}$ to be the Euclidean metric.

B. Partition Similarity Measures

Each data set comes with a sequence of reference labels; the number of true underlying clusters is denoted above with l. In order to measure the degree of concordance between the generated partitions (the resulting dendrograms were cut at an appropriate level) and the true labels, we decided to rely on the notion of the Fowlkes–Mallows (see [16]) as well as the Adjusted Rand index (see [17]). Let $\mathcal{C} = \{C_1, \dots, C_l\}$ and $\mathcal{C}' = \{C'_1, \dots, C'_l\}$ be two l-partitions of the set $\mathcal{X}$ of cardinality n. Moreover, let $m_{u,v} = |C_u \cap C'_v|$, $m_{u,\cdot} = \sum_{v=1}^{l} m_{u,v}$, and $m_{\cdot,v} = \sum_{u=1}^{l} m_{u,v}$. The Fowlkes–Mallows (FM) index is given by:

$$\text{FM-index}(\mathcal{C}, \mathcal{C}') = \frac{\sum_{u=1}^{l} \sum_{v=1}^{l} m_{u,v}^2 - n}{\sqrt{\left(\sum_{u=1}^{l} m_{u,\cdot}^2 - n\right)\left(\sum_{v=1}^{l} m_{\cdot,v}^2 - n\right)}}.$$

On the other hand, the Adjusted Rand (AR) index is given by:

$$\text{AR-index}(\mathcal{C}, \mathcal{C}') = \frac{\binom{n}{2} \sum_{u=1}^{l} \sum_{v=1}^{l} \binom{m_{u,v}}{2} - c\,d}{\frac{1}{2}\binom{n}{2}(c+d) - c\,d},$$

where $c = \sum_{u=1}^{l} \binom{m_{u,\cdot}}{2}$ and $d = \sum_{v=1}^{l} \binom{m_{\cdot,v}}{2}$. The AR-index has zero expected value in the case of two random (uniformly-distributed) partitions, while the FM-index in such a case yields $1/l$. Both indices are bounded from above by 1 in the case of a perfect agreement between two given partitions.

C. Inequity Indices

In [13] we have already investigated the impact of the choice of an inequity measure on the clustering quality in the case of the single and generalized centroid linkage and found no significant differences between the Gini, Bonferroni, and de Vergottini indices. Therefore, here we only inspect the normalized Gini-index [18] which is given by:

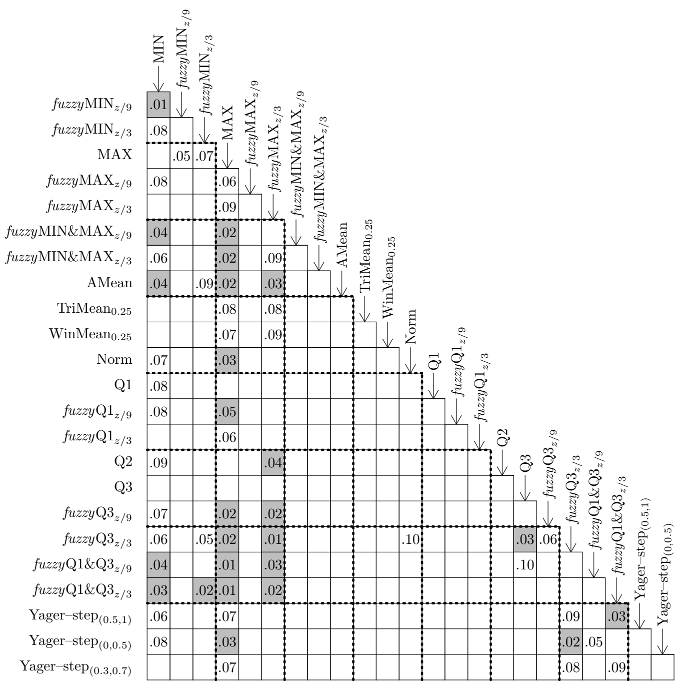
$$\mathrm{G}(c_1, \dots, c_l) = \frac{\sum_{u=1}^{l-1} \sum_{v=i+1}^{l} |c_u - c_v|}{(l-1)\sum_{i=u}^{l} c_u}.$$

D. Results

For each data set we applied the proposed generalized OWA-based linkage with different weighting triangles (see Tab. I) and different Genie correction thresholds $g = 0.1, 0.2, \dots, 1.0$. In each case we computed the FM- and AR-index for 5 random input data set permutations and aggregated the similarity degrees by applying the sample median.

