# Neutrosophic Hierarchical Clustering: A Novel Approach for Handling Uncertainty in Multi-Level Data Organization

**Sitikantha Mallik[1,*], Suneeta Mohanty[1], Bhabani Shankar Prasad Mishra[1]**
[1]School of Computer Engineering, KIIT Deemed to be University, Bhubaneswar, India
Emails: sitikanthamallik@gmail.com; smohantyfcs@kiit.ac.in; bsmishrafcs@kiit.ac.in

**Abstract**

The most important stage of data mining is clustering. Several distinct clustering approaches like grid-based, density-based, partitioning, graph-based, model-based, and hierarchical clustering are used for cluster analysis. We can cluster data objects into hierarchical trees by using the hierarchical clustering approach. Hierarchical clustering, with its agglomerative and divisive types, uses nodes to represent clusters. Agglomerative clustering is favored, and high-quality clusters are essential for successful cluster analysis. Up to this point, numerous alternatives to the clustering technique have been proposed, including the fuzzy k-mean approach. The uncertainty resulting from numerical variations or unpredictable natural occurrences may be handled by any data mining techniques now in use. However, indeterminacy components may be present in current data mining challenges in real-world scenarios. Neutrosophic logic, applicable in various sectors, is gaining traction due to its efficiency and accuracy, attracting investment for its potential to improve human lives. The suggested approach outperforms current methods like fuzzy logic and k-means in its ability to forecast the number of clusters.

**Keywords:** Indeterminacy; Clustering; Hierarchical clustering algorithm; Uncertainty; Silhouette coefficient

## 1. Introduction

Clustering group's similar data points together. However, uncertainty and indeterminacy can make clustering multi-level data challenging. Multi-level data typically refers to data with hierarchical or nested structures, where observations are grouped into various levels of granularity. For instance, in social networks, individuals may be grouped into communities, which are further organized into larger groups or networks. Similarly, in biological systems, genes may be grouped into pathways, which are then organized into biological processes [1].

Uncertainty and indeterminacy arise in multi-level data due to various factors:

1. **Data representation ambiguity**: Multi-level data may be represented differently, i.e., as graphs, trees, nested structures, etc., resulting in ambiguity in calibrating a measure of similarity and defining cluster boundaries.
2. **Incomplete or noisy data**: Multi-level data often contains missing entries, noise, or inconsistencies in the developmental process; this imprecision can significantly pertain to clustering.
3. **Overlapping clusters**: Data points may be parts of several clusters simultaneously, causing the misinterpretation of the boundaries or creating fuzzy patterns.
4. **Varying levels of granularity**: Multi-level data possesses multiple levels of granularity, which contributes to the uncertainty regarding performing clustering on a specific scale.

To cope with these obstacles, the clustering methods on multi-level data with uncertainty and indeterminacy conditions have to accomplish the following features:

- **Incorporating probabilistic models:** Bayesian methods and probabilistic graphical models can explicitly model uncertainty in the clustering process, allowing for more flexible and robust cluster assignments.



DOI: https://doi.org/10.54216/IJNS.260121

- **Fuzzy clustering:** Fuzzy clustering algorithms assign data points to clusters with degrees of membership, allowing for overlapping cluster boundaries and capturing the inherent ambiguity in multi-level data.

- **Hierarchical clustering:** Hierarchical clustering methods recursively partition the data into a hierarchy of clusters, accommodating varying levels of granularity and capturing the nested structure of multi-level data.

- **Ensemble clustering:** Ensemble methods combine multiple clustering algorithms to improve robustness and handle uncertainty in multi-level data.

Overall, clustering in the context of handling uncertainty and indeterminacy in multi-level data requires sophisticated algorithms that can capture the inherent complexity and hierarchical structure of the data while providing meaningful and interpretable clustering.

In the context of hierarchical clustering, there are several motivations for using neutrosophic logic:

1. **Capturing Ambiguity and Vagueness**: In hierarchical clustering, where data may exhibit varying levels of similarity and dissimilarity, neutrosophic logic can provide a more nuanced representation of cluster memberships [2].
2. **Dealing with Overlapping Cluster Boundaries**: Neutrosophic logic enables overlapping cluster boundaries, in contrast to crisp clustering techniques that allocate each data point to a single cluster.
3. **Handling Incomplete or Noisy Data**: Neutrosophic logic can effectively handle incomplete or noisy data by allowing for the representation of partial memberships and indeterminate values...
4. **Expressing Indeterminacy in Cluster Hierarchies**: Hierarchical clustering inherently involves decisions about the appropriate level of granularity at which to partition the data. Neutrosophic logic allows for the representation of indeterminacy in these decisions, enabling the construction of cluster hierarchies that capture the inherent uncertainty in the data.
5. **Integrating Multiple Sources of Information**: A framework for combining data from various sources, such as linguistic, qualitative, and quantitative sources, is provided by neotrophic logic. In hierarchical clustering, neutrosophic logic helps to integrate heterogeneous data types into a single clustering framework, even though the data might start from different sources and modalities.

