# Chapter 13

# Clustering Algorithms

*"Wherever you see a successful business, someone once made a courageous decision."*

— **Peter Drucker**

Cluster analysis is a technique of partitioning a collection of unlabelled objects, with many attributes, into meaningful disjoint groups or clusters. This chapter aims to provide the basic concepts of clustering algorithms.

## Learning Objectives

- Introduce the concepts of clustering
- Highlight the role of distance measures in clustering process
- Provide a taxonomy of clustering algorithms
- Explain hierarchical clustering algorithms
- Explain partitional clustering algorithms
- Briefly explain density-based, grid-based, and probabilistic model-based clustering techniques
- Discuss the validation techniques for clustering algorithms

## 13.1 INTRODUCTION TO CLUSTERING APPROACHES

Cluster analysis is the fundamental task of unsupervised learning. Unsupervised learning involves exploring the given dataset. Cluster analysis is a technique of partitioning a collection of unlabelled objects that have many attributes into meaningful disjoint groups or clusters. This is done using a trial and error approach as there are no supervisors available as in classification. The characteristic of clustering is that the objects in the clusters or groups are similar to each other within the clusters while differ from the objects in other clusters significantly.

The input for cluster analysis is examples or samples. These are known as objects, data points or data instances. All these terms are same and used interchangeably in this chapter. All the samples or objects with no labels associated with them are called unlabelled. The output is

the set of clusters (or groups) of similar data if it exists in the input. For example, the following Figure 13.1(a) shows data points or samples with two features shown in different shaded samples and Figure 13.1(b) shows the manually drawn ellipse to indicate the clusters formed.

