WORKSHEET-1

***Q.1: ans**. – a.*

***Q.2: ans**. – a.*

***Q.3: ans**. – d.*

***Q.4: ans**. – a.*

***Q.5: ans**. – b.*

***Q.6: ans**. – d.*

***Q.7: ans**. – a.*

***Q.8: ans**. – b.*

***Q.9: ans**. – a.*

***Q.10: ans**. – a.*

***Q.11: ans**. – d.*

***Q.12: ans**. – a.*

***Q.13: ans**. - Cluster analysis is an exploratory analysis that tries to identify structures within the data. Cluster analysis is also called segmentation analysis or taxonomy analysis. More specifically, it tries to identify homogenous groups of cases if the grouping is not previously known. Because it is exploratory, it does not make any distinction between dependent and independent variables. The different cluster analysis methods that SPSS offers can handle binary, nominal, ordinal, and scale (interval or ratio) data.*
Cluster analysis is often used in conjunction with other analyses (such as discriminant analysis). The researcher must be able to interpret the cluster analysis based on their understanding of the data to determine if the results produced by the analysis are actually meaningful.
Typical research questions the cluster analysis answers are as follows:

- ***Medicine** – What are the diagnostic clusters? To answer this question the researcher would devise a diagnostic questionnaire that includes possible symptoms (for example, in psychology, anxiety, depression etc.). The cluster analysis can then identify groups of patients that have similar symptoms.*
- ***Marketing** – What are the customer segments? To answer this question a market researcher may conduct a survey covering needs, attitudes, demographics, and behavior of customers. The researcher then may use cluster analysis to identify homogenous groups of customers that have similar needs and attitudes.*
- ***Education** – What are student groups that need special attention? Researchers may measure psychological, aptitude, and achievement characteristics. A cluster analysis then may identify what homogeneous groups exist among students (for example, high achievers in all subjects, or students that excel in certain subjects but fail in others).*
- ***Biology** – What is the taxonomy of species? Researchers can collect a data set of different plants and note different attributes of their phenotypes. A cluster analysis can group those observations into a series of clusters and help build a taxonomy of groups and subgroups of similar plants.*

WORKSHEET-1

***Q.14: ans**. - A cluster is the collection of data objects which are similar to each other within the same group. The data objects of a cluster are dissimilar to data objects of other groups or clusters.*

Clustering Approaches:

***1. Partitioning approach:** The partitioning approach constructs various partitions and then evaluates them by some criterion, e.g., minimizing the sum of square errors. It adopts exclusive cluster separation(each object belongs to exactly one group) and uses iterative relocation techniques to improve the partitioning by moving objects from one group to another. It uses a greedy approach and approach at a local optimum. It finds clusters with spherical shapes in small to medium size databases.*

Partitioning approach methods:

- *k-means*
- *k-medoids*
- *CLARINS*

***2. Density-based approach:** This approach is based on connectivity and density functions. It divides the set of objects into multiple exclusive clusters or a hierarchy of clusters. Density-based methods:*

- *DBSACN*
- *OPTICS*

***3. Grid-based approach:** This approach quantizes objects into a finite number of cells that form a grid structure. Fast processing time and independent of a number of data objects. Grid-based Clustering method is the efficient approach for spatial data mining problems.*

Grid-based approach methods:

- *STING*
- *WaveCluster*
- *CLIQUE*

***4. Hierarchical approach:** This creates a hierarchical decomposition of the data objects by using some measures. Hierarchical approach methods:*

- *Diana*
- *Agnes*
- *BIRCH*

WORKSHEET-1

- *CAMELEON*

Measures for Quality of Clustering:

If all the data objects in the cluster are highly similar then the cluster has high quality. We can measure the quality of Clustering by using the Dissimilarity/Similarity metric in most situations. But there are some other methods to measure the Qualities of Good Clustering if the clusters are alike.

1. Dissimilarity/Similarity metric: *The similarity between the clusters can be expressed in terms of a distance function, which is represented by d(i, j). Distance functions are different for various data types and data variables. Distance function measure is different for continuous-valued variables, categorical variables, and vector variables. Distance function can be expressed as Euclidean distance, Mahalanobis distance, and Cosine distance for different types of data.*

2. Cluster completeness: *Cluster completeness is the essential parameter for good clustering, if any two data objects are having similar characteristics then they are assigned to the same category of the cluster according to ground truth. Cluster completeness is high if the objects are of the same category.*

Let us consider the clustering C1, which contains the sub-clusters s1 and s2, where the members of the s1 and s2 cluster belong to the same category according to ground truth. Let us consider another clustering C2 which is identical to C1 but now s1 and s2 are merged into one cluster. Then, we define the clustering quality measure, Q, and according to cluster completeness C2, will have more cluster quality compared to the C1 that is, Q(C2, Cg) > Q(C1, Cg).