Overall, the motivation for using neutrosophic logic in hierarchical clustering lies in its ability to handle the inherent uncertainty, indeterminacy, and imprecision in multi-level datasets, thereby enabling more robust and flexible cluster analysis. By providing a more expressive representation of cluster memberships and cluster hierarchies, neutrosophic logic enhances the effectiveness of hierarchical clustering algorithms in real-world applications [3].

The rest of the paper is organized as follows: section 2 illustrates clustering concepts and clustering indices. Section 3 provides details about the role of neutrosophic clustering. "Section 4 introduces the proposed approach. Section 5 implies about experimental results and performance of proposed models over existing models. Section 6 summarizes the conclusion and future work.

## 2. Related Work

Clustering is an unsupervised learning task that identifies patterns in unlabeled data. Items within a cluster share similar features, while differing from items in other clusters [4].

### 2.1. Types of clustering algorithm

#### 2.1.1. K-means clustering

Standard k-means clustering is a method that groups data points based on their distance to the nearest centroid, aiming to minimize the overall distance between data points and their assigned centroids.

#### 2.1.2. Fuzzy clustering

Fuzzy k-means clustering allows data points to partially belong to multiple clusters based on their distance to cluster centers. This creates overlapping clusters as membership degrees indicate the likelihood of belonging to each cluster.

#### 2.1.3. Hierarchical clustering

One method for organizing data points based on similarities is called hierarchical clustering. The method is to combine the two data points that are most comparable to each other and figure out which two are nearest to each other. This process is repeated until all data points are sorted into clusters, producing a hierarchical tree of comparable groups known as a dendrogram. Because the number of clusters need not be predetermined, it is appropriate for exploratory data



DOI: https://doi.org/10.54216/IJNS.260121

analysis. Furthermore, the dendrogram created by hierarchical clustering allows users to visually examine and comprehend the relationships between clusters, offering insights into the data's hierarchical structure [5].

### 2.2. K-means Clustering

K-means is an unsupervised learning technique that uses a fixed number of clusters (k) to classify data, aiming to minimize a squared error function [6].

#### 2.2.1. Algorithm

The phases of the algorithm are explained:

1. Set up K pointers in the area containing the objects to be grouped. These points indicate the first set of centroids.
2. The closest centroid should be used to determine where each item belongs in the group.
3. **Update Centroids**: Recalculate cluster centroids based on the average position of their assigned data points.
4. **Iteration**: Iterate assignment and centroid updates until centroids stabilize or the maximum iterations are reached.
5. **Finalization**: Once convergence is achieved, each data point will be assigned to one of the K clusters, and the cluster centroids will represent the final cluster centers.

#### 2.2.2. Limitations of K-means:

- **Assumes spherical clusters and requires the number of clusters to be specified**: This can lead to suboptimal results when clusters are non-spherical or have varying sizes.
- **Sensitive to initial cluster centers and outliers**: Random initialization of centroids can lead to different clustering results, and outliers can significantly affect cluster centroids.
- **Cannot handle uncertainty or overlapping clusters**: K-means assigns each point to exactly one cluster, making it unsuitable for datasets with overlapping clusters or where data points may belong to multiple clusters.

### 2.3 Fuzzy Clustering

Fuzzy clustering lets data points belong to multiple clusters simultaneously, with different levels of membership, in contrast to conventional clustering techniques that allocate every data point to a single cluster.[7]

#### 2.3.1. Fuzzy k-means algorithm

The steps of the fuzzy k-means algorithm are as follows:

1. Specify the number of clusters (kk) and the fuzzification parameter (mm).
2. Initialize cluster centroids randomly.
3. Compute membership values for each data point in every cluster using the following formula:

$$u_{ij} = \left(\frac{\sum_{i=1}^{n} (u_{ij})^m x_i}{\sum_{i=1}^{n} (u_{ij})^m}\right) \quad (1)$$

where, $u_{ij}$ is membership grade of data point i for cluster j

$d_{ij}$ is Euclidean distance between data point i and cluster centroid j

$d_{ik}$ is the Euclidean distance between data point i and cluster centroid k, and m is the fuzzy exponent

4. Update the centroids for each cluster using the following formula:

$$u_{ij} = \left(\frac{1}{\sum_{k=1}^{k} \left(\frac{d_{ij}}{d_{ik}}\right)^{2/m-1}}\right)^{-1} \quad (2)$$

where $c_j$ represents the updated centroid for cluster j, $x_i$ denotes data point i, and n is the total number of data points.

