Chapter - IV

Techniques of Clustering Analysis
4.1 Introduction to Cluster Analysis

Clustering is similar to classification in that data are grouped. However, unlike classification, the groups are not predefined. Instead, the grouping is accomplished by finding similarities between data according to characteristics found in the actual data. The groups are called Clusters. Some authors view clustering as a special type of classification.

Definition

"Set of like elements. Elements from different clusters are not alike."

OR

"The Distance between points in a cluster is less than the distance between a point in the cluster and any point outside it."

A term similar to clustering is database segmentation, where similar tuples in a database are grouped together. This is done to partition or segment the database into components that give the user a more general view of data.

Clustering has been used in many applications domains, including biology, medicine, anthropology, marketing, and economics. Clustering applications include plant and animal classification, disease classification, image processing, pattern recognition, and document retrieval. One of the first domains in which clustering was used in biological taxonomy.

While clustering is applied to real world database, many interesting problems occur:

- Outlier handling is difficult.
- Dynamic data in the database implies that cluster membership may change over time.
- There is not one correct answer to clustering problem.
- Another related issue is that data should be used for clustering.
Some basic features of clustering:

- The number of clusters is not known.
- There may not be any a priori knowledge concerning the clusters.
- Cluster Results are dynamic.

4.2 Desired Features of Cluster Analysis

There are a large number of cluster analysis methods. A list of desired features that an ideal cluster analysis method should have following characteristics [1, 2]. The list is given below:

- **Scalability**

  Data mining problems can be large and therefore it is desirable that a cluster analysis method be able to deal with small as well as large problems gracefully. Ideally, the performance should be linear with the size of the data. The method should also scale well to databases in which the number of attributes is large.

- **Only One Scan of the Dataset**

  Data must be stored on the disk and the cost of the I/O from the disk can then become significant in solving the problem. It is desirable that a cluster analysis method not require more than one scan of the disk-resident data.

- **Ability to Stop and Resume**

  When the dataset is very large, cluster analysis requires considerable processor time to complete the task. In such cases, it is desirable that the task be able to be stopped and then resumed when convenient.

- **Minimal Input Parameter**

  The cluster analysis method should not expect too much guidance from the user. A data mining analyst may be working with a dataset about which his/her knowledge is limited. It is desirable that the user not be expected to have domain knowledge of the data and not be expected to possess insight into clusters that might exist in the data.
Robustness

Most data obtained from a variety of sources has errors. It is desirable that a cluster analysis method be able to deal with noise, outliers and missing values gracefully.

Ability to Discover Cluster Shapes

Clusters come in different shapes and not all clusters are spherical. It is desirable that a cluster analysis method be able to discover cluster shapes other than spherical.

Different Data Types

Many problems have a mixture of data types i.e. numerical, categorical and even textual. It is desirable that a cluster analysis method be able to deal with not only numerical data but also boolean and categorical data.

Result Independent of Data Input Order

It is desirable that a cluster analysis method not be sensitive to data input order. Whatever the order, the result of cluster analysis of the same data should be the same.

4.3 Types of Data

Datasets come in different forms. The data may be quantitative, binary, nominal or ordinal [4-8].

1. Quantitative data is quite common, such as weight, marks, height, price, etc.

2. Binary data is also quite common, such as gender, marital status.

3. Qualitative nominal data is similar to binary data which may take more than two values but has no natural order, for example religion, foods or colors.

4. Qualitative ordinal data is similar to nominal data except that the data has an order associated with it, for example, grades A, B, C, D, sizes S, M, L, and XL. The problem of measuring distance between ordinal variables is
different than for nominal variables since the order of the values is important.

Other data types of data are also possible. For example, data may include text strings or a sequence of web pages visited. Cluster analysis methods do not accept categorical data [3].

4.4 Types of Cluster Analysis Methods

There are many clustering methods available and each of them may give a different grouping of a dataset. The choice of a particular method will depend on the type of output desired, the known performance of method with particular types of data, the hardware and software facilities available and the size of the dataset. In general, clustering methods may be divided into two categories based on the cluster structure which they produce. The non-hierarchical methods divide a dataset of N objects into M clusters, with or without overlap [9-10].

Classification of different clustering algorithms is shown in following figure. Clustering algorithms may be viewed as hierarchical or partitional.