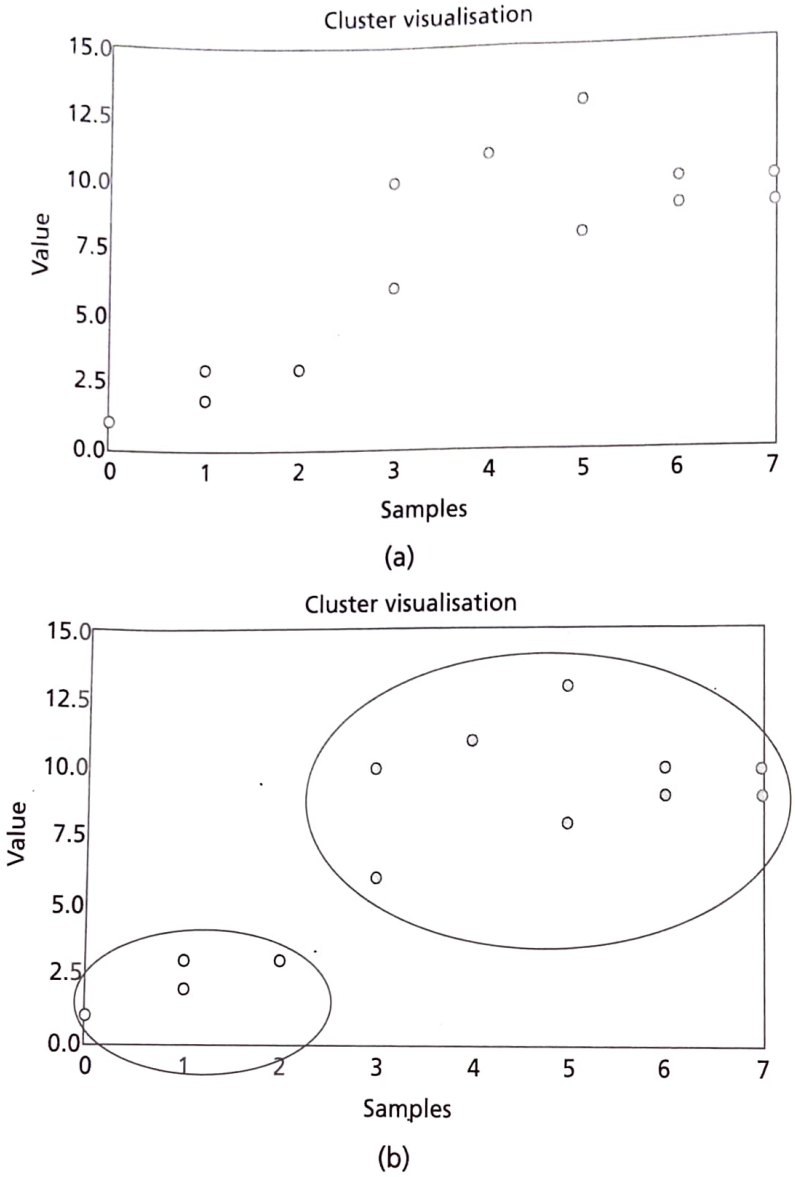


**Figure 13.1:** (a) Data Samples (b) Clusters' Description

Visual identification of clusters in this case is easy as the examples have only two features. But, when examples have more features, say 100, then clustering cannot be done manually and automatic clustering algorithms are required. Also, automating the clustering process is desirable as these tasks are considered difficult by humans and almost impossible. All clusters are represented by centroids. For example, if the input examples or data is (3, 3), (2, 6) and (7, 9), then the centroid is given as $\left(\frac{3+2+7, 3+6+9}{3}\right) = (4, 6)$. The clusters should not overlap and every cluster should represent only one class. Therefore, clustering algorithms use trial and error method to form clusters that can be converted to labels. Thus, the important differences between classification and clustering are given in Table 13.1.

Table 13.1: Differences between Classification and Clustering

| S.No. | Clustering | Classification |
|---|---|---|
| 1. | Unsupervised learning and cluster formation are done by trial and error as there is no supervisor | Supervised learning with the presence of a supervisor to provide training and testing data |
| 2. | Unlabelled data | Labelled data |
| 3. | No prior knowledge in clustering | Knowledge of the domain is must to label the samples of the dataset |
| 4. | Cluster results are dynamic | Once a label is assigned, it does not change |

## *Applications of Clustering*

1. Grouping based on customer buying patterns
2. Profiling of customers based on lifestyle
3. In information retrieval applications (like retrieval of a document from a collection of documents)
4. Identifying the groups of genes that influence a disease
5. Identification of organs that are similar in physiology functions
6. Taxonomy of animals, plants in Biology
7. Clustering based on purchasing behaviour and demography
8. Document indexing
9. Data compression by grouping similar objects and finding duplicate objects

## *Challenges of Clustering Algorithms*

A huge collection of data with higher dimensions (i.e., features or attributes) can pose a problem for clustering algorithms. With the arrival of the internet, billions of data are available for clustering algorithms. This is a difficult task, as scaling is always an issue with clustering algorithms. Scaling is an issue where some algorithms work with lower dimension data but do not perform well for higher dimension data. Also, units of data can post a problem like some weights in kg and some in pounds can pose a problem in clustering. Designing a proximity measure is also a big challenge.

The advantages and disadvantages of the cluster analysis algorithms are given in Table 13.2.

Table 13.2: Advantages and Disadvantages of Clustering Algorithms

| S.No. | Advantages | Disadvantages |
|---|---|---|
| 1. | Cluster analysis algorithms can handle missing data and outliers. | Cluster analysis algorithms are sensitive t initialization and order of the input data. |
| 2. | Can help classifiers in labelling the unlabelled data. Semi-supervised algorithms use cluster analysis algorithms to label the unlabelled data and then use classifiers to classify them. | Often, the number of clusters present in th data have to be specified by the user. |

(Continu

| S.No. | Advantages | Disadvantages |
|---|---|---|
| 3. | It is easy to explain the cluster analysis algorithms and to implement them. | Scaling is a problem. |
| 4. | Clustering is the oldest technique in statistics and it is easy to explain. It is also relatively easy to implement. | Designing a proximity measure for the given data is an issue. |

## 13.2 PROXIMITY MEASURES

Scan for *'Additional Information on Proximity Measures'*

Clustering algorithms need a measure to find the similarity or dissimilarity among the objects to group them. Similarity and dissimilarity are collectively known as proximity measures. Often, the distance measures are used to find similarity between two objects, say $i$ and $j$.