5. Continue iterating through steps 3 and 4 until the algorithm converges or the maximum number of iterations is reached.
6. Finally, output the identified clusters along with the membership grades for each data point.



DOI: https://doi.org/10.54216/IJNS.260121

#### 2.3.2. Limitations

- **Requires a priori specification of the number of clusters**: Like K-means, fuzzy clustering requires specifying the number of clusters in advance, which can be challenging.
- **May converge to suboptimal solutions depending on initialization**: Initialization can significantly affect the final clustering result, potentially leading to suboptimal solutions.
- **Assumes Gaussian distributions and does not handle arbitrary cluster shapes well:** Fuzzy clustering is assumption of Gaussian distributions might not hold true for all datasets, and it may struggle with non-Gaussian or irregularly shaped clusters.

### 2.4. Hierarchical Clustering

Hierarchical clustering groups similar objects into clusters using a tree-like structure (dendrogram). It iteratively merges or splits clusters based on their similarity until a stopping point is reached.

#### 2.4.1. Types of hierarchical clustering

Hierarchical clustering has two primary methods: agglomerative and divisive.

**1. Agglomerative Hierarchical Clustering**

- Initially, each data point forms its own individual cluster.
- At each iteration, the two most similar clusters are merged into a single cluster, reducing the total number of clusters by one.
- The process iterates until everything converges into one cluster, creating a dendrogram visualizing the cluster hierarchy.
- The choice of similarity or dissimilarity measure and the method used to compute it (e.g., single linkage, complete linkage, and average linkage) determine the resulting clusters.

**2. Divisive Hierarchical Clustering**

- Divisive hierarchical clustering starts with all data points belonging to a single cluster.
- At each iteration, the algorithm identifies the cluster that is least homogeneous or most dissimilar and splits it into two smaller clusters.
- The process continues recursively until each data point is in its own cluster, resulting in a dendrogram.

#### 2.4.2. Hierarchical clustering algorithm

The hierarchical clustering algorithm proceeds as follows:

1. **Initialization:** Treat each data point as an individual cluster**.**
2. **Compute Pairwise Distances:** Calculate the distance between each pair of clusters or data points using metrics such as Euclidean distance, Manhattan distance, etc.
3. **Merge Closest Clusters:** Identify the two closest clusters based on the computed distances and merge them into a single cluster. Repeat this step until all data points form a single cluster or a predefined stopping criterion is met.
4. **Update Distance Matrix:** After merging clusters, update the distance matrix to reflect the revised distances between clusters.
5. **Repeat:** Continue steps 2–4 until the desired number of clusters is achieved or the stopping condition is satisfied.
6. **Output:** The final output is typically represented as a dendrogram, a tree-like diagram illustrating the hierarchical arrangement of clusters.



DOI: https://doi.org/10.54216/IJNS.260121

### 2.4.3. Limitations of Hierarchical Clustering

- **Produces a static hierarchy, which may not be suitable for all datasets**: The hierarchical structure may not capture all the nuances of the data, and a static hierarchy might not be appropriate for dynamic datasets.
- **Computationally expensive for large datasets**: As the number of data points increases, the computational complexity of hierarchical clustering grows significantly.
- **Often relies on distance-based metrics, which may not adequately capture uncertainty or overlapping clusters**: Similar to K-Means, hierarchical clustering struggles with non-spherical clusters and overlapping clusters.

## 3. Clustering using Neutrosophic Algorithm

Computational Intelligence (C.I.) is a subfield of artificial intelligence (A.I.) that encompasses various techniques and methodologies designed to enable machines to mimic human intelligence in solving complex problems. It focuses on developing algorithms and models inspired by natural intelligence, including neural networks, evolutionary algorithms, and fuzzy systems.

An extension of classical logic, neutrosophic logic enables the handling and representation of ambiguous, partial, and inconsistent data. The idea of truth-values is expanded by neutrosophic logic to encompass a third value known as "indeterminate" or "neither true nor false," in addition to "true" and "false." [8].

