Engineering and Technology Journal e-ISSN: 2456-3358
Volume 09 Issue 03 March-2024, Page No.-3558-3564
DOI: 10.47191/etj/v9i03.01, I.F. – 8.227

Interval-valued Data Group Average Clustering (IUPGMA)

Sérgio Mário Lins Galdino[1], Jornandes Dias da Silva[1], Cícero José da Silva[1], Willames de Albuquerque Soares[1], Juan Carlos Oliveira de Medeiros[2]
[1]University of Pernambuco, Polytechnic School, Recife-PE, Brazil
[2]Federal University of Ceará, Electrical Engineering Department, Campus Sobral-CE, Brazil

ABSTRACT: In this paper, we deal with a particular type of information, namely interval-valued data. We face the problem of clustering data units described by intervals of the real data set (interval data). Currently, clustering methods rely on dissimilarity measures for interval-valued data uses representative point distance. The Data Group Average Clustering UPGMA is one of the popular algorithms to construct a phylogenetic tree according to the distance matrix created by the pairwise distances among taxa. A phylogenetic tree is used to present the evolutionary relationships among the interesting biological species based on the similarities in their genetic sequences. Interval-valued Data Group Average Clustering (IUPGMA) extends Group Average clustering to interval-valued data. Based on the Range Euclidean Metric it is a reliable alternative to be used to uncertainty quantification from interval-valued data. They contain more information than point-valued data, and such informational advantages could be exploited to yield more efficient analysis.

KEYWORDS: hierarchical clustering, unsupervised machine learning, interval valued-data, interval arithmetic, range Euclidean metric

I. INTRODUCTION

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups (clusters). Clustering can therefore be formulated as a multi-objective optimization problem.

Some popular clustering algorithms includes K-Means Clustering, Hierarchical Clustering (We'll discuss here), Mean-Shift Clustering, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), and Expectation-Maximization (EM) Clustering using Gaussian Mixture Models (GMM).

Hierarchical clustering is an approach for identifying groups in the dataset. It does not require us to pre-specify the number of clusters to be generated as is required by the k-means approach. It has an added advantage over K-means clustering in that it results in tree called a dendrogram. To determine how close together two clusters are, we can use a few different methods including: *Complete linkage* clustering find the max distance between points belonging to two different clusters. *Single linkage* clustering finds the minimum distance between points belonging to two different clusters. *Mean linkage clustering* finds all pairwise distances between points belonging to two different clusters and then calculate the average. *Centroid linkage* clustering find the centroid of each cluster and calculate the distance between the centroids of two different clusters. *Ward's* minimum variance method minimize the total.

Group Average, also called the unweighted pair-group method, is perhaps the most popular of all the hierarchical cluster techniques. Better known as UPGMA is treated as a clustering technique that uses the (unweighted) arithmetic averages of the measures of dissimilarity. Note that the unweighted term indicates that all distances contribute equally to each average that is computed and does not refer to the math by which it is achieved. Thus, the simple averaging in WPGMA produces a weighted result and the proportional averaging in UPGMA produces an unweighted result. UPGMA is a distance method and therefore needs a distance matrix [1].

Mather [2] suggests that the Group Average method is the safest to use as an exploratory method, although he goes on to suggest that several methods should be tried and the one with the largest cophenetic correlation be selected for further investigation.

Interval-valued data are frequent in real life; examples include maximum and minimum daily temperatures, maximum and minimum asset prices in a trading period, high and low blood pressures, bid and ask prices, saving and lending interest rates, and so on. They contain more information than point-valued data, and such informational advantages could be exploited to yield more efficient analysis.

In this paper, we deal with a particular type of information, namely interval-valued data. We face the problem of clustering data units described by variables whose values are intervals of the real data set (interval data) [3-4]. We will

focus on IUPGMA-Group Average Clustering agglomerative method for Interval-valued Data. The Range Euclidean Metric for interval-valued data is used to compare two vectors of intervals.

The rest of the paper is organized as follows: Section II presents basic concepts of interval arithmetic and distance measures for interval data. Section III describes interval-valued input distance matrix for clustering. Section IV introduces the approach IUPGMA for clustering interval data, Section V provides the conclusion.