Distance measures are known as dissimilarity measures, as these indicate how one object is different from another. Measures like cosine similarity indicate the similarity among objects. Distance measures and similarity measures are two sides of the same coin, as more distance indicates more similarity and vice versa. Distance between two objects, say $i$ and $j$, is denoted by the symbol $D_{ij}$.

The properties of the distance measures are:

1. $D_{ij}$ is always positive or zero.
2. $D_{ij} = 0$, i.e., the distance between the object to itself is 0.
3. $D_{ij} = D_{ji}$. This property is called symmetry.
4. $D_{ij} \leq D_{ik} + D_{kj}$. This property is called triangular inequality.

If all these conditions are satisfied, then the distance measure is called a metric.

Based on the data types of the attributes of an object, distance measures vary. The concept of data types is discussed in Chapter 2. It can be recollected that the data types are divided into categorical and quantitative variables. Quantitative variables are numbers and are of two types – nominal and ordinal. For example, gender is a nominal variable as gender can be enumerated as Gender = {male, female}. Ordinal variables look like nominal variables but have an inherent order present in the enumeration. For example, temperature is a nominal variable as temperature can be enumerated as Temperature = {low, medium, high}. One can observe the inherent order present, that is, medium > low and low < medium. Quantitative variables are real or Integer numbers or binary data. In binary data, the attributes of the object can take a Boolean value. Objects whose attributes take binary data are called binary objects.

Let us review some of the proximity measures.

### *Quantitative Variables*

Some of the qualitative variables are discussed below.

**Euclidean Distance** It is one of the most important and common distance measures. It is also called as $L_2$ norm. It can be defined as the square root of squared differences between the coordinates of a pair of objects.

The Euclidean distance between objects $x_i$ and $x_j$ with $k$ features is given as follows:

$$\text{Distance}(x_i, x_j) = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2} \tag{13.1}$$

The advantage of Euclidean distance is that the distance does not change with the addition of new objects. But the disadvantage is that if the units change, the resulting Euclidean or squared Euclidean changes drastically. Another disadvantage is that as the Euclidean distance involves a square root and a square, the computational complexity is high for implementing the distance for millions or billions of operations involved.

**City Block Distance** City block distance is known as Manhattan distance. This is also known as boxcar, absolute value distance, Manhattan distance, Taxicab or $L_1$ norm. The formula for finding the distance is given as follows:

$$\text{Distance}(x_i, x_j) = \sum_{k=1}^{n}\left|x_{ik} - x_{jk}\right| \tag{13.2}$$

**Chebyshev Distance** Chebyshev distance is known as maximum value distance. This is the absolute magnitude of the differences between the coordinates of a pair of objects. This distance is called supremum distance or $L_{max}$ or $L_\infty$ norm. The formula for computing Chebyshev distance is given as follows:

$$\text{Distance}(x_i, x_j) = \max_k\left|x_{ik} - x_{jk}\right| \tag{13.3}$$

**Example 13.1:** Suppose, if the coordinates of the objects are (0, 3) and (5, 8), then what is the Chebyshev distance?

**Solution:** The Euclidean distance using Eq. (13.1) is given as follows:

$$\text{Distance}(x_i, x_j) = \sqrt{(0-5)^2 + (3-8)^2}$$
$$= \sqrt{50} = 7.07$$

The Manhattan distance using Eq. (13.2) is given as follows:

$$\text{Distance}(x_i, x_j) = \left|(0-5) + (3-8)\right| = 10$$

The Chebyshev distance using Eq. (13.3) is given as follows:

$$\text{Max}\left\{\left|0-5\right|, \left|3-8\right|\right\} = \text{Max}\{5, 5\} = 5$$

**Minkowski Distance** In general, all the above distance measures can be generalized as:

$$\text{Distance}(x_i, x_j) = \left(\sum\left|x_{ik} - x_{jk}\right|^r\right)^{\frac{1}{r}} \tag{13.}$$

This is called Minkowski distance. Here, $r$ is a parameter. When the value of $r$ is 1, the distance measure is called city block distance. When the value of $r$ is 2, the distance measure is called Euclidean distance. When, $r$ is $\infty$, then this is Chebyshev distance.

