SNS COLLEGE OF TECHNOLOGY

Coimbatore-35.

An Autonomous Institution

COURSE NAME : 19CST203 - DATA ANALYTICS

II YEAR /IV SEMESTER

CLUSTERING:

Clustering is a type of unsupervised learning method of machine learning. In the unsupervised learning method, the inferences are drawn from the data sets which do not contain labelled output variable. It is an exploratory data analysis technique that allows us to analyze the multivariate data sets.

Clustering is a task of dividing the data sets into a certain number of clusters in such a manner that the data points belonging to a cluster have similar characteristics. Clusters are nothing but the grouping of data points such that the distance between the data points within the clusters is minimal. Clustering is done to segregate the groups with similar traits.

CLUSTERING METHODS:

1. **Partitioning-based Clustering**

   Partitioning objects into k number of clusters where each partition makes/represents one cluster, these clusters hold certain properties such as each cluster should consist of at least one data object and each data object should be classified to exactly one cluster.

   These methods are broadly classified to optimize a targeted benchmark similarity function such that distance becomes a significant parameter to consider first. The examples are;

- K-means clustering, (understand K-means clustering from here in detail)
- CLARANS (Clustering Large Applications based upon Randomized Search)

Moreover, Partitioning clustering algorithms are the form of non-hierarchical that generally handle statics sets with the aim of exploring the groups exhibited in data via optimization techniques of the objective function, making the quality of partition better repeatedly.

Partitioning-based clustering is highly efficient in terms of simplicity, proficiency, and easy to deploy, and computes all attainable clusters synchronously.

2. **Hierarchical-based Clustering**

Depending upon the hierarchy, these clustering methods create a cluster having a tree-type structure where each newly formed clusters are made using priorly formed clusters, and categorized into two categories: Agglomerative (bottom-up approach) and Divisive (top-down approach***).*** The examples of Hierarchical clustering are

- CURE (Clustering Using Representatives)
- BIRCH (Balanced Iterative Reducing Clustering and using Hierarchies)

The agglomerative clustering method is achieved by locating each point in a cluster, initially and then merging two points closest to it where points represent an individual object or cluster of objects. The divisive clustering first considers the complete population as one cluster and then segments into smaller groups.

3. **Density-based Clustering**

These methods of clustering recognize clusters of dense regions that possess some similarity and are distinct from low dense regions of the space. These methods have sufficient accuracy and the high ability to combine two clusters. Its examples include

- DBSCAN (Density-based Spatial Clustering of Applications with Noise)
- OPTICS (Ordering Points to Identify Clustering Structure)

These methods implement distance measures between the objects in order to cluster the objects. In most of the cases, clusters, produced using this method, are spherical in shape, so sometimes it becomes hard to identify arbitrary shaped clusters.

Moreover, clusters are produced in all directions as long as the density, residing neighbourhood, surpass some threshold.

Density-based methods save data sets from outliers, the entire density of a point is treated and deciphered for determining features or functions of a dataset that can impact a specific data point.

Some algorithms like OPTICS, DenStream, etc deploy the approach that automatically filtrates noise (outliers) and generates arbitrary shaped clusters.

3. Density-based Clustering

These methods of clustering recognize clusters of dense regions that possess some similarity and are distinct from low dense regions of the space. These methods have sufficient accuracy and the high ability to combine two clusters. Its examples include

- DBSCAN (Density-based Spatial Clustering of Applications with Noise)
- OPTICS (Ordering Points to Identify Clustering Structure)

These methods implement distance measures between the objects in order to cluster the objects. In most of the cases, clusters, produced using this method, are spherical in shape, so sometimes it becomes hard to identify arbitrary shaped clusters.

Moreover, clusters are produced in all directions as long as the density, residing neighbourhood, surpass some threshold.

Density-based methods save data sets from outliers, the entire density of a point is treated and deciphered for determining features or functions of a dataset that can impact a specific data point.

Some algorithms like OPTICS, DenStream, etc deploy the approach that automatically filtrates noise (outliers) and generates arbitrary shaped clusters.