II. INTERVAL ANALYSIS

Interval arithmetic is a method for determining absolute errors of an algorithm, considering all data errors and rounding [5]. Interval arithmetic makes systematic calculations through intervals $[x] = [\underline{x}, \overline{x}]$ limited to representable machine numbers $\underline{x}, \overline{x} \in \mathbb{F}$, instead of real numbers x. Arithmetic operations $+, -, \times, \div$ are defined using intervals. Interval algorithms produce interval results guaranteed to contain the true solution. If $\odot$ denotes any of these arithmetic operation for real numbers x and y, then the corresponding operation for arithmetic on interval numbers $[x]$ and $[y]$ is

$$[x]\odot[y] = \{x \odot y | x \in [x], y \in [y]\}.$$

Thus, the interval $[x]\odot[y]$ resulting from the operations contain every possible number that can be found as $x \odot y$ for each $x \in [x]$, and each $y \in [y]$:

$$[x] + [y] = [\underline{x}, \overline{x}] + [\underline{y}, \overline{y}] = [\underline{x} + \underline{y}, \overline{x} + \overline{y}]$$

$$[x] - [y] = [\underline{x}, \overline{x}] - [\underline{y}, \overline{y}] = [\underline{x} - \overline{y}, \overline{x} - \underline{y}]$$

$$[x] \cdot [y] = [\underline{x}, \overline{x}] \cdot [\underline{y}, \overline{y}]$$

$$= \left[\boldsymbol{min}\left(\underline{x}\underline{y}, \underline{x}\overline{y}, \overline{x}\underline{y}, \overline{x}\overline{y}\right), \boldsymbol{max}\left(\underline{x}\underline{y}, \underline{x}\overline{y}, \overline{x}\underline{y}, \overline{x}\overline{y}\right)\right]$$

$$\frac{[x]}{[y]} = [\underline{x}, \overline{x}] \cdot \frac{1}{[\underline{y}, \overline{y}]} = [\underline{x}, \overline{x}] \cdot \left[\frac{1}{\overline{y}}, \frac{1}{\underline{y}}\right], \text{ if } 0 \notin [\underline{y}, \overline{y}] \quad (1)$$

The interval arithmetic operations are defined for exact calculation [5]. Machine computations are affected by rounding errors. Therefore, the formulas were modified in order to consider the called directed rounding [6].

Throughout this paper, all matrices are denoted by bold capital letters ($\mathbf{A}$), vectors by bold lowercase letters ($\boldsymbol{a}$), and scalar variables by ordinary lowercase letters (a). Interval variables are enclosed in square brackets ($[\boldsymbol{A}], [\boldsymbol{a}], [a]$). Underscores and overscores denote lower and upper bounds, respectively. A real interval [x] is a nonempty set of real numbers

$$[x] = [\underline{x}, \overline{x}] = \{\tilde{x} \in R : \underline{x} \leq \tilde{x} \leq \overline{x}\} \quad (2)$$

where $\underline{x}$ and $\overline{x}$ are called the *infimum (inf)* and *supremum (sup)*, respectively, and $\tilde{x}$ is a point value belonging to an interval variable $[x]$. The set of all intervals $\mathbb{R}$ is denoted by $I(\mathbb{R})$ where

$$I(R) = \{[\underline{x}, \overline{x}] : \underline{x}, \overline{x} \in R : \underline{x} \leq \overline{x}\}. \quad (3)$$

A. *Order Relations of Intervals*

The important issue in using interval data for decision problems is the choice of an appropriate interval order relation. Unlike real numbers that are ordered by a strict transitive relation "<" (if $a < b$ and $b < c$, then $a < c$ for any a, b, and $c \in \mathbb{R}$), the ranking of intervals is not symmetric, and as consequence, in many situations, the definitions cannot differentiate two intervals in general. Theoretically intervals can only have partial order in $I(\mathbb{R})$. According to Moore et al. (2009) [7], two transitive order relations can be defined for intervals: (i) $[x] \leq [y] \Leftrightarrow \overline{x} \leq \underline{y}$, and (ii) $[x] \subseteq [y] \Leftrightarrow \underline{y} \leq \underline{x}$ and $\overline{x} \leq \overline{y}$ (set inclusion). Let $[x]$ and $[y]$ be a pair of arbitrary intervals. These can be classified as follows: non-overlapping intervals; partially overlapping intervals; completely overlapping intervals. In contrast to real numbers, it is not straightforward to define a total order relation for intervals. As a result, researchers have defined order relations in different ways. Most of these definitions cannot specify the order relations properly for completely overlapping intervals. A detailed description and comparison between these and other ranking definitions is given in Karmakar and Bhunia (2012) [8].