### *Binary Attributes*

Binary attributes have only two values. Distance measures discussed above cannot be applied to find distance between objects that have binary attributes. For finding the distance among objects with binary objects, the contingency Table 13.3 can be used. Let $x$ and $y$ be the objects consisting of N-binary objects. Then, the contingency table can be constructed by counting the number of matching of transitions, 0–0, 0–1, 1–0 and 1–1.

**Table 13.3:** Contingency Table

| Attributes Matching | 0 | 1 |
|---|---|---|
| 0 | a | b |
| 1 | c | d |

In other words, '$a$' is the number of attributes where $x$ attribute is 0 and $y$ attribute is 0. '$b$' is the number of attributes where $x$ attribute is 0 and $y$ attribute is 1, '$c$' is the number of attributes where $x$ attribute is 1 and $y$ attribute is 0 and '$d$' is the number of attributes where $x$ attribute is 1 and $y$ attribute is 1.

**Simple Matching Coefficient (SMC)** SMC is a simple distance measure and is defined as the ratio of number of matching attributes and the number of attributes. The formula is given as:

$$\frac{a+d}{a+b+c+d} \tag{13.5}$$

The values of $a$, $b$, $c$, and $d$ can be observed from the Table 13.4.

**Jaccard Coefficient** Jaccard coefficient is another useful measure for and is given as follows:

$$J = \frac{d}{b+c+d} \tag{13.6}$$

**Example 13.2:** If the given vectors are $x = (1, 0, 0)$ and $y = (1, 1, 1)$ then find the SMC and Jaccard coefficient?

**Solution:** It can be seen from Table 13.2 that, $a = 0$, $b = 2$, $c = 0$ and $d = 1$.

The SMC using Eq. (13.5) is given as $\frac{a+d}{a+b+c+d} = 0+1/3 = 0.33$

Jaccard coefficient using Eq. (13.6) is given as $J = \frac{d}{b+c+d} = 1/3 = 0.33$

**Hamming Distance** Hamming distance is another useful measure that can be used for knowing the sequence of characters or binary values. It indicates the number of positions at which the characters or binary bits are different.

For example, the hamming distance between $x = (1\ 0\ 1)$ and $y = (1\ 1\ 0)$ is 2 as $x$ and $y$ differ in two positions. The distance between two words, say wood and hood is 1, as they differ in only one character. Sometimes, more complex distance measures like edit distance can also be used.

### Categorical Variables

In many cases, categorical values are used. It is just a code or symbol to represent the values. For example, for the attribute Gender, a code 1 can be given to female and 0 can be given to male.

To calculate the distance between two objects represented by variables, we need to find only whether they are equal or not. This is given as:

$$\text{Distance}(x, y) = \begin{cases} 1 & \text{if } x \neq y \\ 0 & \text{if } x = y \end{cases} \tag{13.7}$$

### Ordinal Variables

Ordinal variables are like categorical values but with an inherent order. For example, designation is an ordinal variable. If job designation is 1 or 2 or 3, it means code 1 is higher than 2 and code 2 is higher than 3. It is ranked as $1 > 2 > 3$.

Let us assume the designations of office employees are clerk, supervisor, manager and general manager. These can be designated as numbers as clerk = 1, supervisor = 2, manager = 3 and general manager = 4. Then, the distance between employee $X$ who is a clerk and $Y$ who is a manager can be obtained as:

$$\text{Distance}(X, Y) = \frac{|position(X) - position(Y)|}{n - 1} \tag{13.8}$$

Here, position (X) and position(Y) indicate the designated numerical value. Thus, the distance between $X$ (Clerk = 1) and $Y$ (Manager = 3) using Eq. (13.8) is given as:

$$\text{Distance}(X, Y) = \frac{|position(X) - position(Y)|}{n - 1} = \frac{|1 - 3|}{4 - 1} = \frac{2}{3} \approx 0.66$$

### Vector Type Distance Measures

For text classification, vectors are normally used. Cosine similarity is a metric used to measure how similar the documents are irrespective of their size. Cosine similarity measures the cosine of the angle between two vectors projected in a multi-dimensional space. The similarity function for vector objects can be defined as:

$$sim(X, Y) = \frac{X \bullet Y}{\|X\| \times \|Y\|} = \frac{\sum_{i=1}^{n} X_i \times Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2} \times \sqrt{\sum_{i=1}^{n} Y_i^2}} \tag{13.9}$$

The numeration is the dot product of the vectors $A$ and $B$. The denominator is the product of the norm of vectors $A$ and $B$.