Neutrosophic logic offers several advantages that make it a valuable and unique tool for handling uncertainty, ambiguity, and imprecision in information and decision-making. Some of the key benefits of neutrosophic logic include:

1. **Representation of Indeterminacy:** By introducing an additional truth-value, "indeterminate," neutrosophic logic makes it possible to accurately represent situations in which information is neither entirely true nor incorrect. It recognizes that ambiguity and uncertainty exist in real-world situations, something that classical logic frequently ignores.
2. **Handling Contradictions:** When contradictory data coexists, neutrosophic logic can handle the problem, allowing for a more flexible and nuanced approach to decision-making. It allows for circumstances in which both features may coexist to some extent rather than imposing an option between true and false.[9]
3. **Fuzzy Set and Interval Set Integration:** By using neutrosophic logic to connect fuzzy sets and interval sets, data can be better represented by accounting for both membership and non-membership degrees.

Degrees of truth, falsehood, and indeterminacy can be associated with propositions, variables, and sets in neutrosophic logic. This makes it possible for representation and reasoning to be more flexible, especially in circumstances when conventional logic could find it difficult to convey the underlying ambiguity or uncertainty [10].

### 3.1. Neutrosophic Dissimilarity Measure:

The neutrosophic dissimilarity measure is used to quantify the distance or dissimilarity between neutrosophic elements, which are represented using three membership functions: truth (T), indeterminacy (I), and falsity (F). One commonly used neutrosophic dissimilarity measure is based on the Euclidean distance between neutrosophic elements.

Neutrosophic dissimilarity measures are used in situations where there is uncertainty or indeterminacy associated with the data. Here is a definition for a neutrosophic dissimilarity measure that considers both indeterminacy and distance between neutrosophic elements [11]:

Let $A = (a, a\tilde{}, a\hat{})$ and $B = (b, b\tilde{}, b\hat{})$ be two neutrosophic elements, where $a, a\tilde{}, a\hat{}$ and $b, b\tilde{}, b\hat{}$ represent the truth-membership, indeterminacy-membership, and falsity-membership degrees, respectively.

1. **Euclidean Distance Calculation**: Calculate the Euclidean distance between the truth-membership, indeterminacy-membership, and falsity-membership degrees of the two neutrosophic elements:

$$d_{truth} = \sqrt{(a-b)^2}, d_{indeterminacy} = \sqrt{(a\tilde{}-b\tilde{})^2}, d_{falsity} = \sqrt{(a\hat{}-b\hat{})^2}$$

2. **Normalization**: Normalize the distances to ensure they are within the range $[0,1]$ :

$$d'_{truth} = \frac{d_{truth}}{max(a,b)}, d'_{indeterminacy} = \frac{d_{indeterminacy}}{max(a\tilde{},b\tilde{})}, d'_{falsity} = \frac{d_{falsity}}{max(a\hat{},b\hat{})}$$



DOI: https://doi.org/10.54216/IJNS.260121

3. **Combination with Indeterminacy:** Since indeterminacy reflects the uncertainty, we need to weigh the truth and falsity distances based on the indeterminacy:

$$d_{weighted\_truth} = (1 - \tilde{a}) \cdot d'_{truth}, \; d_{weighted\_falsity} = (1 - \tilde{a}) \cdot d'_{falsity}$$

4. **Overall Dissimilarity Measure:** Combine the weighted truth and falsity distances:

$$d_{overall} = \alpha \cdot d_{weighted\_truth} + (1 - \alpha) \cdot d_{weighted\_falsity} \quad (3)$$

where $\alpha$ is a parameter to balance between truth and falsity distances, usually set between 0 and 1.

This neutrosophic dissimilarity measure incorporates both the indeterminacy and the distances between the truth-membership, indeterminacy-membership, and falsity-membership degrees of neutrosophic elements, providing a comprehensive assessment of dissimilarity for clustering purposes. Adjusting the parameter $\alpha$ allows for customization based on the specific characteristics of the data.

### 3.2. Clustering Evaluation Metrics

Clustering indices, also known as cluster validity measures or clustering evaluation metrics, are quantitative measures used to assess the quality of clustering results. They provide a way to objectively evaluate how well a clustering algorithm has partitioned the data into clusters.

Here are some commonly used clustering indices:

#### 3.2.1 PBM-index

PBM (Privacy, Bias and Model Performance) Index is a composite metric used to evaluate machine-learning models in terms of their privacy preservation, mitigation of bias, and overall performance.

It consists of mainly three components:

1. **Privacy**: Measures the extent to which the model preserves the privacy of sensitive information in the dataset. This includes considerations such as data anonymization, differential privacy techniques, and compliance with privacy regulations.