Definition. Given two intervals $[x], [y] \in I(\mathbb{R})$, $[x] \leq [y]$, *iff* $m([x]) \leq m([y])$, where $m(\mathcal{X})$ is a point within the interval $\mathcal{X} \in \{[x], [y]\}$, usually the *midpoint*, *infimum*, and *supremum*. We propose the following order relation: $[x] \leq [y]$ is determined by choosing the interval *infimum* that captures the "*minimum*" between the two intervals, i.e., the interval with the lowest *infimum*.

B. *Range of interval-valued function*

The range of an interval-valued function can be expressed in interval form as

$$range\big(f([\boldsymbol{x}])\big) = f([x_1], [x_2], \dots, [x_n])$$
$$= [inf\ (f([x_1], [x_2], \dots, [x_n]), sup\ (f([x_1], [x_2], \dots, [x_n]))] \quad (4)$$

where the *inf* and *sup* are taken for all $x_i \in [x]_i (i = 1, \dots n)$.

Finding the range of a multi-variable function over a box is a fundamental problem

encountered in numerous applications. The main focus of interval arithmetic is the simplest way to calculate upper and lower endpoints for the range of values of a function in one or more variables. These endpoints are not necessarily the *supremum* or *infimum*, since the precise calculation of those values can be difficult or impossible. In special cases the exact range can be found in a straightforward way [7],[9].

C. *Range Euclidean Distance*

The Range Euclidean Distance between interval vectors $[p]$ and $[q]$ is the interval length of the lines segment connecting them ($\overline{[p][q]}$).

In cartesian coordinates, if $[p] = ([p_1], [p_2], \dots, [p_n])$ and $[q] = ([q_1], [q_2], \dots, [q_n])$ are two interval vectors in Euclidean n-space (i.e., $I(\mathbb{R}^n)$), then the distance $[d_2]$ from $[p]$ to $[q]$, or from $[\boldsymbol{q}]$ to $[\boldsymbol{p}]$ is given by the Interval Pythagorean formula:

$$d_2([\boldsymbol{p}],[\mathbf{q}]) \;=\; d_2([\mathbf{q}],[\mathbf{p}]) \;=$$
$$=\sqrt{([q_1]-[p_1])^2+([q_2]-[p_2])^2+\cdots+([q_n]-[p_n])^2}$$
$$=\sqrt{\sum_{i=1}^{n}([q_i]-[p_i])^2}$$

(5)

Consider the function x^2 as a monotonically decreasing function for $x < 0$ and a monotonically increasing function for $x > 0$.

The range corresponding to the interval $[x]^2$ can be calculated by applying the function to its endpoints:

$$[x_1,x_2]^2=\begin{cases}[x_1^2,x_2^2] & x_1\geq 0\\ [x_2^2,x_1^2] & x_2<0\\ 0,max\{x_1^2,x_2^2\} & otherwise\end{cases}$$

(6)

III. INTERVAL-VALUED INPUT DISTANCE MATRIX

Table I provides the oils dataset originally presented by Ichino [10].The categories are 6 types of oil (linseed, perilla, cottonseed, sesame, camellia, and olive) and 2 fats (beef and hog); N = 8 multi-valued objects. There are four interval-valued variables, namely, *Specific Gravity* (in g/cm^3), *Freezing Point* (in℃), *Iodine Value*, and *Saponification*.

The **iodine value** (or *iodine adsorption value*) in chemistry is the mass of iodine in grams that is consumed by 100 grams of a chemical substance.

https://en.wikipedia.org/wiki/Iodine_value (Accessed January 20, 2024).

Saponification value (or *saponification number*, or *Koettstorfer number*, also referred to as *sap* in short) represents the number of milligrams of potassium hydroxide required to saponify 1g of fat under the conditions specified.

https://en.wikipedia.org/wiki/Saponification_value (Accessed January 20, 2024).

Table I. interval-valued observations: oils and fats.