For every unique pair of weighting triangles and $g = 1.0$ (no Genie correction), we compared the FM-index distributions. Figure 3 gives the significant p-values as reported by the Wilcoxon (paired) signed rank test (null hypothesis: pairwise differences between the 29 FM-indices are symmetric around 0). The smallest p-values are observed for $\mathit{fuzzy}\mathrm{MAX}_{z/3}$ and $\mathit{fuzzy}\mathrm{Q3}_{z/3}$ (0.007), MIN and $\mathit{fuzzy}\mathrm{MIN}_{z/9}$ (0.01), as well as MAX and $\mathit{fuzzy}\mathrm{Q1\&Q3}_{z/9}$ (0.01).

Figure 4 depicts the box-and-whiskers plots for the FM- and the AR-index distributions as a function of different weighting triangles. In each case we report the results obtained for the Gini-index threshold of $g = 1$ (no Genie correction) and the threshold that yielded the best median value. Moreover, we performed the signed rank test in order to determine if the introduction of the Genie correction leads to a significant improvement in the clustering quality. A bold, red arrow indicates p-value ≤ 0.05. Please note that the Genie correction in each case improved the (raw) median FM- and AR-indices, even though the (very conservative) Wilcoxon test finds the differences significant in only some of the cases. Interestingly, we observe the most striking improvement in the case of the original Genie algorithm [7], where from the globally worst



	MIN	*fuzzy*MIN$_{z/9}$	*fuzzy*MIN$_{z/3}$	MAX	*fuzzy*MAX$_{z/9}$	*fuzzy*MAX$_{z/3}$	*fuzzy*MIN&MAX$_{z/9}$	*fuzzy*MIN&MAX$_{z/3}$	AMean	TriMean$_{0.25}$	WinMean$_{0.25}$	Norm	Q1	*fuzzy*Q1$_{z/9}$	*fuzzy*Q1$_{z/3}$	Q2	Q3	*fuzzy*Q3$_{z/9}$	*fuzzy*Q3$_{z/3}$	*fuzzy*Q1&Q3$_{z/9}$	*fuzzy*Q1&Q3$_{z/3}$	Yager–step$_{(0.5,1)}$	Yager–step$_{(0,0.5)}$
*fuzzy*MIN$_{z/9}$	.01																						
*fuzzy*MIN$_{z/3}$	.08																						
MAX		.05	.07																				
*fuzzy*MAX$_{z/9}$	.08			.06																			
*fuzzy*MAX$_{z/3}$				.09																			
*fuzzy*MIN&MAX$_{z/9}$	.04			.02																			
*fuzzy*MIN&MAX$_{z/3}$	.06			.02		.09																	
AMean	.04		.09	.02		.03																	
TriMean$_{0.25}$				.08		.08																	
WinMean$_{0.25}$				.07		.09																	
Norm	.07			.03																			
Q1	.08																						
*fuzzy*Q1$_{z/9}$	.08			.05																			
*fuzzy*Q1$_{z/3}$				.06																			
Q2	.09					.04																	
Q3																							
*fuzzy*Q3$_{z/9}$	.07			.02		.02																	
*fuzzy*Q3$_{z/3}$	.06		.05	.02		.01						.10					.03	.06					
*fuzzy*Q1&Q3$_{z/9}$	.04			.01		.03											.10						
*fuzzy*Q1&Q3$_{z/3}$	.03		.02	.01		.02																	
Yager–step$_{(0.5,1)}$	.06			.07															.09		.03		
Yager–step$_{(0,0.5)}$	.08			.03															.02	.05			
Yager–step$_{(0.3,0.7)}$				.07															.08		.09		

Fig. 3: p-values of the Wilcoxon signed rank test for the difference between the FM-index distributions for each weighting triangle; $g = 1$; only p-values ≤ 0.1 are reported.

performing clustering routine we jump to the best one. Also recall that this algorithm is very fast to compute.

