Hybrid Clustering: Combining K-Means and Interval valued data-type Hierarchical Clustering

Sérgio Mário Lins Galdino and Jornandes Dias da Silva

University of Pernambuco, Polytechnic School, Rua Benfica, 455, CEP:50720-001, Recife-PE, Brazil; galdino@poli.br; jornandesdias@poli.br

Abstract: In this paper, we describe a hybrid clustering procedure which is well-suited when we deal with a large data set. It combines the K-Means clustering to handle large data sets, and an Interval valued data-type Hierarchical Clustering (IHCA). The Hierarchical Cluster Analysis is especially helpful when we want to detect the appropriate number of clusters. The hybrid clustering procedure relies on the following schema: First, we use the K-Means algorithm in order to create pre-clusters (e.g., 30), they contain a few examples and second, we start the IHAC from these pre-clusters (summarized by interval data vectors- they contain more information than point-valued data, and such informational advantages could be exploited to yield more efficient analysis) to create the dendrogram. The main goal of this paper is show that hybrid cluster analysis is appropriate. A simple case study demonstrates the procedure for combining K-means/IHCA, which finds representative groups and thus, proves the efficiency of approach.

Keywords: hybrid clustering; interval valued data-type; hierarchical clustering; interval arithmetic; Range Euclidean metric; unsupervised machine learning

1 Introduction

Cluster analysis is a form of exploratory data in which observations are divided into different groups (or classes or clusters) that share common characteristics in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). Clustering can therefore be formulated as a multi-objective optimization problem. Cluster analysis is an unsupervised machine learning technique.

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

Two Step Cluster is an algorithm primarily designed to analyze large datasets. The two-step clustering (called also “Hybrid Clustering”) under Tanagra is described on *Tanagra - Data Mining and Data Science Tutorials*[1]. A PCA (Principal Component Analysis) computed from the original variables. This pre-treatment cleans the dataset by removing the irrelevant information such as noise, etc. In this tutorial, the approach on a large dataset with 500,000 observations and 68 variables using Tanagra 1.4.27 and R 2.7.2. The two-step clustering procedure relies on the following schema: first, the K-means algorithm is used in order to create pre-clusters (e.g., 50), they contain a few examples; second, starting the HCA from these pre-clusters to create the dendrogram.

SPSS Statistics 11.5 and later releases offer a two-step clustering method (SPSS 2001, 2004). SPSS Two Step clustering was developed by Chiu, Fang, Chen, Wang and Jeris 2001 for the analysis of large data sets. The procedure consists of two steps [1]:

Step 1) The procedure begins with the construction of a Cluster Features (CF) Tree. The tree begins by placing the first case at the root of the tree in a leaf node that contains variable information about that case. Each successive case is then added to an existing node or forms a new node, based upon its similarity to existing nodes and using the distance measure as the similarity criterion. A node that contains multiple cases contains a summary of variable information about those cases. Thus, the CF tree provides a capsule summary of the data file.

Step 2) The leaf nodes of the CF tree are then grouped using an agglomerative clustering algorithm. The agglomerative clustering can be used to produce a range of solutions. To determine which number of clusters is "best", each of these cluster solutions is compared using Schwarz's Bayesian Criterion (BIC) or the Akaike Information Criterion (AIC) as the clustering criterion.

In this paper, we deal with Hybrid Clustering. The algorithm integrates IHCA (Interval-valued data HCA) [2] and K-Means clustering algorithms to cluster. We will focus on Interval Ward Clustering (IWard)) method [3] [4]. The Range Euclidean Metric for interval-valued data is used to compare two vectors of intervals. The classical Ward method is also applied for comparison.

The rest of paper is organized as follows: Section 2 presents basic concepts of interval arithmetic and distance measures for interval data. Section 3 describes Interval Ward clustering. Section 4 presents the Hybrid Clustering case study. Section 5 provides the conclusion and discusses future work.

[1] http://data-mining-tutorials.blogspot.com/2009/06/two-step-clustering-for-handling-large.html [Accessed January 1, 2024]

2 Interval Analysis

Interval arithmetic is a method for determining absolute errors, considering all data errors and rounding [5]. Interval arithmetic makes systematic calculations through intervals $[x] = \left[\underline{x}, \overline{x}\right]$ 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. For each $x \in [x]$, and each $y \in [y]$:

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

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

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

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

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:

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

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:

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

1.1 Order Relations of Intervals

The important issue in using interval data for decision problems is the choice of an appropriate interval order relation. According to Moore et al. [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 [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*.