Sample (Label)	Features			
	Specific gravity	Freezing point	Iodine value	Saponification value
Linseed oil (A)	[0.930, 0.935]	[−27, −18]	[170, 204]	[118, 196]
Perilla oil (B)	[0.930, 0.937]	[−5, −4]	[192, 208]	[188, 197]
Cottonseed oil (C)	[0.916, 0.918]	[−6, −1]	[99, 113]	[189, 198]
Sesame oil (D)	[0.92, 0.926]	[−6, −4]	[104, 116]	[187, 193]
Camellia oil (E)	[0.916, 0.917]	[−21, −15]	[80, 82]	[189, 193]
Olive oil (F)	[0.914, 0.919]	[0, 6]	[79, 90]	[187, 196]
Beef tallow (G)	[0.86, 0.87]	[30, 38]	[40, 48]	[190, 199]
Hog fat (H)	[0.858, 0.864]	[22, 32]	[53, 77]	[190, 202]

In our example, the distance matrix is an 8×8 table with the lines and rows representing the objects (i.e., fats and oils) under consideration. As the distance between objects A and B (in this case $[13,\ 90.632]$) is the same as between B and A, the distance matrix is symmetrical. Furthermore, since the distance between an object and itself include zero, one must only look at either the lower or upper non-diagonal elements (see Table II lower non-diagonal distance matrix).

IV. INTERVAL-VALUED HIERARCHICAL CLUSTERING

CLUSTERING

Hierarchical Clustering Algorithms

Given a set of N objects (i.e., observations, individuals, cases, or data rows) to be clustered, and an NxN distance (or similarity) matrix, the basic process of Johnson's [11] hierarchical clustering is this:

1) Start by assigning each object to its own cluster, so that if you have N objects, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters equal the distances (similarities) between the objects they contain.
2) Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one less cluster.
3) Compute distances (similarities) between the new cluster and each of the old clusters.
4) Repeat steps 2 and 3 until all items are clustered into a single cluster of size N.

Step 3 can be done in different ways, which is what distinguishes HCA (Hierarchical Clustering Algorithms). Should be stressed that in Step 2, we will use *infimum* to find the closest pair of clusters. Should be stressed that in Step 2, we will use *infimum* of interval to find the closest pair of clusters. The range metrics gives make possible others merge points as explored (see [3]).

A. *IUPGMA*

The IUPGMA (**I**nterval-valued data **U**nweighted **P**air-**G**roup **M**ethod with **A**rithmetic mean) algorithm produces rooted dendrograms. At each step, the nearest two clusters are combined into a higher-level cluster. The distance between any two clusters $[\mathcal{A}]$ and $[\mathcal{B}]$, each of size (i.e., cardinality) $|[\mathcal{A}]|$ and $|[\mathcal{B}]|$, is taken to be the average of all distances $d([x],[y])$ between pairs of interval-valued data objects $[x]$ in $[\mathcal{A}]$ and $[y]$ in $[\mathcal{B}]$, that is, the mean distance between elements of each cluster:

$$\frac{1}{|[\mathcal{A}]|\cdot|[\mathcal{B}]|}\sum_{[x]\in[\mathcal{A}]}\sum_{[y]\in[\mathcal{B}]}d([x],[y])$$

(7)

In other words, at each clustering step, the updated distance between the joined clusters $[\mathcal{A}]\cup[\mathcal{B}]$ and a new cluster $[\mathcal{X}]$ is given by the proportional averaging of the $d([\mathcal{A}],[\mathcal{X}])$ and $d([\mathcal{B}],[\mathcal{X}])$ distances:

$d([\mathcal{A}]\cup[\mathcal{B}],[\mathcal{X}]) =$

$$\frac{||[\mathcal{A}]|| \cdot d([\mathcal{A}],[\mathcal{X}]) + ||[\mathcal{B}]|| \cdot d([\mathcal{A}],[\mathcal{X}])}{||[\mathcal{A}]|| + ||[\mathcal{B}]||} \quad (8)$$

The IUPGMA is similar to its weighted variant, the IWPGMA method. Unweighted means that that all distances contribute equally to each average that is computed and does not refer to the used algebraic approach by which it is achieved. IWPGMA and UPGMA produces a weighted and unweighted results, respectively.