Table II gives the top median values of the FM-index and the AR-index for each Gini-index threshold inspected. In each case, the corresponding weighting triangles are listed. We observe that the single linkage (MIN) and its generalization ($\mathit{fuzzy}\mathrm{MIN}_{z/9}$) work best when combined with small Gini-thresholds ($g = 0.2$) and give the highest median value of the similarity measures.

Lastly, Fig. 5 depicts the violin plots for the FM- and AR-index distributions for:

- the average (AMean) linkage with $g = 1$;
- $\mathit{fuzzy}\mathrm{MIN}_{z/9}$ with $g = 1$;
- the single (MIN) linkage with $g = 0.2$ (recommended in [7] for practical use);
- the "best" OWA weighting triangle (on a per-data set level) in the case of $g = 1$;
- the "best" OWA weighting triangle and the "best" Gini-index threshold (on a per-data set level).

The figure provides us with the upper bound for the FM- and AR-index if we had possessed an oracle (or a set of experts) deciding which $\triangle$ or g to use prior to starting the clustering procedure.

V. Conclusion

OWA-based linkage was stated in [2] but has not been extensively studied up till now. In this paper we have investigated the effects of choosing different OWA-based linkages on data clustering. The obtained results indicate that by choosing an appropriate weighting triangle, we may improve the overall segmentation quality but none of the linkages is perfect on all the data sets. Moreover, we have proposed the incorporation of a correction for inequality of the cluster size distribution, similar to the one from our recently-introduced Genie algorithm. It turns out that such a modification yields an even better segmentation.

An adaptive selection of the Gini-index threshold as well as the best weighting triangle is left for further research. Due to its simple form and a wide range of linkage possibilities and also taking into account the computational complexity, in this paper only OWA-based clustering was considered. However, the presented setting could be generalized by means of the Choquet integral – this research direction is also worth a deeper investigation in future studies.

Acknowledgment

This study was partially supported by the National Science Center, Poland, research project 2014/13/D/HS4/01700.

References

[1] R. Xu and D. C. Wunsch II, *Clustering*. Wiley-IEEE Press, 2009.

[2] E. Nasıbov and C. Kandemır-Cavas, "OWA-based linkage method in hierarchical clustering: Application on phylogenetic trees," *Expert Systems with Applications*, vol. 38, pp. 12 684–12 690, 2011.

[3] R. R. Yager, "On ordered weighted averaging aggregation operators in multicriteria decision making," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 18, no. 1, pp. 183–190, 1988.

[4] R. R. Yager, J. Kacprzyk, and G. Beliakov, Eds., *Recent Developments in the Ordered Weighted Averaging Operators*. Springer, 2011.

[5] R. R. Yager and J. Kacprzyk, Eds., *The ordered weighted averaging operators. Theory and applications*. Norwell: Kluwer Academic Publishers, 1997.

[6] M. Grabisch, J.-L. Marichal, R. Mesiar, and E. Pap, *Aggregation functions*. Cambridge University Press, 2009.

[7] M. Gagolewski, M. Bartoszuk, and A. Cena, "Genie: A new, fast, and outlier-resistant hierarchical clustering algorithm," *Information Sciences*, vol. 363, pp. 8–23, 2016.

[8] G. Mayor and T. Calvo, "On extended aggregation functions," in *Proc. IFSA 1997*. Prague: Academia, 1997, vol. 1, pp. 281–285.

[9] T. Calvo, G. Mayor, J. Torrens, J. Suner, M. Mas, and M. Carbonell, "Generation of weighting triangles associated with aggregation functions," *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems*, vol. 8, no. 4, pp. 417–451, 2000.

[10] Z. Xu, "An overview of methods for determining OWA weights," *International Journal of Intelligent Systems*, vol. 20, pp. 843–865, 2005.