**Example 13.3:** If the given vectors are $A = \{1, 1, 0\}$ and $B = \{0, 1, 1\}$, then what is the cosine similarity?

**Solution:** The dot product of the vector is $1 \times 0 + 1 \times 1 + 0 \times 1 = 1$. The norm of the vectors $A$ and $B$ is $\sqrt{2}$.

So, the cosine similarity using Eq. (13.9) is given as $\frac{1}{\sqrt{2}\sqrt{2}} = \frac{1}{2} = 0.5$

### Advantages

1. Simple
2. Easy to implement

### Disadvantages

1. It is sensitive to initialization process as change of initial points leads to different clusters.
2. If the samples are large, then the algorithm takes a lot of time.

### How to Choose the Value of k?

It is obvious that $k$ is the user specified value specifying the number of clusters that are present. Obviously, there are no standard rules available to pick the value of $k$. Normally, the $k$-means algorithm is run with multiple values of $k$ and within group variance (sum of squares of samples with its centroid) and plotted as a line graph. This plot is called Elbow curve. The optimal or best value of $k$ can be determined from the graph. The optimal value of $k$ is identified by the flat or horizontal part of the Elbow curve.

### Complexity

The complexity of $k$-means algorithm is dependent on the parameters like $n$, the number of samples, $k$, the number of clusters, $\Theta(nkId)$. $I$ is the number of iterations and $d$ is the number of attributes. The complexity of $k$-means algorithm is $O(n^2)$.

**Example 13.5:** Consider the following set of data given in Table 13.9. Cluster it using $k$-means algorithm with the initial value of objects 2 and 5 with the coordinate values (4, 6) and (12, 4) as initial seeds.

**Table 13.9:** Sample Data

| Objects | X-coordinate | Y-coordinate |
|---|---|---|
| 1 | 2 | 4 |
| 2 | 4 | 6 |
| 3 | 6 | 8 |
| 4 | 10 | 4 |
| 5 | 12 | 4 |

**Solution:** As per the problem, choose the objects 2 and 5 with the coordinate values. Hereafter, the objects' id is not important. The samples or data points (4, 6) and (12, 4) are started as two clusters as shown in Table 13.10.

Initially, centroid and data points are same as only one sample is involved.

**Table 13.10:** Initial Cluster Table

| Cluster 1 | Cluster 2 |
|---|---|
| (4, 6) | (12, 4) |
| Centroid 1 (4, 6) | Centroid 2 (12, 4) |

Iteration 1: Compare all the data points or samples with the centroid and assign to the nearest sample. Take the sample object 1 (2, 4) from Table 13.9 and compare with the centroid of

the clusters in Table 13.10. The distance is 0. Therefore, it remains in the same cluster. Similarly, consider the remaining samples. For the object 1 (2, 4), the Euclidean distance between it and the centroid is given as:

Dist (1, centroid 1) = $\sqrt{(2-4)^2 + (4-6)^2} = \sqrt{8}$

Dist (1, centroid 2) = $\sqrt{(2-12)^2 + (4-4)^2} = \sqrt{100} = 10$

Object 1 is closer to the centroid of cluster 1 and hence assign it to cluster 1. This is shown in Table 13.11. Object 2 is taken as centroid point.

For the object 3 (6, 8), the Euclidean distance between it and the centroid points is given as:

Dist (3, centroid 1) = $\sqrt{(6-4)^2 + (8-6)^2} = \sqrt{8}$

Dist (3, centroid 2) = $\sqrt{(6-12)^2 + (8-4)^2} = \sqrt{52}$

Object 3 is closer to the centroid of cluster 1 and hence remains in the same cluster 1.

