SEEJPH Volume XXVI, S2,2025, ISSN: 2197-5248; Posted:03-02-25

# Cluster Analysis: Theory, Methodology, and Applications

**Sanja Nikolic, Ph.D[1],Tanja Sekulic, Ph.D[2],Branko Medic, D.Sc[3].**

[1]*ORCID: 0000-0001-9632-2458,College for Vocational Education of Preschool Teachers and Coaches drsanjanikolic294@gmail.com*
[2]*Technical College of Applied Sciences,tsekulicvts@gmail.com*
[3]*Academy of Vocational Studies,medic.tesol@gmail.com*

## Introduction

Cluster analysis is a statistical technique used to group objects into sets called clusters, based on their similarities. Clustering is one of the fundamental tasks in data analysis and is widely applied in various fields, including market research, biomedical studies, pattern recognition, and big data analysis. The primary goal of cluster analysis is to categorize data into groups such that objects within each cluster are as similar as possible, while objects from different clusters are as dissimilar as possible. Cluster analysis is a technique in statistics used to group objects based on their similarity. The main objective of this method is to link data into groups (clusters), where objects within each cluster are as similar as possible, while objects from different clusters are as dissimilar as possible. Cluster analysis is valuable in various fields such as market analysis, biomedical studies, biotechnology, pattern recognition, and many other areas where identifying data structure is required. Academics and market researchers often encounter situations best addressed by defining groups of homogeneous objects, whether they are individuals, companies, products, or even their behaviors. Strategic decisions based on identifying groups within a population, such as segmentation and targeted marketing, would not be possible without an objective methodology. This same need arises in other areas, from the physical to the social sciences. In all cases, researchers seek the natural structure among observations based on multiple profiles. The most commonly used technique for this purpose is cluster analysis. It aims to maximize internal homogeneity and external heterogeneity of clusters. An important feature of cluster analysis is the fact that it is not a method of strict statistical inference, where the selected sample is necessarily considered representative of a given population. Cluster analysis is a method for determining structural characteristics of measured properties on a strict mathematical but not statistical basis. Therefore, for the results of cluster analysis to be meaningful, it is necessary to establish assumptions related to the representativeness of the sample and multicollinearity of the variables. In cluster analysis, the group membership of objects is unknown, as is the final number of groups. The goal of cluster analysis is to identify homogeneous groups or clusters.

**Keywords**: cluster method, hierarchical clustering, clustering evaluation

## 1. Clustering Methods

There are several key clustering methods, each with specific advantages and limitations depending on the type of data and the objectives of the analysis. The term "cluster" comes from the English word "cluster," meaning a group of similar things, a bunch, or a heap. A common issue faced by researchers across many areas of study is how to organize observed data into meaningful structures, i.e., how to develop taxonomies. In other words, cluster analysis is a research technique used for data analysis with the goal of sorting different objects into groups such that the degree of association between two objects is maximized if they belong to the same group, and minimized if they belong to different groups. Cluster analysis simply uncovers structures within data without explaining why these structures exist. We encounter grouping in everyday life:

1. A small group of people sitting at the same table in a restaurant can be classified as one group;
2. Biologists must organize different animal species before meaningful differences between them can be established;

3. In market segmentation, consumer clusters are formed in a country, and then a separate business plan is made for each cluster;
4. In marketing, cluster analysis is used to analyze product or service characteristics, customer attitudes, demographic factors, etc.

Cluster analysis can be effectively used for data reduction. For example, if testing a new product in different cities, clusters of similar cities are formed, and one city is chosen from each cluster for testing, to avoid analyzing every city. Furthermore, if cluster analysis reveals an unexpected grouping of observation units, there is a likelihood that certain relationships between the units, previously unknown, have been discovered and need to be investigated. It is important to know that the more variables are included in the analysis and the more independent they are, the harder it is to find an appropriate model for grouping the observation units.

The first forms of cluster analysis appeared in the early 20th century, but significant literature on this subject has developed since the 1960s. The rapid development of computers, combined with the fundamental importance of classification as a scientific procedure, has led to the popularity of this method. Psychologists sometimes refer to it as a "poor man's factor analysis." Within cluster analysis, there are many different algorithms that generally address the same problems.