[11] B. Jamison, S. Orey, and W. Pruitt, "Convergence of weighted averages of independent random variables," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 4, no. 1, pp. 40–44, 1965.

[12] G. Beliakov, M. Gagolewski, and S. James, "Penalty-based and other representations of economic inequality," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 24(Suppl.1), pp. 1–23, 2016.

[13] M. Gagolewski, A. Cena, and M. Bartoszuk, "Hierarchical clustering via penalty-based aggregation and the Genie approach," in *Modeling Decisions for Artificial Intelligence*, ser. Lecture Notes in Artificial Intelligence, V. Torra *et al.*, Eds. Springer, 2016, vol. 9880, pp. 191–202.

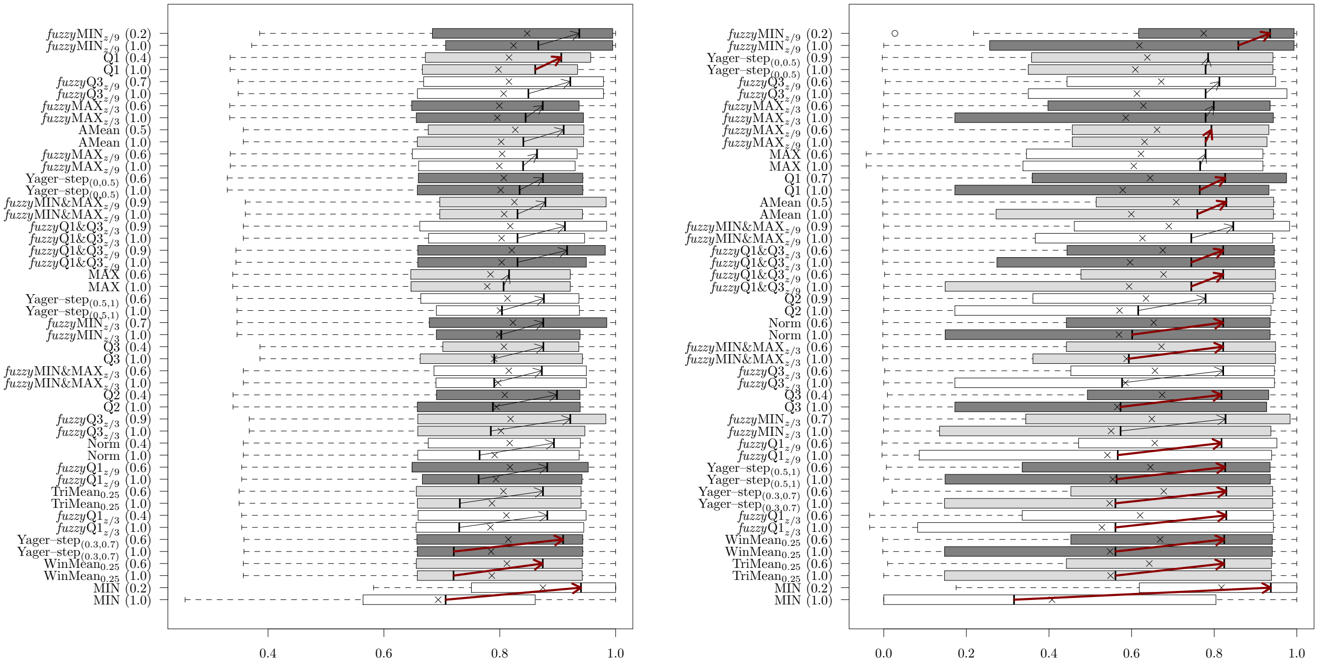
[14] C. F. Olson, "Parallel algorithms for hierarchical clustering," *Parallel Computing*, vol. 21, pp. 1313–1325, 1995.

[15] A. Ultsch, "Clustering with SOM: U*C," in *Workshop on Self-Organizing Maps*, 2005, pp. 75–82.