1.2 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 practical problem encountered in numerous applications. In special cases the exact range can be found in a straightforward way [7] [9].

1.3 Range Euclidean Distance

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

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

$$d_2([\boldsymbol{p}], [\mathbf{q}]) = d_2([\mathbf{q}], [\mathbf{p}]) = \sqrt{([q_1] - [p_1])^2 + ([q_2] - [p_2])^2 + \cdots + ([q_n] - [p_n])^2} = \sqrt{\sum_{i=1}^{n} ([q_i] - [p_i])^2} \quad (5)$$

3 Interval-valued Hierarchical Clustering

The hierarchical cluster analysis has two methods: one is the "bottom-up" agglomerative method, and the other is the "up-bottom" divisive method. The agglomerative method starts from a single point of data and merges adjacent points step by step according to a given rule until all data points are combined into one class. While the divisive method is to treat the whole data set as a whole first and partition it according to certain rules until all data points are separated from each other. These two pathways are inverse operations, and the dendrogram obtained under the same rules are the same. Because that agglomerative hierarchical clustering is more commonly used, it is used here. The steps of an agglomerative hierarchical clustering are as follows [10-16].

Consider each data point as a single-point cluster to forms N clusters;

1) Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have N-1 clusters;
2) Compute distances (similarities) between the new cluster and each of the old clusters;
3) Repeat steps 2 and 3 until all items are clustered into a single cluster of size N;
4) Draw a dendrogram;
5) Select a trim threshold to obtain the cluster classification from the dendrogram.

A Dendrogram is a type of tree diagram showing hierarchical relationships between different sets of data. It contains the memory of hierarchical clustering algorithm, so just by looking at the Dendrogram you can tell how the cluster is formed.

Note:

a) Distance between data points represents dissimilarities;
b) Height of the blocks represents the distance between clusters.

Hierarchical clustering cluster analysis (HCA), is a whole family of algorithms that differ by distance updating. The seven popular methods include Single Linkage, Complete Linkage, Simple Average (WPGMA -Weighted Pair Group Method Average), Group Average (UPGMA -Unweighted Pair Group Method Average), Median (WPGMC -Weighted Pair Group Method Centroid), Centroid(UPGMC -Unweighted Pair Group Method Centroid), and Ward's Minimum Variance Method. They are implemented in standard numerical and statistical software such as Octave, MATLAB, SciPy, Mathematica, R.

We will use The Hybrid Clustering algorithm integrates IHCA/HCA and k-means clustering algorithms to cluster interval-valued data. We will focus on Interval Ward Clustering (IWard) method [3] [4]. The Range Euclidean Metric for interval-valued data is used to compare two vectors of intervals. The classical Ward method is also applied for comparison.

Clustering algorithms is used for problem solutions in extensive practical applications across diverse domains [17-19]. Some recent clustering research contributions on clustering methods and application include [20-23].

4 Case Study: Hybrid Clustering

Clustering is related to the unsupervised learning, where we use a cluster algorithm on unlabeled data and try to form groups of similar data items.

The Iris Dataset contains four features (length and width of sepals and petals) of 50 samples of three species of Iris (Iris-setosa, Iris-virginica and Iris-versicolor). The dataset is often used in data mining, classification and clustering examples and to test algorithms. Table III shows the classical Iris Data Set.

For clustering we have no such labeled data like Iris setosa, Iris virginica and Iris versicolor species. We grouped similar irises based on Two-step algorithm to make 3 groups (8 schemas). We associate each cluster to a particular specie and the accuracy is calculated.

3.1 Hybrid Clustering on Iris dataset

The Hybrid Clustering is carried out in a two-step clustering procedure relies on the following schema:

- **First Step:** The K-Means algorithm is used in order to create pre-clusters;
- **Second Step:** IHCA/HCA from pre-clusters.

Kmeans () in R

Now, using Iris Dataset, we can perform the cluster analysis. Important note: We'll still need to drop the class attribute Information (Iris-setosa, Iris-versicolor, Iris-virginica). These are their corresponding class labels and are not useful in clustering.

We used the followings R commands:

```
set.seed(10) \#Set the seed for reproducibility

kmeans(data,centers=30,algorithm="MacQueen",iter.max=400)
```

3.2 Two Cases

- ***First Step:*** the K-means algorithm is used in order to create pre-clusters;
- ***Second Step:*** HCA 2 Cases:
  - Case 1- IWard (in Octave) - 30 representative interval valued data vector components (Scenario: K30_IW3);
  - Case 2- Ward - 30 representative vectors of mean values components (Scenario: K30_W3).