**What is not cluster analysis?**

1. Supervised classification (There is information about class labels – classification),
2. Simple division (Division of students by the first letter of their last name),
3. Survey results (Grouping is the result of external specification, and others).

**In marketing, cluster analysis is used for:**

1. Segmenting the market and identifying target markets,
2. Product positioning and new product development,
3. Selection of market testing.

**Some important notes regarding the use of cluster analysis:**

1. Most cluster analysis methods are relatively simple statistical procedures and do not have significant support in standard statistical reasoning (e.g., determining significance).
2. Some methods were developed and are useful within specific scientific disciplines, while they may not be of major importance in others.
3. Different clustering methods may, and often do, lead to different final solutions.

The result of cluster analysis always represents the classification of objects into groups, which, depending on the technique used, may lead to different solutions. One important criterion may be the "psychological" meaningfulness of the obtained solution.

Along with cluster analysis, a specific terminology has also developed. Units that are grouped into clusters are typically referred to as objects, entities, cases, or operational taxonomic units (OTUs). Grouping is performed based on certain variables, attributes, characteristics, or features.

### 1.1. Hierarchical Clustering

Hierarchical clustering is based on the formation of a hierarchy of clusters, where objects or clusters are gradually merged (agglomerative approach) or divided (divisive approach). In the agglomerative approach, each object is initially treated as an individual cluster, and then the most similar clusters are merged until one large cluster is formed. This process can be represented through a dendrogram, a

graphical depiction of the hierarchy. The advantage of this method is that it does not require a predefined number of clusters, while the disadvantage is that it is computationally more demanding, especially for large datasets (Johnson, 1967).

### 1.2.K-means Clustering

K-means is the most popular clustering method, based on assigning objects to clusters such that the internal variance within the clusters is minimized. This method requires the predefined specification of the number of clusters (k). The algorithm functions by randomly selecting k cluster centers, and objects are assigned to clusters based on their proximity to the centers stabilize (Lloyd, 1982). The advantages of K-means clustering include its simplicity of implementation and speed. However, its main limitation is that it requires a predefined number of clusters, which can be challenging when this number is not known.

### 1.2. DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

DBSCAN is a density-based clustering method that is particularly suited for data with unclear boundaries or noisy elements. DBSCAN identifies clusters as groups of objects that reside in areas with high data density, while objects located in low-density regions are not assigned to any cluster and are treated as noise. This method is useful in situations where the number of clusters is not known in advance and is especially effective with unstructured data (Ester et al., 1996). DBSCAN is particularly valuable when data contains noise, as it helps in identifying outliers, objects that do not fit into any cluster.

### 1.4. Mean-Shift Clustering

Mean-shift clustering is a method that does not require the predefined specification of the number of clusters and uses shifting windows to identify dense regions in the data. This method is particularly useful for working with data that has irregular shapes and heterogeneous distributions. The algorithm shifts the centers of the windows towards regions of highest data density, allowing the identification of clusters (Cheng et al., 1995).

## 2. Clustering Evaluation

One of the key challenges in cluster analysis is evaluating the quality of the clusters. Several methods exist for assessing clustering quality:

### 2.1. Internal Evaluation

Internal evaluation relies on internal clustering criteria, such as the minimization of intra-cluster variance. For example, in K-means clustering, the minimization of the total squared distance between objects and their cluster centers is used. A measure of success in this context is the lower variance within the clusters (MacQueen, 1967).

### 2.2. External Evaluation

External evaluation uses external reference labels to compare the obtained clusters with pre-known classifications. Indices such as the Rand index, F-measure, and Normalized Mutual Information (NMI) are used to compare clustering results with actual object labels (Vinh et al., 2010).

### 2.3. Silhouette Coefficient

The silhouette coefficient is a popular measure that indicates how well an object has been assigned to its cluster relative to other clusters. The coefficient value ranges from -1 to 1, where higher values indicate better assignment of objects to clusters (Rousseeuw, 1987).