![Classification of Clustering Methods](image)

Figure 4.1: Classification of Clustering Methods

a) Hierarchical Methods

Hierarchical methods obtain a nested partition of the objects resulting in tree of clusters. These methods either start with one cluster and then spilt into smaller and smaller clusters (called Divisive) or start with each object in an
individual cluster and then try to merge similar clusters into larger and large clusters (called agglomerative). With hierarchical clustering, the desired number of clusters is not input.

There are two approach of hierarchical methods are possible i.e. Agglomerative approach and Divisive approach.

The **Agglomerative Approach** is the bottom-up approach, where each object at the start is a cluster by itself and the nearby clusters are repeatedly merged resulting in larger and larger clusters until some stopping criterion is met or all the objects are merged into a single large cluster which is the highest level of the hierarchy.

The **Divisive Approach** is the top-down approach; all the objects are put in a single cluster to start. The method then repeatedly performs splitting of clusters resulting in smaller and smaller clusters until a stopping criterion is reached or each cluster has only one object in it.

**Advantages of Hierarchical Approach**

1. The hierarchical approach can provide more insight into the data by showing a hierarchy of clusters than flat cluster structure created by a partitioning method like the K-Means method.
2. Hierarchical methods are conceptually simpler and can be implanted easily.
3. In some applications only proximity data is available and then the hierarchical approach may be better.
4. Hierarchical methods can provide clusters at different levels of granularity.

**Disadvantages of Hierarchical Approach**

- The Hierarchical methods do not include a mechanism by which objects that have been incorrectly put in a cluster may be reassigned to another cluster.
- The time complexity of hierarchical methods can be shown to be O(n^3).
The distance matrix requires $O(n^3)$ space and becomes very large for a large number of objects.

Different distance metrics and scaling of data can significantly change the results.

b) Partitional Clustering

With partitional clustering, the algorithm creates only one set of clusters. These approaches use the desired number of clusters to derive how the final set is created. It is assumed that each cluster has at least one object and each object belongs to only one cluster. Given objects, these methods make $k \leq n$ clusters of data and use an iterative relocation method [2,3].

K-Means Method

K-Means is the simplest and most popular classical clustering method that is easy to implement. The Classical method can only be used if the data about all the objects is located in the main memory. The method is called K-Means since each of the clusters is represented by the mean of the objects within it. It is also called the centroid methods since at each step the centroid point of each cluster is assumed to be known and each of the remaining points are allocated to the cluster whose centroid is close to it.

1. The K-Means method may be described as follows:

   Select the number of clusters. Let this number be $k$.

2. Pick $k$ seeds as centroids of the $k$ clusters. The seeds may be picked randomly unless the user has some insight into the data.

3. Compute the Euclidean distance of each object in the dataset from each of the centroids.

4. Allocate each object to the cluster it is nearest to based on the distances computed in the previous step.

5. Compute the centroids of the clusters by computing the means of the attribute values of the objects in the cluster.
6. Check if the stopping criterion has been met. If yes, go to step 7. If not, go to step 3.

7. One may decide to stop sat this stage or to spilt a cluster or combine two clusters heuristically until a stopping criterion is met.

The Method is scalable and efficient, and is guaranteed to find a local minimum. Although the K-Means method is most widely known and used, there are a number of issues related to the method that should be understood.

1. The K-Means method needs to compute Euclidean distances and of the attribute values of objects within a cluster.

2. The K-Means method implicitly assumes spherical probability distributions.

3. The K-Means method can be sensitive to outliers.

4. The K-Means method does not deal with overlapping clusters [2-8].

4.5 Different Clusterers

Package weka.clusterers

<table>
<thead>
<tr>
<th>Class Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clusterer</td>
</tr>
<tr>
<td>DensityBasedClusterer</td>
</tr>
<tr>
<td>EM</td>
</tr>
<tr>
<td>FarthestFirst</td>
</tr>
<tr>
<td>MakeDensityBasedClusterer</td>
</tr>
<tr>
<td>SimpleKMeans</td>
</tr>
</tbody>
</table>
1. **DensityBasedClusterer**

   Abstract clustering model that produces (for each test instance) an estimate of the membership in each cluster (i.e. a probability distribution).

   ```java
   java.lang.Object
   L weka.clusterers.Clusterer
   L weka.clusterers.DensityBasedClusterer
   ```

1. **EM Simple EM (Expectation Maximisation) Class**

   EM assigns a probability distribution to each instance which indicates the probability of it belonging to each of the clusters. EM can decide how many clusters to create by cross validation, or you may specify apriori how many clusters to generate.

