**UNIT-5**

**Cluster Analysis**

**Contents:**

* **Basic Concepts and Algorithms:**
* **Clustering for Understanding**
* **Biology**
* **Information Retrieval**
* **Climate**
* **Psychology and Medicine**
* **Business**
* **Clustering for Utility**
* **Summarization**
* **Compression**
* **Efficiently Finding Nearest Neighbors**
* **Overview:**

**2.1 What is Cluster Analysis?**

**2.2 Different types of Clustering’s**

* **Hierarchical versus Partitional**
* **Exclusive versus Overlapping versus Fuzzy**
* **Complete versus Partial**

**2.3 Different types of Clusters**

* **Well- Separated**
* **Prototype-Based**
* **Graph Based**
* **Density Based**
* **Shared-Property (Conceptual Clusters)**
* **K-Means**

**3.1 The Basic K-Means Algorithm**

* **Assigning Points to the closest centroid**
* **Centroid and Objective Functions**
* **Data in Euclidean Space**
* **Document Data**
* **Choosing Initial Centroid**
* **Time and Space Complexity**

**3.2 K-Means: Additional Issues**

* **Handling Empty Clusters**
* **Outliers**
* **Reducing the SSE with Post processing**
* **Updating Centroid Incrementally**

**3.3 Bisecting K-Means**

**3.4 K-Means and different types of Clusters**

**3.5 Strengths and Weaknesses**

**3.6 K-means as an Optimization Problem**

* **Agglomerative Hierarchical Clustering**

**4.1 Basic Agglomerative Hierarchical Clustering Algorithm**

* **Defining Proximity between clusters**
* **Time and Space Complexity**

**4.2 Specific Techniques**

* **Sample data**
* **Single Link or MIN**
* **Complete Link or MAX or Clique**
* **Group Average**
* **Ward Method and centroid Methods**

**4.3 The Lance-Williams Formula for Cluster Proximity**

**4.4 Key Issues in Hierarchical Clustering**

**4.5 Strengths and Weaknesses**

**5. DBSCAN**

**5.1 Traditional Density: Center-Based Approach**

**5.2 The DBSCAN Algorithm**

**5.3 Strengths and Weaknesses**

**6. Cluster Evaluation:**

**6.1 Overview**

**6.2Unsupervised Cluster Evaluation using Cohesion and Separation**

**6.3 Unsupervised Cluster Evaluation using the Proximity Matrix**

**6.4 Unsupervised Evaluation of Hierarchical Clustering**

**6.5Determining the correct number of clusters**

**6.6 Clustering Tendency**

**6.7 Supervised Measures of Cluster Validity**

**6.8 Assessing the Significance of Cluster Validity Measures**

* **Basic Concepts and Algorithms:**

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters).

* **Clustering for Understanding**

Classes or conceptually meaningful groups of objects that share common characteristics, play an important role in how people analyze and describe the world.

* Dividing objects into groups (clustering)
* Assigning Particular objects to these groups (Classification)

**Examples:**

* **Biology**
* Biologists have spent many years creating taxonomy (hierarchical Classification) of living things all: kingdom, phylum, class, order, family genes, and species.
* More Recently, Biologists have applied clustering to analyze the large amounts of genetic information that are now available. For example, clustering has been used to find groups of genes that have similar functions**.**
* **Information Retrieval:**
* The WWW consists of billions of web pages, and the results of a query to a search engine can return of thousands of pages.
* Clustering can be used to group these results into a small number of clusters, each of captures a particular aspect of a query.
* For example, a query of “movie” might returns web pages grouped into categories such as reviews, trailers, stars and theaters. Each category can be divided into sub categories.
* **Climate:**
* Understanding the Earths climate requires finding patterns in the atmosphere and ocean. So here cluster analysis has been to find the patterns in the atmospheric pressure of polar regions and areas of the ocean that have a significant impact on land climate.
* **Psychology and Medicine:**
* An illness or condition frequently has a number of variations, and cluster analysis can be used to identify these sub categories.
* For example, clustering has been used to identify different types of depression. Cluster analysis can also be used to detect patterns in the spatial or temporal distribution of a disease.
* **Business:**
* Business collect large amounts of information on current and potential customers. Clustering can be used to segment customers into a small number of groups for additional analysis and marketing activities.
* **Clustering for Utility:**
* Cluster analysis provides an abstraction from individual data objects to the clusters in which those data objects reside.
* Additionally, some clustering techniques characterize each cluster in terms of cluster prototype i.e., a data object that is representative of the other objects in the cluster.
* These cluster prototypes can be used as the basis for a number of data analysis or data processing techniques.

**Techniques:**

* **Summarization:**
* Many data analysis techniques, such as regression or PCA, have time or space complexity of O(m2) or higher (where m is the number of objects).
* However, instead of applying the algorithm to the entire data set, it can be applied to a reduced data set consisting only of cluster prototypes.
* Depending on the type of analysis, the number of prototypes, and the accuracy with which the prototypes represent the data, the results can be comparable to those that would have been obtained if all the data could have been used.
* **Compression:**
* Cluster prototypes can also be used for data compression. In particular, a table is created that consists of the prototypes for each cluster ie., each prototype is assigned an integer value that it is position (index) in the table.
* This type of compression is known as vector quantization and is often applied to image, sound, video data.
* **Efficiently Finding Nearing Neighbors:**
* Finding nearest neighbors can require computing the pairwise distance between all points. Often clusters and their cluster prototypes can be found much more efficiently.
* If objects are relatively close to the prototype of their cluster, then we can use prototypes to reduce the number of distance computations that are necessary to find nearest neighbors of an object.