## 3. Applications of Clustering

Clustering has broad applications across different disciplines, including:

- **Market Analysis**: Clustering is used for market segmentation to identify groups of consumers with similar preferences and behaviors (Kaufman & Rousseeuw, 1990).
- **Pattern Recognition**: Clustering is employed in pattern recognition within large datasets, such as sensor data, biometric data, and online user behavior (Aggarwal & Reddy, 2014).

- **Biomedical Studies**: In biomedical research, clustering is used for analyzing genetic data and identifying similar patterns in patient health data (Kerr & Churchill, 2001).
- **Anomaly Detection**: DBSCAN and other methods are frequently used for detecting anomalies or unusual data points within large datasets (Hodge & Austin, 2004).

### 3.1. Recent Studies and Research

Recent studies have focused on improving the efficiency of clustering in the context of big data, as well as applying deep learning for clustering unstructured data. For example, Lloyd & Rupp (2020) discuss the optimization of K-means clustering for working with big data, while Xu & Wunsch (2022) explore the application of the DBSCAN algorithm in geospatial data analysis, which results in better clustering outcomes for data with low density and noise.

### 3.2. Important Decisions in Cluster Analysis

Some critical decisions to be made when performing cluster analysis include:

1. The choice of the sample to be subjected to cluster analysis.
2. Identifying the set of relevant variables that will represent the characteristics of the objects (entities).
3. Determining the transformation of the original data.
4. Choosing the method for measuring distance/similarity between objects (entities).
5. Selecting the method for connecting objects into clusters.
6. Validating the obtained results.

Many of these decisions are associated with the choice of an appropriate statistical algorithm or technique. For example, if an important characteristic of political parties is omitted, the grouping result is likely to be suboptimal. The concept of distance or similarity is central to many statistical techniques. Distance measures refer to the dissimilarity (distance) between two objects with respect to a measured characteristic (e.g., the distance between two parties in terms of membership numbers).

**Similarity measures** indicate how close two objects are to each other. For closely related objects, distance measures are small, while similarity measures are large. In cluster analysis, these concepts are critical, as they form the basis for cluster formation. The choice of a distance measure should be based on the characteristics of the variables as well as the algorithm used for clustering.

Cluster analysis is also known by several other names:

1. Q-analysis,
2. Typology construction,
3. Classification analysis,
4. Numerical taxonomy.

This diversity in terminology arises due to the use of the clustering method in various disciplines such as psychology, biology, sociology, and economics. Despite different names depending on the discipline, all methods share a common dimension: classification according to natural relationships.

Cluster analysis methods in mathematical taxonomy can be divided into two main groups:

1. **Agglomerative methods** (methods that group taxonomic units into clusters based on similar characteristics).
2. **Divisive methods** (methods that divide a set of taxonomic units into multiple clusters)

