Chapter 2
APPLICATIONS OF CLUSTERING

2.1 Introduction
Clustering analysis has been applied in a wide variety of fields such as Biology, Medicine, Psychology, Economics, Sociology, and Astrophysics. The main goal of clustering analysis is to partition a given set of objects or data sets into homogeneous groups based on their features or characteristics such that objects within a group are more similar to each other and more different from those in other groups (Chen et al. 1996). Clustering can group the genes into biologically relevant clusters with similar expression patterns. The genes clustered together tend to be functionally related. Hence, clustering can reveal the co-expression of genes which were previously uncharacterized or unnoticed. In recent years, clustering analysis has become a valuable and useful tool for analysis of microarray or gene expression data. Eisen et al. (1998) applied a variant of hierarchical clustering to identify groups of co-expressed yeast genes. Alon et al. (1999) used a two-way clustering technique to detect groups of correlated genes and tissues. Self-organizing maps were used by Tamayo et al. (1999) to identify clusters in the yeast cell cycle data set and human hematopoietic differentiation data set.

In the subsequent sections we present applications of clustering in various fields in brief and existing and standard methods of clustering along with similarity measures.

2.2 Applications of clustering
Modern clustering owes its development to high speed computers and to a small group of numerical taxonomists. To apply computers, it is necessary to have precise definitions of meaning of cluster, the data type and similarity measures before using computers. In this section, we briefly list out clustering applications in various fields.
**Medicine**

The principle clustering problem in Medicine is the clustering of diseases. Webster’s dictionary defines a disease as an impairment of the normal state of living being that interrupts or modifies the performance of the vital functions, being a response to environmental factors (such as malnutrition, industrial hazards, climatic conditions etc.), to specific infective agents(such as worms, bacteria, viruses etc.) or to the inherent defects of the organism. Knusman and Toeller (1972) discover three groups of diabetes mellitus, using factor analysis (statistical tool). Winkel and Tygstrup (1971) identified two distinct groups of 400 cirrhosis patients, but left 70% the patients unclassified. Baron and Fraser (1968), in an experimental test of clustering methods on 50 cirrhosis patients measured on 330 characteristics, showed that single linkage algorithm conforms less well to previous diagnosis than average linkage algorithm. Hayhoe *et al.*, (1964) identified four groups in 140 cases of leukemia and proposed diagnostic criteria to distinguish four groups. Manning and Watson (1966) divided 99 heart patients into three groups agreeing substantially with physicians diagnoses of univalvular lesions, multivalvular lesions and arteriosclerotic, hypertensive, or pulmonary disease.

**Psychiatry**

Diseases of the mind are more elusive than diseases of the body, and the classification of such diseases is in an uncertain state. There is an agreement on the existence of paranoia, schizophrenia, and depression (such categories can be seen in Kant’s classification published in 1790), but clear diagnostic criteria are not available, as Katz *et al.* (eds.,1970) remark. A characteristic difficulty of classification of mental illness is the subjective, subtle, and variable character of the symptoms. Numerical techniques have gained more acceptances in this area than in medical diagnosis. One of the earliest known contributions to clustering, by Zubin(1938), discussed a method of discovering subgroups of schizophrenic patients. Some papers such as those of Everitt *et al.*, (1971) and of Paykel (1970) seek clusters of patients, others such as those of Hautaluoma (1971) and of Lorr *et al* (1963), seek clusters of symptoms (syndromes).

**Archaeology and Anthropology**

The field worker finds large number of objects such as stone tools, funeral objects, pieces of pottery, ceremonial statues, or skulls that he would like to divide into groups of
similar objects, each group produced by the same civilization. Clustering techniques are surveyed in Weiner and Huizinger (eds., 1972) and in Hodson et al (eds., 1971). Boyce (1969) studied a number of average linkage techniques on 20 skulls, and Hodson (1969,1970) considered a wide range of techniques on three interesting datasets – broaches, stone tools, and copper tools.

**Phytosociology**

Phytosociology concerns the plant and animal species. It bares the same relation to taxonomy that epidemiology bares to the classification of disease. Clustering detects similar quadrats as being the same type of habitation. An article by Whittaker (1962) contains a survey of traditional approaches. Lieth and Moore (1970) reorder the data matrix so that similar species are close in the new ordering and similar quadrats are close.