2. **Bias**: Assesses the presence of bias in the model's predictions or decisions, particularly regarding protected attributes such as race, gender, or age. Techniques to address bias include fairness-aware algorithms, bias detection, and mitigation strategies.

3. **Model Performance:** Evaluates the overall effectiveness of the model in terms of accuracy, precision, recall, F1 score, or other relevant performance metrics. This component ensures that the model meets the desired objectives while maintaining privacy and fairness.

The PBM index is described as follows:

$$PBM(K) = \left(\frac{1}{K} \times \frac{E_1}{E_K} \times D_K\right)^2 \quad (4)$$

such that

$$E_K = \sum_{k=1}^{K} E_k \;, \quad (5)$$

$$E_k = \sum_{j=1}^{n} u_{kj} \parallel x_j - z_k \parallel \quad (6)$$

and

$$D_K = max^K{}_{i,j=1} \parallel z_i - z_j \parallel \quad (7)$$

where $E_k$ is the total distance from each cluster's points to their barycenter

$E_1$ is the total distance from each point in the data set to its barycenter

k is the number of clusters

#### 3.2.2 Silhouette Score

Silhouette Score = (y-x) / max (y, x)

where, x = average intra-cluster distance i.e., the average distance between each point within a cluster

y = average inter-cluster distance i.e., the average distance between all clusters



DOI: https://doi.org/10.54216/IJNS.260121

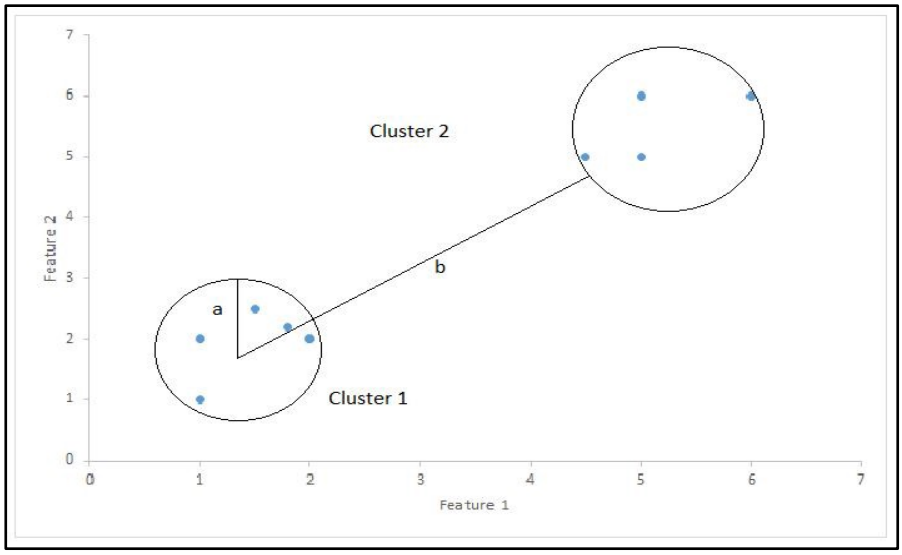


**Figure 1.** Silhouette Score

The silhouette coefficient is determined by taking the mean of the nearest cluster and intra-cluster distances for each sample. The silhouette coefficient has a range of [-1,1]. The closer the silhouette coefficients are to +1, the more separation there is between the clusters. Similarly, values near zero denote overlap or poorly separated clusters. Samples with a negative one, on the other hand, might have been inadvertently assigned to a different cluster [12].

### 3.2.3. Davies Bouldin index

Davies Bouldin index is based on two fundamental distances: the between and with-cluster distances The ratio of within-cluster to between-cluster lengths is used to assess proximity, and the distances between each cluster and its nearest neighbor are averaged to determine the score. Clusters that have less dispersion and greater space between them will therefore score higher. A lower DB Index indicates better clustering quality, with the optimal value being 0.[13]

$$\delta_k = \frac{1}{n_k}\sum_{i\in I_k} \quad \| M_i^{\{k\}} - G^{\{k\}} \| \tag{8}$$

$$\Delta_{kk'} = d\left(G^{\{k\}}, G^{\{k'\}}\right) = \| G^{\{k'\}} - G^{\{k\}} \| \tag{9}$$

Where,

$\delta_k$ = mean distance of points belonging to cluster $C_k$ to their barycenter, $G^k$

$\Delta_{kk'}$ = distance between the barycenters $G_k$ and $G_k{}'$ of clusters $C_k$ and $C_k{}'$