SEEJPH Volume XXVI, S2,2025, ISSN: 2197-5248; Posted:03-02-25

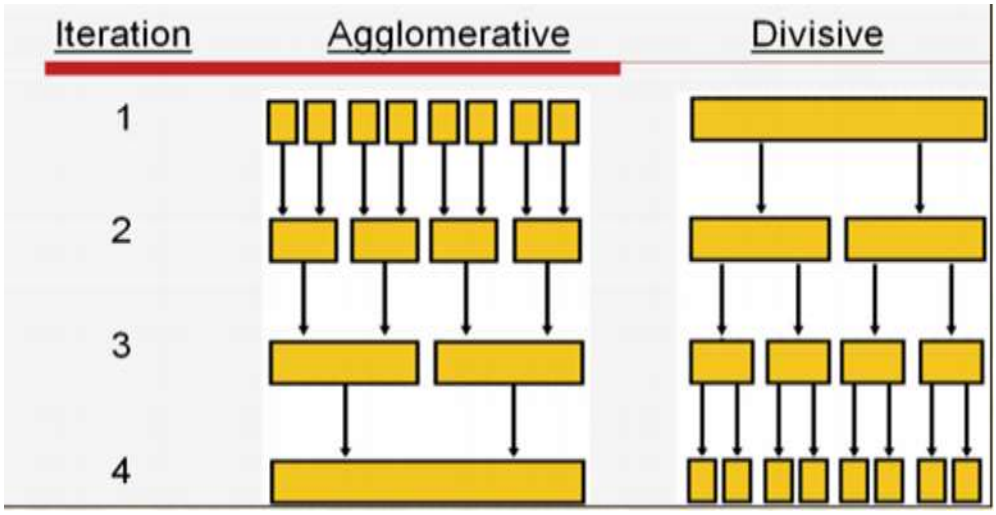


**Figure 1.** Agglomerative and Divisive Methods

**Table 1:** Grouping of Objects Based on Measured Characteristics

| | **Measured characteristics of objects (variables)** | | | |
|---|---|---|---|---|
| **Object** | **$VAR_1$** | **$VAR_2$** | **$VAR_3$** | **$VAR_k$** |
| **Object 1** | $X_{11}$ | $X_{12}$ | $X_{13}$ | $X_{1k}$ |
| **Object 2** | $X_{21}$ | $X_{22}$ | $X_{23}$ | $X_{2k}$ |
| **Object 3** | $X_{31}$ | $X_{32}$ | $X_{33}$ | $X_{3k}$ |
| **...** | | | | |
| **Object N** | $X_{N1}$ | $X_{N2}$ | $X_{N3}$ | $X_{Nk}$ |

In the above case, cluster analysis would aim to determine how objects are grouped based on some of their measured characteristics. Suppose the objects consist of political parties, numbered from 1 to N. The relevant characteristics of each party are represented by numerical variables, numbered from 1 to k. At least one variable with some measured characteristics of the objects is required. For example, assume we have three variables representing the relevant characteristics of each party:

P1 = position on the government-opposition dimension;

P2 = number of members;

P3 = position on the unitary-regional dimension.

For instance, the value X11 could be the average rating of a group of 100 respondents determining the position of the first party on the government-opposition dimension.

In this case, cluster analysis would attempt to identify groups of parties that are most similar to each other based on these characteristics, i.e., the least different. Therefore, it is necessary to psychologically interpret

the resulting clusters. An identical situation arises when the units of analysis are individual respondents for whom we have some measured characteristics. In this case, the input data matrix consists of respondents and their data in the variables.

**Table 2:** Grouping of Respondents Based on Measured Characteristics

| | **Measured characteristics of respondents (variables)** | | | |
|---|---|---|---|---|
| **Respondents** | **$VAR_1$** | **$VAR_2$** | **$VAR_3$** | **$VAR_k$** |
| **Respondent1** | $X_{11}$ | $X_{12}$ | $X_{13}$ | $X_{1k}$ |
| **Respondent2** | $X_{21}$ | $X_{22}$ | $X_{23}$ | $X_{2k}$ |
| **Respondent 3** | $X_{31}$ | $X_{32}$ | $X_{33}$ | $X_{3k}$ |
| **...** | | | | |

In the above case, we are interested in how respondents are grouped based on their characteristics (V1 to Vk). Another possible situation, based on the content of the input matrix, can represent the case where variables (which in this case represent objects or entities) are in the columns, and respondents who evaluated each object based on some characteristic are in the rows. These respondents serve as the source of information about the similarity or dissimilarity of the variables.

4. **Cluster Analysis for Large Datasets (K-means Cluster Analysis)**

Within SPSS, there is a specific algorithm for classifying objects into clusters, called K-means cluster analysis (or Quick Cluster), which is suitable for working with a large number of objects (e.g., 200 or more). Unlike hierarchical cluster analysis, which results in the successive merging of objects into larger clusters, K-means only provides one solution for a pre-defined number of clusters. In this procedure, the final number of clusters must be defined in advance, after which the program assigns new objects to the nearest cluster. The algorithm for this type of cluster analysis is the nearest centroid sorting method (Anderberg, 1973). According to this algorithm, an object is assigned to the cluster whose center (centroid) is the closest. If the cluster centroids are known (i.e., the average values for each of the k variables), the object is assigned to the cluster with the smallest distance. If the cluster centroids are unknown, they are iteratively estimated from the data.