Table I shows k=30 from k-means results and $\langle \boldsymbol{C} \rangle : [min, max]$ component wise representative interval vector data.

Table I
Intervals from Iris Data Set K-means

C	**Iris sample label (types of irises)**	$\langle \boldsymbol{C} \rangle$
1	137, 149	⟨[6.2,6.3], [3.4,3.4], [5.4,5.6], [2.3,2.4]⟩
2	64, 74, 79, 92	⟨[6.0,6.1], [2.8,3.0], [4.5,4.7], [1.2,1.5]⟩
3	104, 112, 117, 129, 133, 138	⟨[6.3,6.5], [2.7,3.1], [5.3,5.6], [1.8,2.2]⟩
4	62, 72, 75, 98	⟨[5.9,6.4], [2.8,3], [4.0,4.3], [1.3,1.5]⟩
5	63, 69, 88	⟨[6.0, 6.3], [2.2, 2.3], [4.0, 4.5], [1, 1.5]⟩
6	15, 16,19, 34	⟨[5.5, 5.8], [3.8, 4.4], [1.2, 1.7], [0.2, 0.4]⟩
7	102, 114, 143	⟨[5.7, 5.8], [2.5, 2.7], [5.0, 5.1], [1.9, 2.0]⟩
8	111, 116, 142, 146, 148	⟨[6.40,6.9], [3.0,3.2], [5.1,5.3], [2.0,2.3]⟩
9	24, 27, 44, 45	⟨[5.0,5.1], [3.3,3.8], [1.6,1.9], [0.4,0.6]⟩
10	9, 14, 39, 42	⟨[4.3,4.5], [2.3,3.0], [1.1,1.4], [0.1,0.3]⟩
11	54, 56, 60, 65, 67, 68, 70, 83, 85, 89, 90, 91, 93, 95, 96, 97, 100, 107	⟨[4.9,5.8], [2.3,3.0], [3.6,4.5], [1.0,1.7]⟩
12	106, 119, 123, 136	⟨[7.6,7.7], [2.6,3], [6.1,6.9], [2,2.3]⟩
13	110, 118, 132	⟨[7.2,7.9], [3.6,3.8], [6.1,6.7], [2,2.5]⟩
14	3, 7, 23, 43, 48	⟨[4.4,4.7], [3.2,3.6], [1.0,1.4], [0.2,0.3]⟩
15	52, 57, 86	⟨[6.0,6.4], [3.2,3.4], [4.5,4.7], [1.5,1.6]⟩
16	58, 61, 80, 81, 82, 94, 99	⟨[4.9,5.7], [2.0,2.6], [3.0,3.8], [1.0,1.1]⟩
17	11, 21, 28, 32, 37, 49	⟨[5.2,5.5], [3.4,3.7], [1.3,1.7], [0.2,0.4]⟩
18	73, 84, 120, 134, 135, 147	⟨[6.0,6.3], [2.2,2.8], [4.9,5.6], [1.4,1.9]⟩
19	105, 113, 121, 125, 140, 141, 144, 145	⟨[6.5,6.9], [3.0,3.3], [5.4,5.9], [2.1,2.5]⟩
20	51, 53, 55, 59, 66, 76, 77, 78, 87	⟨[6.5,7.0], [2.8,3.2], [4.4,5.0], [1.3,1.7]⟩
21	1, 5, 8, 18, 29, 36. 40, 41, 50	⟨[5, 5.2], [3.2, 3.6], [1.2, 1.5], [0.2, 0.3]⟩

22	109	⟨[6.7,6.7], [2.5,2.5], [5.8,5.8], [1.8,1.8]⟩
23	2, 4, 10, 13, 26, 31, 35, 38, 46	⟨[4.6,5], [3,3.1], [1.4,1.6], [0.1,0.3]⟩
24	101	⟨[6.3,6.3], [3.3,3.3], [6.0,6.0], [2.5,2.5]⟩
25	115, 122	⟨[5.6,5.8], [2.8,2.8], [4.9,5.1], [2,2.4]⟩
26	6, 17, 20, 22, 33, 47	⟨[5.1,5.4], [3.7,4.1], [1.3,1.7], [0.1,0.4]⟩
27	150	⟨[5.9,5.9], [3.0,3.0], [5.1,5.1], [1.8,1.8]⟩
28	71, 124, 127, 128, 139	⟨[5.9,6.3], [2.7,3.2], [4.8,4.9], [1.8,1.8]⟩
29	103, 108, 126, 130, 131	⟨[7.1,7.4], [2.8,3.2], [5.8,6.3], [1.6,2.1]⟩
30	12, 25, 30	⟨[4.7,4.8], [3.2,3.4], [1.6,1.9], [0.2,0.2]⟩

C - K-Means cluster label

Types of irises = Iris-setosa (1-50), Iris-versicolor (51-100), Iris-virginica (101-150)

⟨C⟩:[min, max] component wise representative data

In sequel we focus on the relative impact of each one HCA method choice.