The mean value of all the clusters for the quantities, $M_k$ is calculated in DB index as seen in [14, eq. (10)]. DB Index is denoted by $C$.

$$C = \frac{1}{K}\sum_{k=1}^{K} \quad M_k = \frac{1}{K}\sum_{k=1}^{K} \quad \max_{k'\neq k}\left(\frac{\delta_k+\delta_{k'}}{\Delta_{kk'}}\right) \tag{10}$$

## 4. Neutrosophic Hierarchical Clustering Algorithm

Agglomerative Neutrosophic Hierarchical Clustering (ANHC) is an extension of the agglomerative hierarchical clustering algorithm to handle neutrosophic data. It aims to group neutrosophic elements into clusters in a hierarchical manner, capturing the inherent uncertainty and indeterminacy present in the data. Here is an overview of the process and dendrogram construction in ANHC:

**Agglomerative Neutrosophic Hierarchical Clustering Process:**

1. **Initialization**:

- Start with each neutrosophic element as its own singleton cluster.

$$C_i = \{x_i\}, i = 1,2,\dots,n$$

where $C_i$ represents the $i$-th cluster containing the $i$-th neutrosophic element $x_i$.



DOI: https://doi.org/10.54216/IJNS.260121

**2. Compute Neutrosophic Dissimilarity**:

Calculate the neutrosophic dissimilarity between all pairs of clusters using a suitable neutrosophic dissimilarity measure:

$$D(C_i, C_j) = d(x_i, x_j), \forall i, j$$

where $D(C_i, C_j)$ is the dissimilarity between clusters $C_i$ and $C_j$, and $d(x_i, x_j)$ is neutrosophic dissimilarity measure between elements $x_i$ and $x_j$. For Euclidean-based measure:

$$d(x_i, x_j) = \sqrt{(T_i - T_j)^2 + (I_i - I_j)^2 + (F_i - F_j)^2}$$

where $T, I$, and $F$ are the truth, indeterminacy, and falsity membership degrees, respectively.

**3. Merge Closest Clusters**:

- Identify the pair of clusters with the smallest neutrosophic dissimilarity.

$$(C_a, C_b) = arg \min_{i,j} D(C_i, C_j)$$

- Merge these clusters into a single cluster.

$$C_{new} = C_a \cup C_b$$

**4. Update Neutrosophic Dissimilarities**:

- Recalculate the neutrosophic dissimilarity between the newly formed cluster and all other clusters.

$$D(C_{new}, C_k) = \frac{D(C_a, C_k) + D(C_b, C_k)}{2}$$

**5. Repeat**:

- Repeat steps 3 and 4 until a single cluster containing all neutrosophic elements is formed.

$$D(C_{new}, C_k) = \frac{D(C_a, C_k) + D(C_b, C_k)}{2}$$

**Dendrogram Construction:**

During the ANHC process, a dendrogram is constructed to visualize the hierarchical relationships between clusters. Here is how the dendrogram is constructed:

- **Vertical Axis**: The vertical axis of the dendrogram represents the neutrosophic dissimilarity or distance between clusters.
- **Horizontal Axis**: The horizontal axis represents the neutrosophic elements or clusters being merged at each step of the clustering process.
- **Merge Points**: At each merge step, a new merge point is added to the dendrogram, indicating the distance at which the clusters were merged.
- **Branches**: Branches in the dendrogram connect clusters at different levels of the hierarchy, showing the order in which clusters were merged and the distances at which they were merged.
- **Leaf Nodes**: Each leaf node in the dendrogram represents an individual neutrosophic element or cluster at the lowest level of the hierarchy.

By analyzing the dendrogram, researchers can interpret the hierarchical relationships between clusters and identify meaningful clusters at different levels of granularity. The dendrogram provides valuable insights into the structure of the neutrosophic data and helps guide the selection of an appropriate number of clusters based on the desired level of detail or resolution.

In short, using neutrosophic logic in hierarchical clustering helps make better decisions, improves accuracy, creates flexible systems, and combines different types of data more effectively. This approach allows data analysts and scientists to make smarter and more reliable choices, leading to better results, efficiency, and adaptability in different fields.



DOI: https://doi.org/10.54216/IJNS.260121

## 5. Experimental Results Analysis and Discussion

This study compares the suggested model's classification performance to that of current techniques. It evaluates the efficacy of certain computational intelligence and machine learning methodologies. Statistical indicators are also used to assess how accurate the cluster number forecast is.