**Table 3:** For example, for the cluster of successful individuals, the centers may be.

| variable | V1 | V2 | V3 | V4 |
|---|---|---|---|---|
| center | 12.5 | 11.0 | 12.0 | 10.7 |

For each cluster, the center represents the arithmetic mean of all the variables calculated based on the objects that form the cluster. Typically, for this purpose, solutions obtained using one of the previously mentioned methods of hierarchical cluster analysis are used. These cluster centers are called "initial cluster centers." New objects are standardized on a scale with M and σ of the original dataset. For each new object, the Euclidean distance from the initial cluster centers is calculated, and the object is assigned to the nearest cluster. After all new objects have been assigned, it is possible to recalculate the cluster centers. These centers are called "final cluster centers." The cluster centers based on which the classification of new objects into the designated

clusters is performed are called "classification cluster centers" and there are (number of clusters x number of variables = number of centers).

**Table 4:** For example, for 3 clusters and 4 variables, we have 12 values.

| **v**ariables**:** | **V1** | **V2** | **V3** | **V4** |
|---|---|---|---|---|
| klaster 1 | 8 | 7 | 6.5 | 4 |
| cluster 2 | 9 | 9 | 8 | 7 |
| cluster 3 | 12.5 | 11.0 | 12.0 | 10.7 |

It is possible to check how successfully these centers separate the individual clusters from each other on individual variables. One option is to calculate the Euclidean distances between each pair of clusters (the result is a matrix of Euclidean distances). Another option, which takes variability into account, is variance analysis between clusters for each variable, which examines the variance between clusters and within clusters.

**Example1:**
For five different light bulb manufacturers (taxonomic units), the lighting provided by the bulbs was tested, in the following order: 60W, 75W, 100W, and 150W. Based on the illumination data for the given light bulbs, we can form a hierarchical structure of objects using the single linkage method, and as the distance between objects, we will use Euclidean distance.

- k1 - 60W bulb
- k2 - 75W bulb
- k3 - 100W bulb
- k4 - 150W bulb
- t1 - Light bulb manufacturer: G. Electric
- t2 - Maxi
- t3 - Philips
- t4 - JMC
- t5 - ΕΛΛΑ∑

| | $k_1$ | $k_4$ | $k_3$ | $k_2$ |
|---|---|---|---|---|
| $t_1$ | 145 | 403 | 321 | 321 |
| $t_2$ | 119 | 320 | 250 | 189 |
| $t_3$ | 20.36 | 488 | 274 | 51 |
| $t_4$ | 114 | 482 | 252 | 156 |
| $t_5$ | 148 | 504 | 266 | 190 |

**Table 5. Calculated Euclidean Distances**

Initially, all taxonomic units are considered as individual groups. Then, in the first step, the taxonomic units with the smallest degree of distance are merged. In this case, these are t3 and t4, as their degree of distance is 41.67.In the next step, we calculate the distance between the newly formed group and the other taxonomic units. The distance from this group to t1 is the smaller of the distances d(t3,t1) and d(t4,t1), which is 117.11.

The distance from this group to t2 is again the smaller of the distances 239.94 and 165.41, which is 165.41. The distance from this group to t5 is 54.70 (Table 6).

| | $t_{34}$ | $t_1$ | $t_2$ | $t_5$ |
|---|---|---|---|---|
| $t_{34}$ | 0 | 117,11 | 165,41 | 54,70 |
| $t_1$ | 117,11 | 0 | 112,72 | 101,04 |
| $t_2$ | 165,41 | 112,72 | 0 | 186,95 |
| $t_5$ | 54,70 | 101,04 | 186,95 | 0 |

Repeated procedures result in Tables 7 and 8, from which the dendrogram in Figure 2 is derived.

| | $t_{345}$ | $t_1$ | $t_2$ |
|---|---|---|---|
| $t_{345}$ | 0 | 101,04 | 165,41 |
| $t_1$ | 101,04 | 0 | 112,72 |
| $t_2$ | 165,41 | 112,72 | 0 |

**Table. 7**

| | $t_{3451}$ | $t_2$ |
|---|---|---|
| $t_{3451}$ | 0 | 112.72 |
| $t_2$ | 112,72 | 0 |