   The cross validation performed to determine the number of clusters is done in the following steps:

   1. The number of clusters is set to 1.
   2. The training set is split randomly into 10 folds.
   3. EM is performed 10 times using the 10 folds the usual CV way.
   4. The loglikelihood is averaged over all 10 results.
   5. If loglikelihood has increased the number of clusters is increased by 1 and the program continues at step 2.

   The number of folds is fixed to 10, as long as the number of instances in the training set is not smaller than 10. If this is the case the number of folds is set equal to the number of instances.

   Valid options are:

   - `-N` Specify the number of clusters to generate. If omitted, EM will use cross validation to select the number of clusters automatically.
   - `-I` Terminate after this many iterations if EM has not converged.
   - `-S` Specify random number seed.
   - `-M` Set the minimum allowable standard deviation for normal density calculation.
1. FarthestFirst

Implements the “Farthest First Traversal Algorithm” by Hochbaum and Shmoys 1985: A best possible heuristic for the k-centre problem

Valid options are:

-N Specify the number of clusters to generate.
-S Specify random number seed.

2. MakeDensityBasedClusterer

java.lang.Object

-weka.clusterers.Clusterer

-weka.clusterers.DensityBasedClusterer

-weka.clusterers.MakeDensityBasedClusterer

Class for wrapping a Clusterer to make it return a distribution and density. Fits normal distributions and discrete distributions within each cluster produced by the wrapped clusterer, it Supports the NumberOfClustersRequestable interface.

1. SimpleKMeans

java.lang.Object

-weka.clusterers.Clusterer

-weka.clusterers.SimpleKMeans

Simple k means clustering class. Valid options are:

-N Specify the number of clusters to generate.
-S Specify random number seed.

4.6 Data Clustering and Its Applications

Clustering is a main task of explorative data mining and a common technique for statistical data analysis used in many fields, including machine
learning, pattern recognition, image analysis, information retrieval, and bioinformatics.

Fast retrieval of the relevant information from the databases has always been a significant issue. Different techniques have been developed for this purpose; one of them is Data Clustering. In this chapter Data Clustering is discussed along with its two traditional approaches and their analysis. Some applications of Data Clustering like Data Mining using Data Clustering.

Creating a Cluster

When a system administrator wishes to create a new cluster, the administrator will run a cluster installation utility on the system to become the first member of the cluster. For a new cluster, the database is created and the initial cluster member is added. The administrator will then configure any devices that are to be managed by the cluster software. We now have a cluster with a single member. In the next step of clustering, each node is added to the cluster by means of similarity on the basis of the resources used. The new node automatically receives a copy of the existing cluster database.

Joining a Cluster

Following a restart of a system, the cluster service is started automatically. The system configures and mounts local, non-shared devices. Cluster-wide devices must be left offline while booting because another node may be using them. The system uses a ‘discovery’ process to find the other members of the cluster.

Leaving a Cluster

When leaving a cluster, a cluster member will send a ClusterExit message to all other members on the cluster, notifying them of its intent to leave the cluster. The exiting cluster member does not wait for any responses and immediately proceeds to shutdown all resources and close all connections managed by the cluster software.
Sending a message to the other systems in the cluster when leaving saves the other systems from discovering the absence and having to go to a regroup effort to re-establish the membership.

4.6.1 Introduction

Data clustering is a method in which we make cluster of objects that are somehow similar in characteristics. The criterion for checking the similarity is implementation dependent.

Clustering is often confused with classification, but there is some difference between the two. In classification, the objects are assigned to predefined classes, whereas in clustering the classes are also to be defined.

Precisely, Data Clustering is a technique in which, the information that is logically similar is physically stored together. In order to increase the efficiency in the database systems the number of disk accesses are to be minimized. In clustering, the objects of similar properties are placed in one class of objects and a single access to the disk makes the entire class available.

4.6.1.1 Example to Elaborate the Idea of Clustering*

In order to elaborate the concept a little bit, let us take the example of the library system. In a library, books concerning to a large variety of topics are available. They are always kept in form of clusters. The books that have some kind of similarities among them are placed in one cluster. For example, books on the database are kept in one shelf and books on operating systems are kept in another cupboard, and so on. To further reduce the complexity, the books that cover same kind of topics are placed in same shelf, then the shelf and the cupboards are labelled with the relative name. Now when a user wants a book of specific kind on specific topic, he or she would only have to go to that particular shelf and check for the book rather than checking in the entire library.