Proceed with the next point object 4(10, 4) and again compare it with the centroids in Table 13.10.

Dist (4, centroid 1) = $\sqrt{(10-4)^2 + (4-6)^2} = \sqrt{40}$

Dist (4, centroid 2) = $\sqrt{(10-12)^2 + (4-4)^2} = \sqrt{4} = 2$

Object 4 is closer to the centroid of cluster 2 and hence assign it to the cluster table. Object 4 is in the same cluster. The final cluster table is shown in Table 13.11.

Obviously, Object 5 is in Cluster 3. Recompute the new centroids of cluster 1 and cluster 2. They are (4, 6) and (11, 4), respectively.

**Table 13.11:** Cluster Table After Iteration 1

| Cluster 1 | Cluster 2 |
|---|---|
| (4, 6) | (10, 4) |
| (2, 4) | (12, 4) |
| (6, 8) | |
| Centroid 1 (4, 6) | Centroid 2 (11, 4) |

The second iteration is started again with the Table 13.11.

Obviously, the point (4, 6) remains in cluster 1, as the distance of it with itself is 0. The remaining objects can be checked. Take the sample object 1 (2, 4) and compare with the centroid of the clusters in Table 13.12.

Dist (1, centroid 1) = $\sqrt{(2-4)^2 + (4-6)^2} = \sqrt{8}$

Dist (1, centroid 2) = $\sqrt{(2-11)^2 + (4-4)^2} = \sqrt{81} = 9$

Object 1 is closer to centroid of cluster 1 and hence remains in the same cluster. Take the sample object 3 (6, 8) and compare with the centroid values of clusters 1 (4, 6) and cluster 2 (11, 4) of the Table 13.12.

Dist (3, centroid 1) = $\sqrt{(6-4)^2 + (8-6)^2} = \sqrt{8}$

Dist (3, centroid 2) = $\sqrt{(6-11)^2 + (8-4)^2} = \sqrt{41}$

Object 3 is closer to centroid of cluster 1 and hence remains in the same cluster. Take the sample object 4 (10, 4) and compare with the centroid values of clusters 1 (4, 6) and cluster 2 (11, 4) of the Table 13.12:

Dist (4, centroid 1) = $\sqrt{(10-4)^2 + (4-6)^2} = \sqrt{40}$

Dist (3, centroid 2) = $\sqrt{(10-11)^2 + (4-4)^2} = \sqrt{1} = 1$

Object 3 is closer to centroid of cluster 2 and hence remains in the same cluster. Obviously, the sample (12, 4) is closer to its centroid as shown below:

Dist (5, centroid 1) = $\sqrt{(12-4)^2 + (4-6)^2} = \sqrt{68}$

Dist (5, centroid 2) = $\sqrt{(12-11)^2 + (4-4)^2} = \sqrt{1} = 1$. Therefore, it remains in the same cluster. Object 5 is taken as centroid point.

The final cluster Table 13.12 is given below:

**Table 13.12:** Cluster Table After Iteration 2

| Cluster 1 | Cluster 2 |
|---|---|
| (4, 6) | (10, 4) |
| (2, 4) | (12, 4) |
| (6, 8) | |
| Centroid (4, 6) | Centroid (11, 4) |

There is no change in the cluster Table 13.12. It is exactly the same; therefore, the *k*-means algorithm terminates with two clusters with data points as shown in the Table 13.12.

Scan for *'Additional Examples'*

## 13.5 DENSITY-BASED METHODS

Density-based spatial clustering of applications with noise (DBSCAN) is one of the density-based algorithms. Density of a region represents the region where many points above the specified threshold are present. In a density-based approach, the clusters are regarded as dense regions of objects that are separated by regions of low density such as noise. This is same as a human's intuitive way of observing clusters.

The concept of density and connectivity is based on the local distance of neighbours. The functioning of this algorithm is based on two parameters, the size of the neighbourhood ($\varepsilon$) and the minimum number of points ($m$).

1. Core point – A point is called a core point if it has more than specified number of points ($m$) within $\varepsilon$-neighbourhood.
2. Border point – A point is called a border point if it has fewer than '$m$' points but is a neighbour of a core point.