Table II shows the accuracy of 2 cases Hybrid clustering results: K30_IW3 - starting from 30 representative interval valued data vector components we get from IWard clustering result with 90% of accuracy. K30_W3 - 30 representative vectors of mean values components we get from Ward clustering result with 90% of accuracy. Overall, we see a good agreement between K30_IW3 and K30_W3 accuracy.

It should be highlighted that the Case 2 using IWard (we observed similar results and lesser processing) can allow for a reliable analysis of clustering results because we can track the intervals on second step, by means of interval valued data dendrogram produced.

Table 2

K-means (30 clusters first step) → Ward (3 clusters second step)

Scenario	C	**Iris data label (wrong types of irisis)**	**Accuracy**
K30_IW3	1	1, 3, 8, 12, 13, 19, 22, 24	0.90
	2	2, 4, 5, 7(3), 11(1), 15, 16, 18(4), 20, 25(2), 27(1), 28(4)	
	3	6, 9, 10, 14, 17, 21, 23, 26, 30(3)	
K30_W3	1	1, 3, 8, 12, 13, 19, 22, 24, 29	0.90
	2	2, 4, 5, 7(3), 11(1), 15, 16, 18(4), 20, 25(2), 27(1), 28(4)	
	3	6, 9, 10, 14, 17, 21, 23, 26, 30	

C = (1- Iris-virginica, 2- Iris-versicolor, 3- Iris-setosa)

Conclusions

This paper investigated the hybrid clustering (two-step) method, designed to cluster large data sets. In second step we introduce Case 2 with IWard, for clustering Interval-valued Data, based on the Range Euclidean Metric.

Hybrid Clustering using Interval-valued Data information related to representative interval vector with the uncertainty of input data. Basically, it allows comparison and quantitative judgement over data granulation, with different pre-clusters (e.g., 10, 20, 30, 40, and 50). We can then apply the procedure, by performing experiments on several different datasets - including very large data sets, with known cluster patterns, in order to better understand the technique.

Future studies may use other IHCA cluster methods, for interval-valued data. We also have the ongoing K-[Means] (K-Means tailored for interval valued data) to acquire a Hybrid Clustering method, for interval-valued data. In addition, there are ongoing Hybrid Clustering methods, for interval-valued data that are tailored for Interval-Valued Data.

References

[1] Chiu,T. Fang, D. Chen, J. Wang, Y. and Jeris, C. *A Robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment.* In Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2001, pp. 263-268, 2001

[2] Galdino, S. M. L. *Interval-valued Data Clustering Based on the Range City Block Metric.* In: SMC 2016, 2016, Budapest. The 2016 IEEE International Conference on Systems, Man, and Cybernetics, 2016

[3] Galdino, Sérgio and Dias, Jornandes. *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, pp. 1-6

[4] Dias, Jornandes and Galdino, Sérgio. *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, pp. 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. E. Kearfott, 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] Johnson, S. C. *Hierarchical Clustering Schemes.* Psychometrika, 2:241-254, 1967

[11] Jain, A. K. and Dubes, R. C. *Algorithms for clustering data.* Prentice-Hall, Inc., 1988

[12] Jain, A. Murty, K. M. N. and Flynn, P. J. *Data clustering: a review.* ACM computing surveys (CSUR) 31(3), 264-323, 1999

[13] Kaufman, L. and Rousseeuw, P. J. *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley-Interscience, 2005

[14] Gan, G. Ma, C. and Wu, J. *Data Clustering: Theory, Algorithms, and Applications.* SIAM, 2007

[15] Everitt, B. S. Sabine Landau, S. Leese, M. and Stahl, D. *Cluster analysis*. John Wiley & Sons, 2011

[16] Aggarwal, C. C. and Reddy, C. K. *Data Clustering: Algorithms and Applications.* Chapman & Hall/CRC, 1st edition, 2013