4.6.2 Definitions of Different Parameters Used for Clustering

In clustering frequently used terms are defined as follows:

Cluster
A cluster is an ordered list of objects, which have some common characteristics.

**Clustering** is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters.

**Distance Between Two Clusters**

The distance between two clusters involves some or all elements of the two clusters. The clustering method determines how the distance should be computed [1].

**Similarity**

A similarity measure SIMILAR (Di, Dj) can be used to represent the similarity between the documents. Typical similarity generates values of 0 for documents exhibiting no agreement among the assigned indexed terms, and 1 when perfect agreement is detected. Intermediate values are obtained for cases of partial agreement [1].

**Average Similarity**

If the similarity measure is computed for all pairs of documents (Di, Dj) except when i=j, an average value AVERAGE SIMILARITY is obtainable. Specifically, AVERAGE SIMILARITY = CONSTANT SIMILAR (Di, Dj), where i=1,2,...,n and j=1,2,...,n and i <> j

**Threshold**

The lowest possible input value of similarity required to join two objects in one cluster.

**Similarity Matrix**

Similarity between objects calculated by the function SIMILAR (Di,Dj), represented in the form of a matrix is called a similarity matrix.

**Dissimilarity Coefficient**
The dissimilarity coefficient of two clusters is defined to be the distance between them. The smaller the value of dissimilarity coefficient, the more similar two clusters are.

Cluster Seed

First document or object of a cluster is defined as the initiator of that cluster i.e. every incoming object's similarity is compared with the initiator. The initiator is called the cluster seed.

4.7 Types of Clustering Methods

There are many clustering methods available, and each of them may give a different grouping of a dataset. The choice of a particular method will depend on the type of output desired. The known performance of method with particular types of data, the hardware and software facilities available and the size of the dataset. In general, clustering methods may be divided into two categories based on the cluster structure which they produce. The non-hierarchical methods divide a dataset of N objects into M clusters, with or without overlap.

These methods are sometimes divided into partitioning methods, in which the classes are mutually exclusive, and the less common clumping method, in which overlap is allowed. Each object is a member of the cluster with which it is most similar; however the threshold of similarity has to be defined. The hierarchical methods produce a set of nested clusters in which each pair of objects or clusters is progressively nested in a larger cluster until only one cluster remains. The hierarchical methods can be further divided into agglomerative or divisive methods. In agglomerative methods, the hierarchy is build up in a series of N-1 agglomerations, or Fusion, of pairs of objects, beginning with the un-clustered dataset. The less common divisive methods begin with all objects in a single cluster and at each of N-1 steps divide some clusters into two smaller clusters, until each object resides in its own cluster.

Some of the important Data Clustering Methods are described below.

4.7.1 Partitioning Methods

The partitioning methods generally result in a set of M clusters, each object belonging to one cluster. Each cluster may be represented by a centroid or a
cluster representative; this is some sort of summary description of all the objects contained in a cluster. The precise form of this description will depend on the type of the object which is being clustered. In case where real-valued data is available, the arithmetic mean of the attribute vectors for all objects within a cluster provides an appropriate representative; alternative types of centroid may be required in other cases, e.g., a cluster of documents can be represented by a list of those keywords that occur in some minimum number of documents within a cluster. If the number of the clusters is large, the centroids can be further clustered to produce hierarchy within a dataset.

Single Pass

A very simple partition method, the single pass method creates a partitioned dataset as follows:

1. Make the first object the centroid for the first cluster.
2. For the next object, calculate the similarity, S, with each existing cluster centroid, using some similarity coefficient.
3. If the highest calculated S is greater than some specified threshold value, add the object to the corresponding cluster and re-determine the centroid; otherwise, use the object to initiate a new cluster. If any objects remain to be clustered, return to step 2.

As its name implies, this method requires only one pass through the dataset; the time requirements are typically of order $O(N\log N)$ for order $O(\log N)$ clusters. This makes it a very efficient clustering method for a serial processor. A disadvantage is that the resulting clusters are not independent of the order in which the documents are processed, with the first clusters formed usually being larger than those created later in the clustering run.

