CMR INSTITUTE OF TECHNOLOGY	USN	

Internal Assesment Test – III, May 2019

Sub:	DATA WAREHOUSING AND DATA MINING							Code:	17MCA442
Date:	14-05-2019	Duration:	90 mins	Max Marks:	50	Sem:	IV	Branch:	MCA

	Answer **ONE FULL QUESTION** from each part	Marks	OBE CO	OBE RBT
	Part – I			
1	List some applications of clustering	10	CO7	L2
	(OR)			
2	Explain the requirements of clustering.	10	CO7	L2
	Part – II			
3	Explain how to find the dissimilarity for interval-scaled variables and ratio-scaled variables.	10	CO7	L2
	(OR)			
4	Explain how to find the dissimilarity for binary variables and variables of mixed type.	10	CO7	L2
	Part – III			
5	Explain about the k-means and k-medoids clustering. Illustrate the strength and weakness of k-means in comparison with the k-medoids algorithm.	10	CO7	L2
	(OR)			
6	Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8): (a) Compute the Euclidean distance between the two objects. (b) Compute the Manhattan distance between the two objects.	10	CO7	L3
	PART – IV			
7	Explain the BIRCH and ROCK clustering.	10	CO7	L2
	(OR)			
8	Explain the DBSCAN and OPTICS clustering.	10	CO7	L2
	Part – V			
9	Explain the measures for quality and validity cluster analysis.	10	CO7	L2
	(OR)			
10	Data cubes and multidimensional databases contain categorical, ordinal, and numerical data in hierarchical or aggregate forms. Based on what you have learned about the clustering methods, which clustering method would you choose that finds clusters in large data cubes effectively and efficiently. Justify your answer.	10	CO7	L4

CMR Institute of Technology, Bengalore – 560 037

Department of Computer Applications

Scheme and Solutions for Internal Assessment Test – III

17MCA442 – DATA WAREHOUSING AND DATA MINING **Semester / Section: IV** **Date of Test: 14-05-2019**

---

1. List some applications of clustering. **(10)**

- Clustering is the process of grouping the data into classes or clusters, so that:
  - objects within a cluster (intra-cluster) have high similarity in comparison to one another
  - very dissimilar to objects in other clusters (inter-cluster).
- Applications:
- Biology: taxonomy of living things: kingdom, phylum, class, order, family, genus and species
- Information retrieval: document clustering
- Land use: Identification of areas of similar land use
- Marketing: discover distinct groups from customer bases, and develop targeted marketing programs
- City-planning: Identifying groups of houses according to their house type, value, and geographical location
- Earth-quake studies: Observed earth quake epicenters should be clustered along continent faults
- Climate: understanding earth climate, find patterns of atmospheric and ocean
- Economic Science: market research

Scheme:

Definition of clustering: 2 Marks

Applications: 8 Marks

2. Explain the requirements of clustering. **(10)**

- typical requirements (Features) of clustering in data mining:

i. Scalability: Highly scalable clustering algorithms are needed to handle millions of objects

ii. Ability to deal with different types of attributes: to handle different types of data (binary, categorical, mixtures of these)

iii. Discovery of clusters with arbitrary shape:

- It is important to develop algorithms that can detect clusters of arbitrary shape (usually spherical shapes based on Euclidean or Manhattan distance measures)

iv. Minimal requirements for domain knowledge to determine input parameters: The clustering results can be quite sensitive to input parameters and are often difficult to determine

v. Ability to deal with noisy data: Some clustering algorithms are sensitive to data which has missing, unknown, or erroneous data and may lead to clusters of poor quality

vi. Incremental clustering and insensitivity to the order of input records:

- Some clustering algorithms cannot incorporate newly inserted data
- Some clustering algorithms are sensitive to the order of input data

vii. High dimensionality: Many clustering algorithms are good at handling low-dimensional data

- Finding clusters of data objects in high dimensional space is challenging

---

viii. Constraint-based clustering: Real-world applications may need to perform clustering under various kinds of constraints (Example: Choose a new ATM location)
ix. Interpretability and usability: clustering results to be interpretable, comprehensible, and usable

Scheme:
For all the individual requirements: 10 Marks

3. Explain how to find the dissimilarity for interval-scaled variables and ratio-scaled variables. **(10)**

(i) Interval-Scaled Variables: are continuous measurements of a roughly linear scale, like weight and height and weather temperature.