[17] Karmenova, Markhaba. et al. *An approach for clustering of seismic events using unsupervised machine learning.* Acta Polytechnica Hungarica 19.5 (2022): 7-22

[18] Kolnhofer-Derecskei, A. Reicher, Regina Zs. and Szeghegyi, A. *The X and Y generations*' characteristics comparison. Acta Polytechnica Hungarica 14.8 (2017): 107-125

[19] Phung, L. X. Nguyen, T. K. and Truong, S. H. *The Enhancement of the Overall Group Technology Efficacy using Clustering Algorithm for Cell Formation.* Acta Polytechnica Hungarica 21.2 (2024)

[20] Guo, L. Qin, W. Cai, Z. * and Xing Su, X. *Hybrid Clustering Algorithm Based on Improved Density Peak Clustering.* Appl. Sci. 14, 715, 2024, https://doi.org/10.3390/app14020715

[21] Rizalde, A. R. Mubarak, H. A. Ramadhan, G. and Fatan, M.A. *Comparison of K-Means, BIRCH and Hierarchical Clustering Algorithms in Clustering OCD Symptom Data.* Public Research Journal of Engineering, Data Technology and Computer Science:PREDATECS, Vol. 1, No. 2, January 2024, https://doi.org/10.57152/predatecs.v1i2.1106

[22] Muhammad, Izhar. et al. *Clustering Fake News with K-Means and Agglomerative Clustering Based on Word2Vec.* International Journal of Mathematics and Computer Research, Volume 12, Issue 02, pp. 3999-4007, February 2024

[23] Han, W. Zhang, S. Gao, H. and Bu, D. *Clustering on hierarchical heterogeneous data with prior pairwise relationships*. BMC Bioinformatics 25, 40, January 2024, https://doi.org/10.1186/s12859-024-05652-6

[24] Kelly, M. Longjohn, R. and Nottingham, K. UCI Machine Learning Repository, https://archive.ics.uci.edu [Accessed January 2, 2024]

Table 3
Iris Data Set [24]

S	Iris-setosa Values(cm)	S	Iris-versicolor Values(cm)	S	Iris-virginica Values(cm)
1	5.10 3.50 1.40 0.20	51	7.00 3.20 4.70 1.40	101	6.30 3.30 6.00 2.50
2	4.90 3.00 1.40 0.20	52	6.40 3.20 4.50 1.50	102	5.80 2.70 5.10 1.90
3	4.70 3.20 1.30 0.20	53	6.90 3.10 4.90 1.50	103	7.10 3.00 5.90 2.10
4	4.60 3.10 1.50 0.20	54	5.50 2.30 4.00 1.30	104	6.30 2.90 5.60 1.80
5	5.00 3.60 1.40 0.20	55	6.50 2.80 4.60 1.50	105	6.50 3.00 5.80 2.20
6	5.40 3.90 1.70 0.40	56	5.70 2.80 4.50 1.30	106	7.60 3.00 6.60 2.10
7	4.60 3.40 1.40 0.30	57	6.30 3.30 4.70 1.60	107	4.90 2.50 4.50 1.70
8	5.00 3.40 1.50 0.20	58	4.90 2.40 3.30 1.00	108	7.30 2.90 6.30 1.80
9	4.40 2.90 1.40 0.20	59	6.60 2.90 4.60 1.30	109	6.70 2.50 5.80 1.80
10	4.90 3.10 1.50 0.10	60	5.20 2.70 3.90 1.40	110	7.20 3.60 6.10 2.50
11	5.40 3.70 1.50 0.20	61	5.00 2.00 3.50 1.00	111	6.50 3.20 5.10 2.00
12	4.80 3.40 1.60 0.20	62	5.90 3.00 4.20 1.50	112	6.40 2.70 5.30 1.90
13	4.80 3.00 1.40 0.10	63	6.00 2.20 4.00 1.00	113	6.80 3.00 5.50 2.10
14	4.30 3.00 1.10 0.10	64	6.10 2.90 4.70 1.40	114	5.70 2.50 5.00 2.00
15	5.80 4.00 1.20 0.20	65	5.60 2.90 3.60 1.30	115	5.80 2.80 5.10 2.40
16	5.70 4.40 1.50 0.40	66	6.70 3.10 4.40 1.40	116	6.40 3.20 5.30 2.30
17	5.40 3.90 1.30 0.40	67	5.60 3.00 4.50 1.50	117	6.50 3.00 5.50 1.80
18	5.10 3.50 1.40 0.30	68	5.80 2.70 4.10 1.00	118	7.70 3.80 6.70 2.20
19	5.70 3.80 1.70 0.30	69	6.20 2.20 4.50 1.50	119	7.70 2.60 6.90 2.30
20	5.10 3.80 1.50 0.30	70	5.60 2.50 3.90 1.10	120	6.00 2.20 5.00 1.50
21	5.40 3.40 1.70 0.20	71	5.90 3.20 4.80 1.80	121	6.90 3.20 5.70 2.30
22	5.10 3.70 1.50 0.40	72	6.10 2.80 4.00 1.30	122	5.60 2.80 4.90 2.00
23	4.60 3.60 1.00 0.20	73	6.30 2.50 4.90 1.50	123	7.70 2.80 6.70 2.00
24	5.10 3.30 1.70 0.50	74	6.10 2.80 4.70 1.20	124	6.30 2.70 4.90 1.80
25	4.80 3.40 1.90 0.20	75	6.40 2.90 4.30 1.30	125	6.70 3.30 5.70 2.10
26	5.00 3.00 1.60 0.20	76	6.60 3.00 4.40 1.40	126	7.20 3.20 6.00 1.80
27	5.00 3.40 1.60 0.40	77	6.80 2.80 4.80 1.40	127	6.20 2.80 4.80 1.80
28	5.20 3.50 1.50 0.20	78	6.70 3.00 5.00 1.70	128	6.10 3.00 4.90 1.80
29	5.20 3.40 1.40 0.20	79	6.00 2.90 4.50 1.50	129	6.40 2.80 5.60 2.10