**Miscellaneous**

In Economics, Fisher (1969) considers input-output matrices in which the rows and columns have the same labels, so that the clustering of rows and columns must occur simultaneously. In Market Research Goronzy (1970) clusters firms by various financial and operating characteristics, while King (1966) does so by stock price behaviour. Frank and Green (1968) reviewed a number of interesting applications. In Linguistics Dyen et al (1967) used the proportion of matched words over a list of 196 meanings as a measure of the distance between two languages, with the aim of reconstructing an evolutionary tree of languages. Kaiser (1966) and Weaver and Hess (1963) consider numerical methods for establishing legislative districts. Abell (1960) finds clusters of galaxies by searching photographic plates of all high galactic latitudes. He lists 2712 such clusters and demonstrates that the clusters are not randomly distributed but exhibit further clustering them. Psychological applications are less common because of the dominance of factor analysis and multidimensional scaling, which are frequently interpreted as classifications. Miller (1969) has 50 Harvard students divide 48 nouns into categories according to similarity of meaning; the nouns are clustered into 5 groups, measuring similarity between two nouns by the proportion of students who place them in the same category. Wiley (1967) uses a factor-analysis-like technique on a similar data set.
2.3 Similarity and its measurements

The basic objective in clustering analysis is to discover natural groupings of the items (or variables). In turn, we must first develop a quantitative scale on which to measure the association between the objects. These scales are referred to as similarity measures and are mainly statistical measures that indicate the distances between each of the objects. This section presents various similarity measures.

Most efforts to produce a rather simple group structure from a complex data necessarily require a measure of “closeness” or “similarity”. Similarity can be understood as the quantity or state of being similar; likeliness or resemblance. Similarity is hard to define however real meaning of similarity is a philosophical question, but in clustering we have to adopt a pragmatic approach. We measure similarity based on features. Sometimes we are given the perfect or exact features to measure similarity. Most of the times we need to

- **Generate features**: Suppose we have to find similar people with regard to their medical conditions. Knowing both their height and weight is not helpful, but their Body Mass Index \[\text{BMI} = \text{Weight in Kilograms}/(\text{Height in meters})^2\].
- **Clean features**: Our features may contain noise or outliers and these are to be cleaned or eliminated.
- **Normalize features**: We may need to transform the features.
- **Reduce features**: We may have too many features, to do efficient similarity measurement, dimensionality reduction may be an essential step.

There is no single “magic” black box for measuring similarity; however there are two useful and general tricks viz., **Feature Projection and Edit distance**.

**Feature Projection**, as an example in Fishers iris data set we have three varieties of same plant species. We take the features as petal length and petal width and we project on to the feature space. **Edit distance**, The second method is to measure similarity between two objects is to measure similarity between two objects is to transform one object into other; and measure how much effort it takes. The measure of the effort or rather ‘cost’ of this transformation becomes similarity. There is often a great deal of subjectivity involved in the choice of similarity measure. Important considerations include the nature of variables (discrete, continuous, and binary) or scale of measurement (nominal, ordinal, interval, and ratio) and subject matter of knowledge.
2.3.1 Definition of Distance Measures: Let $X_i$ and $X_j$ be two objects from the universe of possible objects. The distance (dissimilarity) between $X_i$ and $X_j$ is the real number denoted by $\text{distance}(X_i, X_j)$ or $d(X_i, X_j)$. Two types of distance measures can be considered: Similarity Measures, and Dissimilarity Measures.

**Similarity Measure**

Similarity Measures are used to find similar pairs of objects from the given data set, $X$. Let $d(X_i, X_j)$ be a similar coefficient. If objects $X_i$ and $X_j$ are alike, then $d(X_i, X_j)$ becomes larger. Otherwise, $d(X_i, X_j)$ becomes smaller. For all objects $X_i$ and $X_j$, a similarity measure needs to satisfy the following conditions:

- $d(X_i, X_j) \geq 0$
- $d(X_i, X_i) = 0$
- $d(X_i, X_j) = d(X_j, X_i)$

- **Pearson’s correlation coefficient**: Pearson’s correlation coefficient measures the similarity between the shapes of two expression patterns. Given two objects $X_i$ and $X_j$, Pearson’s correlation coefficient is defined as:

$$
\text{Person}(X_i, X_j) = \frac{\sum_{d=1}^{n} (X_{id} - \mu_{oi})(X_{jd} - \mu_{oj})}{\sqrt{\sum_{d=1}^{n} (X_{id} - \mu_{oi})^2} \sqrt{\sum_{d=1}^{n} (X_{jd} - \mu_{oj})^2}}
$$

where $\mu_{oi}$ and $\mu_{oj}$ are the means for $X_i$ and $X_j$, respectively. Pearson’s correlation coefficient views each object as a random variable with observations and measures the similarity between two objects by calculating the linear relationship between the distributions of the two corresponding random variables. Pearson’s correlation coefficient is widely used and has proven effective as a similarity measure for gene expression data.