4.7.2 Hierarchical Agglomerative Methods

The hierarchical agglomerative clustering methods are most commonly used. The construction of an hierarchical agglomerative classification can be achieved by the following general algorithm.

1. Find the 2 closest objects and merge them into a cluster
2. Find and merge the next two closest points, where a point is either an individual object or a cluster of objects.

3. If more than one cluster remains, return to step 2

Individual methods are characterized by the definition used for identification of the closest pair of points, and by the means used to describe the new cluster when two clusters are merged.

There are some general approaches for the implementation of this algorithm, these being stored matrix and stored data, are discussed below:

- In the second matrix approach, an N*N matrix containing all pairwise distance values is first created, and updated as new clusters are formed. This approach has at least an \( O(n^2) \) time requirement, rising to \( O(n^3) \) if a simple serial scan of dissimilarity matrix is used to identify the points which need to be fused in each agglomeration, a serious limitation for large N.

- The stored data approach required the recalculation of pairwise dissimilarity values for each of the N-1 agglomerations, and the \( O(N) \) space requirement is therefore achieved at the expense of an \( O(N^3) \) time requirement.

4.7.3 The Single Link Method (SLINK)

The single link method is probably the best known of the hierarchical methods and operates by joining, at each step, the two most similar objects, which are not yet in the same cluster. The name single link thus refers to the joining of pairs of clusters by the single shortest link between them.

4.7.4 The Complete Link Method (CLINK)

The complete link method is similar to the single link method except that it uses the least similar pair between two clusters to determine the inter-cluster similarity (so that every cluster member is more like the furthest member of its own cluster than the furthest item in any other cluster). This method is characterized by small, tightly bound clusters.

4.7.5 The Group Average Method
The group average method relies on the average value of the pair wise within a cluster, rather than the maximum or minimum similarity as with the single link or the complete link methods. Since all objects in a cluster contribute to the inter-cluster similarity, each object is, on average more like every other member of its own cluster than the objects in any other cluster.

4.7.6 Text Based Documents

In the text based documents, the clusters may be made by considering the similarity as some of the key words that are found for a minimum number of times in a document. Now when a query comes regarding a typical word then instead of checking the entire database, only that cluster is scanned which has that word in the list of its key words and the result is given. The order of the documents received in the result is dependent on the number of times that key word appears in the document.

4.8 Clustering Algorithms

Clusters found by different algorithms vary significantly in their properties, and understanding these “cluster models” is key to understanding the differences between the various algorithms. Typical cluster models include:

- **Connectivity Models**: for example hierarchical clustering builds models based on distance connectivity.

- **Centroid Models**: for example the k-means algorithm represents each cluster by a single mean vector.

- **Distribution Models**: clusters are modeled using statistic distributions, such as multivariate normal distributions used by the Expectation-maximization algorithm.

- **Density Models**: for example DBSCAN and OPTICS defines clusters as connected dense regions in the data space.

- **Subspace Models**: in Biclustering (also known as Co-clustering or two-mode-clustering), clusters are modeled with both cluster members and relevant attributes.
Techniques of Clustering Analysis / 126

➢ **Group Models:** some algorithms (unfortunately) do not provide a refined model for their results and just provide the grouping information.

➢ **Graph-Based Models:** a clique, i.e., a subset of nodes in a graph such that every two nodes in the subset are connected by an edge

- **Connectivity based clustering (hierarchical clustering)**

Connectivity based clustering, also known as hierarchical clustering, is based on the core idea of objects being more related to nearby objects than to objects farther away.

Popular choices are known as single-linkage clustering (the minimum of object distances), complete linkage clustering (the maximum of object distances) or UPGMA ("Unweighted Pair Group Method with Arithmetic Mean", also known as average linkage clustering). Furthermore, hierarchical clustering can be agglomerative (starting with single elements and aggregating them into clusters) or divisive (starting with the complete data set and dividing it into partitions).

- **Centroid-based clustering**

In centroid-based clustering, clusters are represented by a central vector, which may not necessarily be a member of the data set. When the number of clusters are fixed to k, k-means clustering gives a formal definition as an optimization problem: find the k cluster centres and assign the objects to the nearest cluster centre, such that the squared distances from the cluster are minimized.