B. *Step-by-Step IUPGMA Clustering*

First step

First clustering: Let us assume that we have eight elements (A, B, C, D, E, F, G). From distance matrix Table II of pairwise distances between them: $d_{(C,D)} = [0.002,\ 20.857]$ has the smallest *infimum* value of distance matrix from Table II, so we join elements C and D.

First distance matrix update: We then proceed to update the initial distance matrix Table II into a new distance matrix Table III, reduced in size by one row and one column because of the clustering of C with D.

$$d_{(C,D)\to A} = \frac{(1 \times d_{CA} + 1 \times d_{DA})}{(1+1)} = \frac{(\,1 \times [58.249,\ 134.54] + 1 \times [55.317,\ 127.1]\,)}{2} = [56.783,\ 130.82]$$

$$d_{(C,D)\to B} = \frac{(1 \times d_{CB} + 1 \times d_{DB})}{(1+1)} = \frac{(\,1 \times [79,\ 109.54] + 1 \times [76,\ 104.5]\,)}{2} = [77.5,\ 107.02]$$

$$d_{(C,D)\to E} = \frac{(1 \times d_{CE} + 1 \times d_{DE})}{(1+1)} = \frac{(\,1 \times [19.235, 39.624] + 1 \times [23.769, 40.262]\,)}{2} = [21.502, 39.943]$$

$$d_{(C,D)\to F} = \frac{(1 \times d_{CF} + 1 \times d_{DF})}{(1+1)} = \frac{(\,1 \times [9.0553, 37.697] + 1 \times [14.56, 39.925]\,)}{2} = [11.807,\ 38.811]$$

Table II. Lower Non-Diagonal Distance Matrix.

	A	B	C	D	E	F	G
B	[13, 90.632]						
C	[58.249, 134.54]	[79, 109.54]					
D	[55.317, 127.1]	[76, 104.5]	[0.002, 20.857]				
E	[88, 145.42]	[110.4, 129.38]	[19.235, 39.624]	[23.769, 40.262]			
F	[82, 151]	[102.0, 129.86]	[9.0553, 37.697]	[14.56, 39.925]	[15, 29.632]		
G	[131.1, 194.12]	[147.9,\ 173.77]	[59.682, 85.82]	[65.513, 88.635]	[55.217, 73.11]	[39.204, 63.938]	
H	[101.2, 182.59]	[117.9, 159.97]	31.827, 72.202]	[37.483, 75.087]	[37.121, 61.799]	[16.124, 51.167]	[5, 42.06]

Min. Distance (linkage) = [0.002, 20.857]

$$d_{(C,D)\to G} = \frac{(1 \times d_{CG} + 1 \times d_{DG})}{(1+1)} = \frac{(\,1 \times [59.682, 85.82] + 1 \times [65.513, 88.635]\,)}{2} = [62.597, 87.227]$$

$$d_{(C,D)\to H} = \frac{(1 \times d_{CH} + 1 \times d_{DH})}{(1+1)} = \frac{(\,1 \times [31.827, 72.202] + 1 \times [37.483,\ 75.087]\,)}{2} = [34.655, 73.644]$$

Second step

Second clustering: We now reiterate the two previous steps starting from the new distance matrix Table III. Here, $d_{(G,H)} = [5,\ 42.06]$ has the lowest *infimum* value of Table III, so we join elements G and H.

Second distance matrix update: We then proceed to update the matrix Table III into a new distance matrix Table IV, reduced in size by one row and one column because of the clustering of G with H:

$$d_{(G,H)\to A} = \frac{(1 \times d_{GA} + 1 \times d_{HA})}{(1+1)} = \frac{(\,1 \times [131.1, 194.12] + 1 \times [101.23,\ 182.59]\,)}{2} = [116.17,\ 188.36]$$

$$d_{(G,H)\to B} = \frac{(1 \times d_{GB} + 1 \times d_{HB})}{(1+1)} = \frac{(\,1 \times [147.95,\ 173.77] + 1 \times [117.9,\ 159.97]\,)}{2} = [132.93,\ 166.87]$$

$$d_{(G,H)\to (C,D)} = \frac{(1 \times d_{G(C,D)} + 1 \times d_{H(C,D)})}{(1+1)} = \frac{(\,1 \times [65.513,\ 88.635] + 1 \times [37.483,\ 75.087]\,)}{2} = [48.626,\ 80.436]$$