**Dissimilarity Measure**

Dissimilarity Measures are used to find dissimilar pairs of objects from the given data set, $X$. The dissimilarity coefficient, $d(X_i, X_j)$, are small when objects $X_i$ and $X_j$ are alike, otherwise, $d(X_i, X_j)$ become larger. As the similarity measures, the dissimilarity measures need to satisfy the following conditions:
Most of the clustering algorithms use dissimilarity measures to join, or to separate, objects. Some of the measures used in practice:

- **Euclidean Distance**: The Euclidean Distance between points $X_i$ and $X_j$ is given by:
  \[
d(X_i,X_j) = \sqrt{\sum_{k=1}^{n} (X_{ik} - X_{jk})^2}
\]
  where $X_{ik}$ and $X_{jk}$ are the $k^{th}$ coordinates of $X_i$ and $X_j$ respectively.

- **City-block (Manhattan Norm) distance**: The Manhattan Norm between points $X_i$ and $X_j$ is given by:
  \[
d(X_i,X_j) = \frac{1}{n} \sum_{i=1}^{n} |X_{ik} - X_{jk}|
\]
  where $X_{ik}$ and $X_{jk}$ are the $k^{th}$ coordinates of $X_i$ and $X_j$ respectively.

- **Minkowski Distance**: The Minkowski Distance between points $X_i$ and $X_j$ is given by:
  \[
d(X_i,X_j) = \left( \sum_{k=1}^{n} (X_{ik} - X_{jk})^q \right)^{1/q}
\]
  where $X_{ik}$ and $X_{jk}$ are the $k^{th}$ coordinates of $X_i$ and $X_j$ respectively.

- **Chebychev distance**: The Chebychev distance between points $X_i$ and $X_j$ is given by:
  \[
d(X_i,X_j) = \max(k=1,n) |X_{ik} - X_{jk}|
\]

**Similarity of Objects when attributes are binary**

When items cannot be represented by meaningful $n$-dimensional pairs, items are often compared on the basis of the presence or absence of certain characteristics. Similar items have more characteristics in common than dissimilar items. The presence or absence of a certain characteristics can be described mathematically by introducing a binary variable, which assume value 1 if the characteristic is present and value 0 if the characteristic is absent. Several similarity measures and the corresponding rationale are as follows.
Table 2.3.1 Rationale of various similarity measures

<table>
<thead>
<tr>
<th>S.No</th>
<th>Coefficient</th>
<th>Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(a+d)/n</td>
<td>Equal weights for 1-1 matches 0-0 matches</td>
</tr>
<tr>
<td>2</td>
<td>2(a+d)/(2(a+d)+b+c)</td>
<td>Double weight for 1-1 and 0-0 matches</td>
</tr>
<tr>
<td>3</td>
<td>(a+d)/(a+d+2(b+c))</td>
<td>Double weight for unmatched pairs</td>
</tr>
<tr>
<td>4</td>
<td>a/n</td>
<td>No 0-0 matches in numerator</td>
</tr>
<tr>
<td>5</td>
<td>a/(a+b+c)</td>
<td>0-0 matches are treated as irrelevant</td>
</tr>
<tr>
<td>6</td>
<td>2a/(2a+b+c)</td>
<td>No weight for 0-0 matches. Double weight for 1-1 matches</td>
</tr>
<tr>
<td>7</td>
<td>a/(a+2(b+c))</td>
<td>No 0-0 matches in numerator or denominator Double weight for unmatched pairs</td>
</tr>
<tr>
<td>8</td>
<td>a/(b+c)</td>
<td>Ratio of matches to mismatches with 0-0 matches excluded</td>
</tr>
</tbody>
</table>

Note: 
a- Number of attributes where both objects score is “1”; 
b- Number of attributes where object1’s score is “1” and object2’s score is “0”; 
c- Number of attributes where object1’s score is “0” and object2’s score is “1”; 
d- Number of attributes where both objects score is “0”. 
n- Number of attributes

2.4 Distance between Clusters

In connection with cluster analysis we need to distinguish among (i) Distance between two points (well known notion). (ii) Distance between a point to a cluster and (iii) Distance between two clusters.

Distance between a point to a cluster

Let \( d(X_i, X_j) \) be the distance between two points \( X_i \) and \( X_j \), then the distance between a point ‘\( X_i \)’ to a cluster \( C \), is defined as

\[
d(X_i, C) = \max_{X_j \in C} d(X_i, X_j) \quad \text{or} \quad d(X_i, C) = \min_{X_j \in C} d(X_i, X_j) \quad \text{or}
\]

\[
d(X_i, C) = \frac{1}{n} \sum_{X_j \in C} d(X_i, X_j) \ , \text{where } n \text{ is the number of elements in cluster } C
\]

Distance between two clusters

The distance between two clusters \( C_1 \) and \( C_2 \) is given by

\[
D(C_1, C_2) = \min_{X_i \in C_1, X_j \in C_2} d(X_i, X_j) \quad \text{or}
\]
\[ D(C_1, C_2) = \max_{x_i \in C_1, x_j \in C_2} d(x_i, x_j) \] or
\[ D(C_1, C_2) = \frac{1}{n_1 \times n_2} \sum_{x_i \in C_1, x_j \in C_2} d(x_i, x_j) \]

where \( n_1 \) is the number of elements in cluster \( C_1 \) and \( n_2 \) is the number of elements in cluster \( C_2 \).