30	4.70 3.20 1.60 0.20	80	5.70 2.60 3.50 1.00	130	7.20 3.00 5.80 1.60
31	4.80 3.10 1.60 0.20	81	5.50 2.40 3.80 1.10	131	7.40 2.80 6.10 1.90
32	5.40 3.40 1.50 0.40	82	5.50 2.40 3.70 1.00	132	7.90 3.80 6.40 2.00
33	5.20 4.10 1.50 0.10	83	5.80 2.70 3.90 1.20	133	6.40 2.80 5.60 2.20
34	5.50 4.20 1.40 0.20	84	6.00 2.70 5.10 1.60	134	6.30 2.80 5.10 1.50
35	4.90 3.10 1.50 0.10	85	5.40 3.00 4.50 1.50	135	6.10 2.60 5.60 1.40
36	5.00 3.20 1.20 0.20	86	6.00 3.40 4.50 1.60	136	7.70 3.00 6.10 2.30
37	5.50 3.50 1.30 0.20	87	6.70 3.10 4.70 1.50	137	6.30 3.40 5.60 2.40
38	4.90 3.10 1.50 0.10	88	6.30 2.30 4.40 1.30	138	6.40 3.10 5.50 1.80
39	4.40 3.00 1.30 0.20	89	5.60 3.00 4.10 1.30	139	6.00 3.00 4.80 1.80
40	5.10 3.40 1.50 0.20	90	5.50 2.50 4.00 1.30	140	6.90 3.10 5.40 2.10
41	5.00 3.50 1.30 0.30	91	5.50 2.60 4.40 1.20	141	6.70 3.10 5.60 2.40
42	4.50 2.30 1.30 0.30	92	6.10 3.00 4.60 1.40	142	6.90 3.10 5.10 2.30
43	4.40 3.20 1.30 0.20	93	5.80 2.60 4.00 1.20	143	5.80 2.70 5.10 1.90
44	5.00 3.50 1.60 0.60	94	5.00 2.30 3.30 1.00	144	6.80 3.20 5.90 2.30
45	5.10 3.80 1.90 0.40	95	5.60 2.70 4.20 1.30	145	6.70 3.30 5.70 2.50
46	4.80 3.00 1.40 0.30	96	5.70 3.00 4.20 1.20	146	6.70 3.00 5.20 2.30
47	5.10 3.80 1.60 0.20	97	5.70 2.90 4.20 1.30	147	6.30 2.50 5.00 1.90
48	4.60 3.20 1.40 0.20	98	6.20 2.90 4.30 1.30	148	6.50 3.00 5.20 2.00
49	5.30 3.70 1.50 0.20	99	5.10 2.50 3.00 1.10	149	6.20 3.40 5.40 2.30
50	5.00 3.30 1.40 0.20	100	5.70 2.80 4.10 1.30	150	5.90 3.00 5.10 1.80

S = sample label

Values = (sepal length, sepal width, petal length, petal width)