### 5.1. Dataset

The Iris dataset is a well-known resource in statistics and machine learning, featuring measurements of iris flowers from three species: Setosa, Versicolor, and Virginica. It typically includes 150 samples, with 50 instances per species.

### 5.2. Parameters

Performance is assessed using PBM Index, DB index, and silhouette coefficient. Better clustering performance is correlated with higher levels of PBM Index and low DB Index scores.

### 5.3. Experimental Setup

The fuzzy and neutrosophic approaches were developed using Python 3 and executed on a Windows 11 (64-bit) PC with an AMD Ryzen 9 processor and 32GB of RAM.

### 5.4 Comparison analysis

To evaluate the effectiveness of the proposed strategy for predicting the number of clusters, we conducted experiments. Using methods like fuzzy, and neutrosophic, the experiment was run on the dataset.

**Table 1:** Comparing existing and proposed system values

| Sl. No. | Metrics | Methods | |
|---|---|---|---|
| | | **Fuzzy** | **Neutrosophic** |
| **1.** | DB Index | 0.5998014269269159 | 0.5954401418981576 |
| **2.** | PBM Index | 1.0047217756201162e-07 | 1.034081671080926e-07 |
| **3.** | Silhouette Coefficient | 0.5347408431441443 | 0.5357020355143183 |

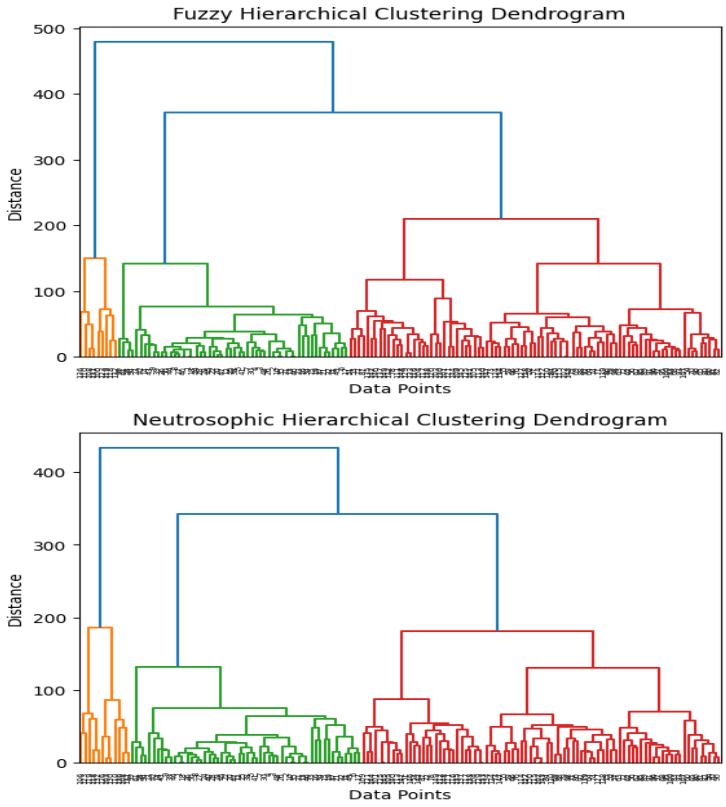


**Figure 2.** Dendrogram construction



DOI: https://doi.org/10.54216/IJNS.260121

Clustering evaluation metrics were analyzed for both the existing and proposed approaches, as shown in Table 1. Traditional methods, such as the fuzzy algorithm, exhibit lower accuracy in clustering indices, whereas the neutrosophic strategy demonstrates higher accuracy. Consequently, the proposed method achieves greater precision in predicting the number of clusters.

All models' dendrogram construction analysis is displayed in Fig 2. We observed that some methods like neutrosophic hierarchical algorithms have performed better than all other algorithms in terms of precision, DB Index, and PBM Index.

## 6. Conclusion and future work

In this paper, we presented a new multilevel algorithm for clustering larger databases with indeterminate data. The algorithm combines a neutrosophic algorithm with unsupervised clustering, especially a hierarchical algorithm. As demonstrated by our experimental results, the multilevel clustering strategy outperforms the standard k-means or fuzzy approaches in producing the optimal clustering solution. By applying neutrosophic logic to multi-level data, the Neutrosophic Hierarchical Clustering Algorithm (NHCA) expands on conventional clustering techniques to manage ambiguity and indeterminacy. When grouping datasets with a variety of properties, such as the Iris dataset, the method performs admirably. The NHCA is a useful tool for data analysis in a variety of sectors because of its capacity to manage uncertainty and hierarchical structures further research could focus on optimizing the computational efficiency of NHCA to handle larger datasets and real-time applications. Extending NHCA to handle additional types of data, such as text or image data could broaden its applicability and relevance in diverse domains. This clustering algorithm handles uncertainty and indeterminacy effectively using neutrosophic logic. It can capture hierarchical structures in the data, providing additional insights. There are still improvements needed for the proposed algorithm as if it has higher computational complexity compared to traditional methods, especially for large datasets. Performance may depend on the choice of neutrosophic dissimilarity measure and other algorithm parameters. For our suggested approach, the outcomes demonstrated exceptional accuracy, recall, and precision.