### 2.5 Clustering Methods

Broadly the clustering algorithms are classified into (i) Hierarchical, (ii) Partitional, (iii) Density based, (iv) Grid based, and (v) Model based algorithms.

#### 2.5.1 Hierarchical Clustering

Hierarchical clustering Algorithm refers to a clustering process that organizes data into large groups, which contain smaller groups, and so on. A hierarchical clustering may be drawn as a tree or dendrogram. The finest grouping is at the bottom of the dendrogram; each sample by itself forms a cluster. The coarsest grouping is at the top of the dendrogram, where all samples are grouped into one cluster. In between there are various number of clusters. Hierarchical clustering algorithms are called agglomerative if they are built dendrograms from the bottom up, and they are called divisive if they build dendrograms from top down.

Agglomerative clustering algorithm is as follows

1. Begin with \( m \) clusters, each consisting of one sample.
2. Repeat step 3 a total of \( n-1 \) times
3. find the most similar clusters \( C_i \) and \( C_j \) and merge \( C_i \) and \( C_j \) into one cluster, if there is a tie, merge first pair found.

On the similar lines, we can write down division clustering algorithm. In turn the following are the various algorithms under the category of hierarchical clustering.

- **Single linkage**

  Single Linkage Algorithm is also known as the minimum method and the nearest neighbor method. The single-linkage algorithm is obtained by defining the distance between two clusters to be the minimum distance between two points such that one point in each cluster.

- **Complete linkage**
The Complete linkage algorithm is also called the maximum method or the farthest neighbor method. It is obtained by defining the distance between two clusters to be the largest distance between a sample in one cluster and a sample in the other cluster.

- **Average linkage**
  The average linkage algorithm is also known as Unweighted Pair Group Method using Arithmetic averages (UPGMA). It is obtained by defining the distance between the two clusters to be the average distance between a point in one cluster and a point in the other cluster.
- **Ward’s method**
  Ward’s method is also called the minimum-variance method. In the Ward’s Method the distance between two clusters \( C_i \) and \( C_j \), is defined as weighted version of the squared Euclidean distance of their mean vector. That is,

\[
D (C_i, C_j) = \frac{2|C_i||C_j|}{|C_i|+|C_j|}\|\mu_i - \mu_j\|^2
\]

Where \( \mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x \) and \( \mu_j = \frac{1}{|C_j|} \sum_{x \in C_j} x \)

List of other methods in this category are

- Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH)
- Clustering Using REpresentatives (CURE)
- RObust Clustering algorithms for boolean and categorical data (ROCK)
- Chameleon and so on

The detailed algorithms of the above hierarchical methods are readily available (c.f. Han and Kamber, 2004)

### 2.5.2 Partitional Clustering

Agglomerative clustering creates a series of nested clusters. This contrasts with partitional clustering in which the goal is usually to create one set of clusters that partitions data into similar groups. Samples close to one another are assumed to be similar and the goal of the partitional clustering algorithms is to group data that are close together.

The varieties of partitional algorithms are (i) Forgy’s algorithm (ii) k-means (iii) k-modes (iv) Partitioning Around Medoids (PAM) (v) Clustering LArge Applications (CLARA)
2.5.3 Density Based Clustering

The Density based methods typically regard clusters as dense regions of objects in the data space that are separated by regions of low density. The various algorithms in density based are

- Density Based Spatial Clustering Applications with Noise (DBSCAN)
- DENsity based CLUstEring (DENCLUE)
- Ordering Points To Identify the Clustering Structure (OPTICS) etc.

2.5.4 Grid Based Clustering

The Grid based approaches quantizes the object space into finite number of cells that form a grid structure on which all of the operations for clustering are performed. Some of the existing algorithms in Grid Based are

- STastical INformation Grid (STING)
- Clustering using Wavelet Transformation

2.5.5 Model Based Clustering Methods

These methods are often based on the assumption that the data are generated by a mixture of underlying probability distributions. The different methods in model based are

- COBWEB
- CLASSIT
- AutoClass
- Comparative learning and Self Organizing Maps

In this thesis we have concentrated on partitional clustering in the analysis of microarray data.