3. Ragbag: *In some situations, there can be a few categories in which the objects of those categories cannot be merged with other objects. Then the quality of those cluster categories is measured by the Rag Bag method. According to the rag bag method, we should put the heterogeneous object into a rag bag category.*

Let us consider a clustering C1 and a cluster $C \in C1$ so that all objects in C belong to the same category of cluster C1 except the object o according to ground truth. Consider a clustering C2 which is identical to C1 except that o is assigned to a cluster D which holds the objects of different categories. According to the ground truth, this situation is noisy and the quality of clustering is measured using the rag bag criteria. we define the clustering quality measure, Q, and according to rag bag method criteria C2, will have more cluster quality compared to the C1 that is, Q(C2, Cg)>Q(C1, Cg).

4. Small cluster preservation: *If a small category of clustering is further split into small pieces, then those small pieces of cluster become noise to the entire clustering and thus it becomes difficult to identify that small category from the clustering. The small cluster preservation criterion states that are splitting a small category into pieces is not advisable and it further decreases the quality of clusters as the pieces of clusters are distinctive. Suppose clustering C1 has split into three clusters, C11 = {d1, . . . , dn}, C12 = {dn+1}, and C13 = {dn+2}.*

Let clustering C2 also split into three clusters, namely C1 = {d1, . . . , dn−1}, C2 = {dn}, and C3 = {dn+1,dn+2}. As C1 splits the small category of objects and C2 splits the big category which is preferred

according to the rule mentioned above the clustering quality measure Q should give a higher score to C2, that is, Q(C2, Cg) > Q(C1, Cg)

***Q.15: ans**. Cluster analysis foundations rely on one of the most fundamental, simple and very often unnoticed ways (or methods) of understanding and learning, which is grouping "objects" into "similar" groups. This process includes a number of different algorithms and methods to make clusters of a similar kind. It is also a part of data management in statistical analysis.*

When we try to group a set of objects that have similar kind of characteristics, attributes these groups are called clusters. The process is called clustering. It is a very difficult task to get to know the properties of every individual object instead, it would be easy to group those similar objects and have a common structure of properties that the group follows. Cluster analysis is a multivariate data mining technique whose goal is to groups objects (eg., products, respondents, or other entities) based on a set of user selected characteristics or attributes. It is the basic and most important step of data mining and a common technique for statistical data analysis, and it is used in many fields such as data compression, machine learning, pattern recognition, information retrieval etc.

Types of Cluster Analysis

The clustering algorithm needs to be chosen experimentally unless there is a mathematical reason to choose one cluster method over another.It should be noted that an algorithm that works on a particular set of data will not work on another set of data. There are a number of different methods to perform cluster analysis. Some of them are,

Hierarchical Cluster Analysis

*In this method, first, a cluster is made and then added to another cluster (the most similar and closest one) to form one single cluster. This process is repeated until all subjects are in one cluster. This particular method is known as **Agglomerative method**. Agglomerative clustering starts with single objects and starts grouping them into clusters.*

***The divisive method** is another kind of Hierarchical method in which clustering starts with the complete data set and then starts dividing into partitions.*

Centroid-based Clustering

In this type of clustering, clusters are represented by a central entity, which may or may not be a part of the given data set. K-Means method of clustering is used in this method, where k are the cluster centers and objects are assigned to the nearest cluster centres

Distribution-based Clustering

WORKSHEET-1

It is a type of clustering model closely related to statistics based on the modals of distribution. Objects that belong to the same distribution are put into a single cluster.This type of clustering can capture some complex properties of objects like correlation and dependence between attributes.

Density-based Clustering

In this type of clustering, clusters are defined by the areas of density that are higher than the remaining of the data set. Objects in sparse areas are usually required to separate clusters.The objects in these sparse points are usually noise and border points in the graph.The most popular method in this type of clustering is DBSCAN.

Applications and Examples

It is the principal job of exploratory data mining, and a common method for statistical data analysis. It is used in many fields, such as machine learning, image analysis, pattern recognition, information retrieval, data compression, bioinformatics and computer graphics.

It can be used to examine patterns of antibiotic resistance, to incorporate antimicrobial compounds according to their mechanism of activity, to analyse antibiotics according to their antibacterial action.

Cluster analysis can be a compelling data-mining means for any organization that wants to recognise discrete groups of customers, sales transactions, or other kinds of behaviours and things. For example, insurance providing companies use cluster analysis to identify fraudulent claims and banks apply it for credit scoring