**Funding:** “This research received no external funding”

**Conflicts of Interest:** “The authors declare no conflict of interest.”

## References

[1] R. T. Ng and J. Han, "Efficient and effective clustering methods for spatial data mining," in *Proc. 20th Int. Conf. Very Large Data Bases (VLDB '94), *San Francisco, CA, USA: Morgan Kaufmann Publishers Inc*., 1994, pp. 144–155.

[2] F. Smarandache, Ed., *Proceedings of the First International Conference on Neutrosophy, Neutrosophic Logic, Neutrosophic Set, Neutrosophic Probability and Statistics, *Univ. of New Mexico, Gallup Campus, Xiquan, Phoenix*, 2002, p. 147.

[3] R. R. Yager and A. T. de Almeida, "On the use of neutrosophic sets in decision making," *Information Sciences*, vol. 572, pp. 358–371, 2021. [Online]. Available: https://doi.org/10.1016/j.ins.2021.06.052

[4] M. N. Murty, A. K. Jain, and P. J. Flynn, "Data clustering: A review," *ACM Comput. Surv*, vol. 31, no. 3, pp. 264–323, 1999.

[5] W. B. Xie, Z. Liu, D. Das, B. Chen, and J. Srivastava, "Scalable clustering by aggregating representatives in hierarchical groups," *Pattern Recognit.*, vol. 136, p. 109230, 2023.

[6] H. Hu, J. Liu, X. Zhang, and M. Fang, "An effective and adaptable k-means algorithm for big data cluster analysis," *Pattern Recognit*., vol. 139, p. 109404, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S003132032300105X

[7] F. M. G. C. A. Cimino, B. Lazzerini, and F. Marcelloni, "A novel approach to fuzzy clustering based on a dissimilarity relation extracted from data using a TS system," *Pattern Recognit*., vol. 39, no. 11, pp. 2077–2091, 2006.

[8] A. Q. Ansari, R. Biswas, and S. Aggarwal, "Neutrosophic classifier: An extension of fuzzy classifier," *Appl. Soft Comput.*, vol. 13, no. 1, pp. 563–573, Jan. 2013. [Online]. Available: https://doi.org/10.1016/j.asoc.2012.08.002

[9] U. Rivieccio, "Neutrosophic logics: Prospects and problems," *Fuzzy Sets Syst*., vol. 159, no. 14, 2008. [Online]. Available: https://doi.org/10.1016/j.fss.2007.11.011



DOI: https://doi.org/10.54216/IJNS.260121

[10] F. Smarandache, A Unifying Field in Logics: Neutrosophic Logic. In: *Neutrosophy, Neutrosophic Set, Neutrosophic Probability: Neutrosophic Logic, Santa Fe, NM, USA: American Research Press*, 2005.

[11] M. G. Gafar, M. Elhoseny, and M. Gunasekaran, "Modeling neutrosophic variables based on particle swarm optimization and information theory measures for forest fires," *J. Supercomput*., vol. 76, pp. 2339–2356, 2020. [Online]. Available: https://doi.org/10.1007/s11227-018-2512-5

[12] M. K. Pakhira, S. Bandyopadhyay, and U. Maulik, "Validity index for crisp and fuzzy clusters," *Pattern Recognit*., vol. 37, no. 3, pp. 487–501, 2004. [Online]. Available: https://doi.org/10.1016/j.patcog.2003.06.005

[13] A. Gallo-Bueno, M. Reynaud, M. Casas-Cabanas, and J. Carrasco, "Unsupervised machine learning to classify crystal structures according to their structural distortion: A case study on Li-argyrodite solid-state electrolytes," *Energy AI*, vol. 9, p. 100159, 2022. [Online]. Available: https://doi.org/10.1016/j.egyai.2022.100159

[14] B. Desgraupes, "Clustering indices," *Univ. Paris Ouest-Lab Modal'X*, vol. 1, no. 34, 2013.



DOI: https://doi.org/10.54216/IJNS.260121