- **Distribution-based Clustering**

The clustering model most closely related to statistics is based on distribution models. Clusters can then easily be defined as objects belonging most likely to the same distribution. The most prominent method is known as expectation-maximization algorithm (or short: *EM-clustering*). Here, the data set is usually modeled with a fixed (to avoid over fitting) number of Gaussian distributions that are initialized randomly and whose parameters are iteratively optimized to fit better to the data set. Distribution-based clustering is a semantically strong method, as it not only provides you with clusters, but also
produces complex models for the clusters that can also capture correlation and dependence of attributes. The use of these algorithms puts an extra burden on the user: to choose appropriate data models to optimize, and for many real data sets, there may be no mathematical model available the algorithm is able to optimize (e.g. assuming Gaussian distributions is a rather strong assumption on the data).

- **Density-based clustering**

  In density-based clustering, clusters are defined as areas of higher density than the remainder of the data set. The most popular density based clustering method is DBSCAN. It only connects points that satisfy a density criterion, in the original variant defined as a minimum number of other objects within this radius. A cluster consists of all density-connected objects (which can form a cluster of an arbitrary shape, in contrast to many other methods) for example DBSCAN and OPTICS defines clusters as connected dense regions in the data space [11-16].

- **Newer developments**

  CLARANS (Ng and Han, 1994), and BIRCH (Zhang et al., 1996).

4.9 **Clustering Applications**

A “clustering” is essentially a set of such clusters, usually containing all objects in the data set. Data clustering has immense number of applications in every field of life. One has to cluster a lot of thing on the basis of similarity either consciously or unconsciously. So the history of data clustering is old as the history of mankind.

In computer field also, use of data clustering has its own value. Specially in the field of information retrieval data clustering plays an important role. Some of the applications are listed below.

4.9.1 **Similarity Searching in Medical Image Database**

This is a major application of the clustering technique. In order to detect many diseases like Tumor etc, the scanned pictures or the x-rays are compared with the existing ones and the dissimilarities are recognized.

We have clusters of images of different parts of the body. For example, the images of the CT Scan of brain are kept in one cluster. To further arrange things,
the images in which the right side of the brain is damaged are kept in one cluster. The hierarchical clustering is used. The stored images have already been analyzed and a record is associated with each image. In this form a large database of images is maintained using the hierarchical clustering.

Now when a new query image comes, it is firstly recognized that what particular cluster this image belongs, and then by similarity matching with a healthy image of that specific cluster the main damaged portion or the diseased portion is recognized. Then the image is sent to that specific cluster and matched with all the images in that particular cluster. Now the image with which the query image has the most similarities, is retrieved and the record associated to that image is also associated to the query image. This means that now the disease of the query image has been detected. Using this technique and some really precise methods for the pattern matching, diseases like really fine tumor can also be detected.

So by using clustering an enormous amount of time in finding the exact match from the database is reduced.

4.9.2 Data Mining

Another important application of clustering is in the field of data mining. Data mining is defined as follows:

**Definition 1:** "Data mining is the process of discovering meaningful new correlation, patterns and trends by sifting through large amounts of data, using pattern recognition technologies as well as statistical and mathematical techniques." [5]

**Definition 2:** Data mining is a "knowledge discovery process of extracting previously unknown, actionable information from very large databases." [5]

**Use of Clustering in Data Mining:** Clustering is often one of the first steps in data mining analysis. It identifies groups of related records that can be used as a starting point for exploring further relationships. This technique supports the development of population segmentation models, such as demographic-based customer segmentation. Additional analyses using standard analytical and other data mining techniques can determine the characteristics of these segments with
respect to some desired outcome. For example, the buying habits of multiple population segments might be compared to determine which segments to target for a new sales campaign.

For example, a company that sells a variety of products may need to know about the sale of all of their products in order to check that what product is giving extensive sale and which is lacking. This is done by data mining techniques. But if the system clusters the products that are giving fewer sales then only the cluster of such products would have to be checked rather than comparing the sales value of all the products. This is actually to facilitate the mining process.

4.10 Concluding remark:

This study comprises the basic concept of clustering by first providing the definition of clustering and then the definition of some related terms. We give some examples to elaborate the concept. Then different approaches to data clustering and also discussed some algorithms to implement that approaches. The partitioning method and hierarchical method of clustering were explained. The applications of clustering are also discussed with the examples of medical images database, data mining using data clustering.

So this study tries to prove the importance of clustering in every area of data mining technique. Clustering is not something really typical to databases but it has its applications in the fields like networking.
REFERENCES