**Table. 8**

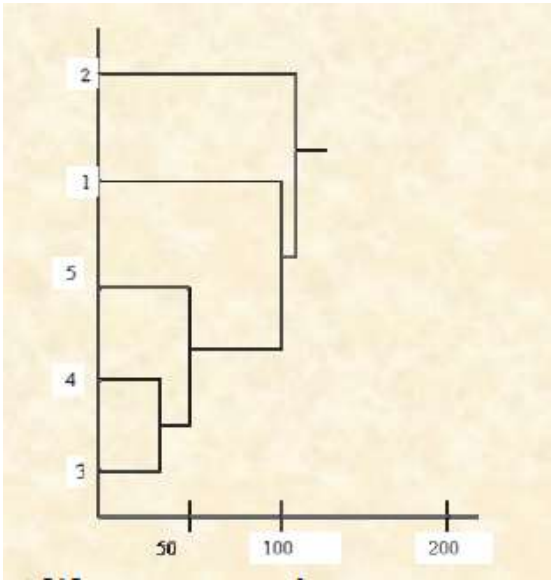


**Figure 2.** Dendogram

## 5. Conclusion

We can observe that the bulbs t3 and t4 are the most similar (i.e., the bulbs manufactured by Philips and JMC have the most similar illumination). The bulbs from Philips and JMC have the most distinct illumination compared to the bulb from the manufacturer Maxi. Furthermore, the Philips and JMC bulbs are most similar to the bulb from the manufacturer ΕΛΛΑ∑. The ΕΛΛΑ∑ bulb is most similar to the group of Philips and JMC bulbs as well as the G. Electric bulb. The G. Electric bulb is most similar to the

Maxi bulb.Clusters play an important role in improving the competitiveness, productivity, and development of small and medium-sized enterprises (SMEs). This is especially significant for transition economies, where companies can become competitive on the international stage by forming clusters. This is particularly relevant for businesses in the textile industry, which have been struggling with competitiveness in the domestic market for many years.Successful clusters offer the following advantages to small and medium-sized enterprises: better access to new skills and knowledge, shared services, partnership support, product branding, development of joint marketing strategies, collaborative work on innovations, more efficient implementation of Quality Management Systems (QMS), co-financing from private and public entities. The government also plays an important role in the process of cluster formation. Cluster policy involves the government's initiative in forming clusters, as well as enhancing business contacts and relations between collaborators based on trade links, innovation links, knowledge flow, and providing specialized infrastructural support.Small and medium-sized enterprises in the textile industry, in order to improve their competitiveness in the market, adopt strategic clustering as a solution. In Serbia, a textile industry cluster has been established, but its activities need to be significantly improved for this form of collaboration to provide the expected results. Cluster analysis is a powerful tool in research that enables the grouping of data into homogeneous and meaningfully connected groups. Its application is not limited to statistical fields but extends across various disciplines, from market research to social sciences. This method allows researchers to discover hidden structures and patterns in data, enabling deeper insights and more informed decision-making. Although cluster analysis does not provide explanations for why certain groups exist, it allows clear and objective classification of objects based on their similarities and differences. However, to ensure valid results, it is crucial to carefully consider the assumptions related to sample representativeness and multicollinearity of variables. In the future, further refinement of cluster analysis methods and their integration with advanced techniques such as machine learning could enable even more precise and efficient results in data analysis.

## Reference


- Aggarwal, C. C., & Reddy, C. K. (2014). *Data Clustering: Algorithms and Applications*. CRC Press.
- Cheng, Y., et al. (1995). Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 17*(8), 790-799.
- Ester, M., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining* (pp. 226-231).
- Hodge, V. J., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review, 22*(2), 85-126.
- Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika, 32*(3), 241-254.
- Kaufman, L., & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley.
- Kerr, M. K., & Churchill, G. A. (2001). Statistical design and the analysis of gene expression microarray data. *Genetics, 158*(1), 169-178.
- Lloyd, S. P. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory, 28*(2), 129-137.
- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1*, 281-297.
- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Graphical Statistics, 1*(1), 53-65.
- Vinh, N. X., et al. (2010). Information theoretic measures for clusterings comparison: is a correction for chance necessary? *Proceedings of the 26th Annual International Conference on Machine Learning*, 1073-1080.
- Xu, R., & Wunsch, D. (2022). *Clustering algorithms: A survey*. IEEE Transactions on Neural Networks and Learning Systems, 33(10), 4304-4318