$$d_{(G,H)\to E} = \frac{(1 \times d_{GE} + 1 \times d_{HE})}{(1+1)}$$
$$= \frac{(\,1 \times [55.217,\ 73.11] + \ 1 \times [37.121,\ 61.799]\,)}{2}$$
$$= [46.169,\ 67.454]$$
$$d_{(G,H)\to F} = \frac{(1 \times d_{GF} + 1 \times d_{HF})}{(1+1)}$$
$$= \frac{(\,1 \times [39.204,\ 63.938] \ + \ 1 \times [16.124,\ 51.167]\,)}{2}$$
$$= [27.664,\ \ 57.552]$$

Third step

Third clustering: We now reiterate starting from the new distance matrix Table IV. Here, $d_{(C,D),F} = [11.807,\ 38.811]$ has the lowest *infimum* value of Table IV so we join elements (C, D) a and F.

Third distance matrix update:} We then proceed to update the matrix Table IV into a new distance matrix Table V, reduced in size by one row and one column because of the clustering of (C, D) with F:

$$d_{(C,D),F\to A} = \frac{(2 \times d_{(C,D),A} + 1 \times d_{F,A})}{(2+1)}$$
$$= \frac{(\,2 \times [56.783,\ 130.82] + \ 1 \times [82,\ 151]\ \ )}{3}$$
$$= [65.188,\ 137.55]$$
$$d_{(C,D),F\to B} = \frac{(2 \times d_{(C,D),B} + 1 \times d_{F,B})}{(2+1)}$$
$$= \frac{(\,2 \times [77.5,\ 107.02] + \ 1 \times [102.07,\ 129.86]\,)}{3}$$
$$= [85.692,\ 114.63]$$
$$d_{(C,D),F\to E} = \frac{(2 \times d_{(C,D),E} + 1 \times d_{F,E})}{(2+1)}$$
$$= \frac{(\,2 \times [21.502,\ 39.943] + \ 1 \times [15,\ 29.632]\,)}{3}$$
$$= [19.335,\ 36.506]$$
$$d_{(C,D),F\to (G,H)} = \frac{(2 \times d_{(C,D),(G,H)} + 1 \times d_{F,(G,H)})}{(2+1)}$$
$$= \frac{(\,2 \times [48.626,\ 80.436] + \ 1 \times [27.664,\ 57.552]\,)}{3}$$
$$= [41.639,\ 72.808]$$

Table III. Grouped Cluster (C,D).

	A	B	C, D	E	F	G
B	[13, 90.632]					
C, D	**[56.783, 130.82]**	**[77.5, 107.02]**				
E	[88, 145.42]	[110.4, 129.38]	**[21.502, 39.943]**			
F	[82, 151]	[102.0, 129.86]	**[11.807, 38.811]**	[15, 29.632]		
G	[131.1, 194.12]	[147.9,\ 173.77]	**[62.597, 87.227]**	[55.217, 73.11]	[39.204, 63.938]	
H	[101.2, 182.59]	[117.9, 159.97]	**[34.655, 73.644]**	[37.121, 61.799]	[16.124, 51.167]	[5, 42.06]

Min. Distance (linkage) = [5, 42.06]

Table IV. Grouped Cluster (G,H).

	A	B	C, D	E	F
B	[13, 90.632]				
C, D	[56.783, 130.82]	[77.5, 107.02]			
E	[88, 145.42]	[110.4, 129.38]	[21.502, 39.943]		
F	[82, 151]	[102.0, 129.86]	[11.807, 38.811]	[15, 29.632]	
G, H	**[116.17, 188.36]**	**[132.93, 166.87]**	**[48.626, 80.436]**	**[46.169, 67.454]**	**[27.664, 57.552]**

Min. Distance (linkage) = [11.807, 38.811]

Table V. Grouped Cluster ((C,D),F).

	A	B	(C, D), F	E
B	[13, 90.632]			
(C, D), F	**[65.188, 137.55]**	**[85.692, 114.63]**		
E	[88, 145.42]	[110.4, 129.38]	**[19.335, 36.506]**	
G, H	[116.17, 188.36]	[132.93, 166.87]	**[41.639, 72.808]**	[46.169, 67.454]

Min. Distance (linkage) = [13, 90.632]

Fourth step

Fourth clustering: We now reiterate starting from the new distance matrix Table V. Here, $d_{A,B} = [13, 90.632]$ has the lowest *infimum* value of Table V, so we join element A with cluster B.