[16] E. Fowlkes and C. Mallows, "A method for comparing two hierarchical clusterings," *Journal of the American Statistical Association*, vol. 78, no. 383, pp. 553–569, 1983.

[17] H. Lawrence and A. Phipps, "Comparing partitions," *Journal of Classification*, vol. 2, pp. 193–218, 1985.

[18] C. Gini, *Variabilità e mutabilità*. Bologna: C. Cuppini, 1912.



(a) FM-index

(b) AR-index

Fig. 4: Box-and-whisker plots for the FM-index and the AR-index distribution for each scenario with $(g < 1.0)$ and without $(g = 1.0)$ the Genie correction.

TABLE II: Top median FM-index and AR-index together with the corresponding weighting scenarios for each Gini-index threshold

	FM-index		**AR-index**	
g	weighting triangle	top median	△	top median
1.0	$fuzzy$MIN$_{z/9}$ Q1	0.8665	$fuzzy$MIN$_{z/9}$	0.8583
0.9	$fuzzy$Q1&Q3 $fuzzy$Q3$_{z/3}$	0.9218	$fuzzy$MIN$_{z/9}$	0.8618
0.8	$fuzzy$Q1&Q3 $fuzzy$Q3	0.9218	$fuzzy$MIN$_{z/9}$	0.8618
0.7	$fuzzy$Q1&Q3 $fuzzy$Q3	0.9218	$fuzzy$MIN$_{z/9}$ $fuzzy$MIN&MAX$_{z/9}$	0.8555
0.6	AMean $fuzzy$Q1&Q3 $fuzzy$Q3$_{z/3}$ $fuzzy$MIN$_{z/9}$ step$_{0.3-0.7}$	0.9150	$fuzzy$MIN$_{z/9}$	0.8813
0.5	$fuzzy$MIN$_{z/9}$	0.9250	$fuzzy$MIN$_{z/9}$	0.9195
0.4	$fuzzy$MIN$_{z/9}$	0.9267	$fuzzy$MIN$_{z/9}$	0.9242
0.3	$fuzzy$MIN$_{z/9}$ MIN	0.9272	MIN	0.9152
0.2	$fuzzy$MIN$_{z/9}$ MIN	0.9402	$fuzzy$MIN$_{z/9}$ MIN	0.9371
0.1	MIN	0.9374	MIN	0.9353

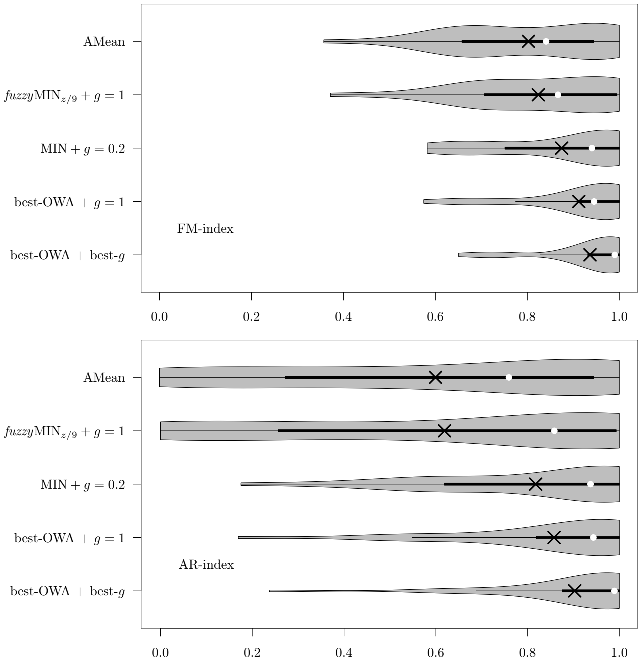


Fig. 5: Violin plots for the FM-index and AR-index distribution for the average linkage (AMean, $g = 1$), $fuzzy$MIN$_{z/9}$ ($g = 1$), the single linkage ($g = 0.2$) [7], the OWA-based linkage with the best weighting scenario chosen individually for each data set without ($g = 1$) and with the Genie correction applied with the best possible threshold found (best-g).