4. Grid-based Clustering

This method follows a grid-like structure, i.e, data space is organized into a finite number of cells to design a grid-structure. Various clustering operations are conducted on such grids (i.e quantized space) and are quickly responsive and do not rely upon the quantity of data objects. Its examples are;

- STING (Statistical Information Grid),
- Wave cluster,

- CLIQUE (Clustering In Quest)

  Computing statistical measurements for the grids consequently increasing the speed of method extensively.

  Also, the performance of grid-based methods is proportional to the grid-size and demands very less space than the actual data stream.

5. **Model-based Clustering**

   These methods deploy a predefined mathematical model for fitting and later on optimizing the data while assuming that the data is hybrid in the form of probability distributions and compute the number of clusters on the basis of standard statistics.

   However, the noise and outliers are taken into account while calculating the standard statistics for having robust clustering. In order to form clusters, these clustering methods are classified into two categories: Statistical and Neural Network approach methods. Its examples are;

- MCLUST (Model-based Clustering)
- GMM (Gaussian Mixture Models)

  The model-based algorithms, that use statistical approaches, follow probability measures for determining clusters, and those algorithms that use neural-network approaches, input and output are associated with unit carrying weights.

2.DISTANCE MEASURE:

Before we define groups of similar data, we must agree on what is similar and what is not similar (dissimilar). We can represent the similarity between two objects by a single number. This will allow us to say, for a particular object, which other objects in the same data set are more similar and which are more dissimilar. A common approach to associate a number with the similarity (and dissimilarity) between two objects is to use distance measures. The most similar objects have the smallest distances between them, and the most dissimilar have the largest distances. The way we compute the distance between objects depends on the scale type of its attributes: whether they are quantitative or qualitative.

DIFFERENCES BETWEEN VALUES OF COMMON ATTRIBUTE TYPES

The difference between two values for the same attribute, here named a and b, will be denoted as d(a, b). For quantitative attributes, one can calculate the absolute difference:

d(a.b) = |a − b|

For example, the difference in age between Andrew (a = 55) and Carolina (b = 37) is |55 − 37| = 18. Note, that even if we change the order of the values (a = 37 and b = 55) the result is the same.

If the attribute type is qualitative, we use distance measures suitable for the given type. If the qualitative attribute has ordinal values, we can measure the difference in their positions as:

d(a, b)=(|posa − posb|)⁄(n − 1)

where n is the number of different values, and posa and posb are the positions of the values a and b, respectively, in a ranking of possible values.

In our data set, education level can be considered an ordinal attribute, with larger values meaning a higher level of education. Thus the distance between the education levels of Andrew and Caroline is |pos1 − pos5|⁄4 = |1 − 5|⁄4 = 1.

Note that ordinal attributes need not be expressed only by numbers. For example, the education level can have values such as “primary”, “high school”, “undergraduate”, “graduate” and “postgraduate”. However, these can be readily transformed into numbers

If a qualitative attribute has nominal values, in order to compute the distance between two values we simply determine if they are equal (in which case the difference, or dissimilarity, will be zero) or not (in

which case the difference will be one).

d(a, b) = {1 , if a ≠ b

0 , if a = b

3.DISTANCE MEASURES FOR OBJECTS WITH QUANTITATIVE ATTRIBUTES

Several distance measures are particular cases of the Minkowski distance. The Minkowski distance for two m-dimensional objects p and q with quantitative attributes is given by:

$$\left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{\frac{1}{p}}$$

where m is the number of attributes, while pk and qk are the values of the k th attribute for objects p and q, respectively. Variants are obtained using different values for r. For example, for the Manhattan distance, r = 1, and for the Euclidean distance, r = 2.

The Manhattan distance is also known as the city block or taxicab distance, since if you are in a city and want to go from one place to another, it will measure the distance traveled along the streets. The Euclidean distance may sound familiar to those who know Pythagoras's theorem, which measures the size of the longest side of a right-angled triangle.