Fourth distance matrix update: We then proceed to update the matrix Table V into a new distance matrix Table VI, reduced in size by one row and one column because of the clustering of A with B:

$$d_{(A,B)\to((C,D),F)} = \frac{(1 \times d_{A,((C,D),F)} + 1 \times d_{B,((C,D),F)})}{(1+1)}$$
$$= \frac{(\,1 \times [65.188,\ 137.55] + \ 1 \times [85.692,\ 114.63]\,)}{2}$$
$$= [75.44,\ 126.09]$$
$$d_{(A,B)\to E} = \frac{(1 \times d_{A,E} + 1 \times d_{B,E})}{(1+1)}$$
$$= \frac{(\,1 \times [88,\ 145.42] + \ 1 \times [110.45,\ 129.38]\,)}{2}$$
$$= [99.226,\ 137.4]$$
$$d_{(A,B)\to(G,H)} = \frac{(1 \times d_{A,(G,H)} + 1 \times d_{B,(G,H)})}{(1+1)}$$
$$= \frac{(\,1 \times [116.17,\ 188.36] + \ 1 \times [132.93,\ 166.87]\,)}{2}$$
$$= [124.55,\ 177.61]$$

Table VI Grouped Cluster (A,B).

	A, B	(C, D), F	E
(C, D), F	**[75.44, 126.09]**		
E	**[99.226, 137.4]**	[19.335, 36.506]	
G, H	**[124.55}, 177.61]**	[41.639, 72.808]	[46.169, 67.454]

Min. Distance (linkage) = [19.335, 36.506]

Fifth step

Fifth clustering: We now reiterate starting from the new distance matrix Table VI. Here, $d_{E,((C,D),F)} = [19.335, 36.506]$ has the lowest *infimum* value of Table VI, so we join element E with cluster $((C, D), F)$.

Fifth distance matrix update: We then proceed to update the matrix Table VI into a new distance matrix Table VII, reduced in size by one row and one column because of the clustering of E with $((C, D), F)$:

$$d_{E,((C,D),F)\rightarrow(A,B)} = \frac{(1 \times d_{E,(A,B)} + 3 \times d_{((C,D),F),(A,B)})}{(1+3)}$$

$$= \frac{(1 \times [99.226, 137.4] + 3 \times [75.44, 126.09])}{4}$$

$$= [81.387, 128.92]$$

$$d_{E,((C,D),F)\rightarrow(G,H)} = \frac{(1 \times d_{E,(G,H)} + 3 \times d_{((C,D),F),(G,H)})}{(1+3)}$$

$$= \frac{(1 \times [46.169, 67.454] + 3 \times [41.639, 72.808])}{4}$$

$$= [42.771, 71.47]$$

Table VII. Grouped Cluster (E,(C,D),F)).

	A, B	E, (C, D), F
E, (C, D), F)	**[81.387, 128.92]**	
G, H	[124.55, 177.61]	**[42.771, 71.47]**

Min. Distance (linkage) = [42.771, 71.47]

Final step

Starting from the new distance matrix Table VII we have $d_{(G,H),(E,((C,D),F))} = [42.771, 71.47]$ the lowest *infimum* value of Table VII, so we join element (G, H) with cluster $(E, ((C, D), F))$. We then proceed to update the matrix into a new distance matrix Table VIII, reduced in size by one row and one column because of the clustering of (G, H) with $(E, ((C, D), F))$:

$$d_{(G,H),(E,((C,D),F))\rightarrow(A,B)} = \frac{(2 \times d_{(G,H),(A,B)} + 4 \times d_{(E,((C,D),F)),(A,B)})}{(2+4)}$$

$$= \frac{(2 \times [124.55, 177.61] + 4 \times [81.387, 128.92])}{6}$$

$$= [95.775, 145.15]$$

Table VII. Grouped Cluster ((G,H),(E,(C,D),F)).

