SNS COLLEGE OF TECHNOLOGY

Coimbatore-35.

An Autonomous Institution

COURSE NAME : 19CST203 - DATA ANALYTICS

II YEAR /IV SEMESTER

CLUSTER VALIDATION:

To find good clustering partitions for a data set, regardless of the clustering algorithm used, the quality of the partitions must be evaluated. In contrast to the classification task, the identification of the best clustering algorithm for a given data set lacks a clear definition.

For predictive tasks like classification, the evaluation of predictive models has a clear meaning: how well the predictive models classify objects. The same cannot be said for clustering partitions, since there are many definitions of what is a good partition. However, some validation measures are frequently used in clustering tasks.

Several cluster validity criteria have been proposed; some automatic and others using expert input. The automatic validation measures for clustering partition evaluation can be roughly divided into three catagories:

- **External indices**:The external criteria uses external information, such as class label, if available, to define the quality of the clusters in a given partition. Two of the most common external measures are the correct-RAND and Jaccard

- **Internal indices**: The internal criteria looks for compactness inside each cluster and/or separation between different clusters. Two of the most common internal measures are the silhouette index, which measures both compactness and separation, and the within-groups sum of squares, which only measures compactness.

- **Relative indices**: The relative criterion compares partitions found by two or more clustering techniques or by different runs of the same technique.

The Jaccard external index, the silhouette and the within-groups sum of squares measures, both internal indices, are discussed next.

CLUSTERING TECHNIQUES:

Since we already know how to measure similarity between records, images, and words, we can proceed to see how to obtain groups/clusters using clustering techniques. There are hundreds of clustering techniques, which can be categorized in various ways. One of them is how the partitions are created, which defines how the data are divided into groups. Most techniques define partitions in one step (partitional clustering), while others progressively define partitions, either increasing or decreasing the number of clusters (hierarchical clustering).

Another criteria is the approach used to define what a cluster is, determining the elements to be included in the same cluster. According to this criterion, the main types of clusters are:

- **Separation-based**: each object in the cluster is closer to every other object in the cluster than to any object outside the cluster

- **Prototype-based**: each object in the cluster is closer to a prototype representing the cluster than to a prototype representing any other cluster

- **Graph-based**: represents the data set by a graph structure associating each node with an object and connecting objects that belong to the same cluster with an edge

- **Density-based**: a cluster is a region where the objects have a high number of close neighbors (i.e. a dense region), surrounded by a region of low density

- **Shared-property**: a cluster is a group of objects that share a property

We will present next three different clustering methods representative of different approaches to do clustering. None is better than the others. Each has its own pros and cons, as we will see. These three methods are:

- **K-means:** the most popular clustering algorithm and a representative of partitional and prototype-based clustering methods

- **DBSCAN**: another partitional clustering method, but in this case density-based

- **Agglomerative hierarchical clustering**: a representative of hierarchical and graph-based clustering methods.