	A, B
(G, H), (E, (C, D), F))	**[95.775,145.15]**

Min. Distance (linkage) = [95.775,145.15]

This process is summarized by the clustering diagram on Figure 1. In that diagram, the columns are associated with the objects and the rows are associated with heights of clustering. The rectangle colour introduced for each level for 'X' cells is placed in a given row if the corresponding objects are merged at that stage in the clustering. We observe partial order from [heights], not every pair of heights interval are disjoint sets.

GROUP AVERAGE (IUPGMA) CLUSTERING DIAGRAM.

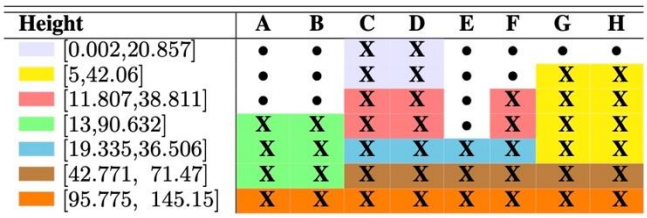

Figure 1: IUPGMA Diagram.

V. CONCLUSIONS

This paper concerns the IUPGMA clustering to interval-valued data based on the Range Euclidean Metric. Interval-valued Data Group Average method is an alternative to be used in uncertainty quantification for interval-valued data. Uncertainty propagation is the quantification of uncertainties in system output(s) propagated from uncertain inputs. Range Euclidean lower and upper bounds are associated with heights of clustering. Note that, conclusions about the proximity of two objects can be drawn only based on the height where branches containing those two objects first are fused. Range metrics can allow for a reliable analysis of the clustering results on interval-valued data.

Interval-valued Data Group Average Clustering can be considered as an extension of classical Group Average method. It can be understood in the context of cluster membership (fuzzy clustering). Basically, it allows partial membership which means that it contains elements that have varying degrees of membership in the cluster. From this, we can understand the difference between UPGMA method and IUPGMA method: UPGMA method contains elements that satisfy precise properties of membership while IUPGMA method contains elements that satisfy imprecise properties of membership.

It is strongly recommend comparing the dendrograms from different representatives to find the closest (most similar) pair of clusters and merge them into a new single cluster on several different datasets with known cluster patterns so that you can get the feel of the technique.

REFERENCES

1. Gan, G., Ma, C. and Wu, J.: *Data Clustering: Theory, Algorithms, and Applications.* ASA-SIAM Series on Statistics and Applied Probability, SIAM, Philadelphia, ASA, Alexandria, VA, 2007.
2. Mather, P.: *Computational Methods of Multivariate Analysis in Physical Geography.* John Wiley & Sons, 1976

3. Galdino, S.M.L., Dias, J.: *Interval-valued Data Ward's Hierarchical Agglomerative Clustering Method: Comparison of Three Representative Merge Points*. In: 2021 International Conference on Engineering and Emerging Technologies (ICEET), 2021, Istanbul. 2021 International Conference on Engineering and Emerging Technologies (ICEET), 2021. p. 1-6.
4. Dias, J. and Galdino, S.M.L.: *Interval-valued Data Ward's Minimum Variance Clustering - Centroid update Formula*. In: 2021 International Conference on Engineering and Emerging Technologies (ICEET), 2021, Istanbul. 2021 International Conference on Engineering and Emerging Technologies (ICEET), 2021. p. 1-6.
5. Moore, R.E.: *Interval Analysis.* Prentice Hall, Englewood Clifs, NJ, USA, 1966.
6. Kulish, U.W. and Miranker, W.L.: The Arithmetic of the Digital Computers: A New Approach. SIAM Review **28**, 1, 1986.
7. Moore, R., Kearfott, E. R. B., and Cloud, M. J: *Introduction to Interval Analysis.* SIAM, Philadelphia, 2009
8. Karmakar, S. and Bhunia, A. K: *A Comparative Study of Different Order Relations of Intervals.* Reliable Computing 16, 38–72., 2012.
9. Hansen, E.R. and Walster, G.W.: *Global Optimization Using Internal Analysis.* Second Edition, Marcel Dekker, Inc., New York, 2004
10. Ichino, M.: *General Metrics for Mixed Features - The Cartesian Space Theory for Pattern Recognition.* In: Proceedings of the 1988 Conference on Systems, Man, and Cybernetics. Pergamon, Oxford, 494-497, 1988
11. Johnson, S.C.: *Hierarchical Clustering Schemes*. Psychometrika, 2:241--254, 1967.