- The measurement unit used can affect the clustering analysis
- To help avoid dependence on the choice of measurement units, the data should be standardized.
- the dissimilarity (or similarity) between the objects described by interval-scaled variables is typically computed based on the distance between each pair of objects
- The most popular distance measure is Euclidean distance, which is defined as

$$d(i,j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \ldots + |x_{ip} - x_{jp}|^2)}$$

where i=(x_{i1}, x_{i2},…, x_{in}) and j =(x_{j1}, x_{j2},…, x_{jn}) are two n-dimensional data objects.

- Another well-known metric is Manhattan (or city block) distance, defined as

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \ldots + |x_{ip} - x_{jp}|$$

- Both the Euclidean distance and Manhattan distance satisfy the following mathematic requirements of a distance function:

d(i,i) = 0	The distance of an object to itself is 0
d(i,j) $\geq$ 0	Distance is a non-negative number
d(i,j) = d(j,i)	Distance is a symmetric function
d(i,j) $\leq$ d(i,k) + d(k,j)	triangular inequality

c. Ratio-Scaled Variables: a positive measurement on a nonlinear scale, approximately at exponential scale approximately following the formula: Ae^{Bt} or Ae^{-Bt} (Example: measure height in centimetres, metres, inches or feet, not possible to have negative ratio)

- three methods to handle ratio-scaled variables for computing the dissimilarity between objects:

i. Treat ratio-scaled variables like interval-scaled variables
ii. Apply logarithmic transformation to a ratio-scaled variable f having value x_{if} for object i by using the formula y_{if} = log(x_{if}). The y_{if} values can be treated as interval valued

iii. Treat x_{if} as continuous ordinal data and treat their ranks as interval-valued

Scheme:

For Interval-scaled variable: 5 Marks

For Ratio-Scaled variable: 5 Marks

4. Explain how to find the dissimilarity for binary variables and variables of mixed type. (10)

(i) Binary Variables

- A binary variable has only two states: 0 or 1
- A binary variable is symmetric if both of its states are equally valuable and carry the same weight (Eg.) Gender
- Dissimilarity that is based on symmetric binary variables is called symmetric binary dissimilarity
- Its dissimilarity (or distance) measure can be used to assess the dissimilarity between objects i and j:

$$d(i,j) = \frac{r+s}{q+r+s+t}$$

- The dissimilarity based on such variables is called asymmetric binary dissimilarity where the number of negative matches, t, is considered unimportant and thus is ignored in the computation
- we can measure the distance between two binary variables based on the notion of similarity instead of dissimilarity, by using Jaccard coefficient

$$sim\ (i,j) = \frac{q}{q+r+s}$$

(ii) Variables of Mixed Types

- A database can contain all of the variables of mixed types
- compute the dissimilarity between objects of mixed variable types:
  - group each kind of variable together, performing a separate cluster analysis for each variable type.
  - process all variable types together, performing a single cluster analysis.
  - combines the different variables into a single dissimilarity matrix, bringing all of the meaningful variables onto a common scale of the interval [0.0,1.0]
- The dissimilarity d(i, j) between objects i and j is defined as

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

- The contribution of variable f to the dissimilarity between i and j ($\delta_{ij}^{(f)}$) is computed dependent on its type:
  - f is binary or nominal: $\delta_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$, or $\delta_{ij}^{(f)} = 1$ otherwise

- f is interval-based: use the normalized distance
- f is ordinal or ratio-scaled: use the distance measure
- compute ranks r_{if} and treat z_{if} as interval-scaled

Scheme:

For binary variables: 5 Marks

For mixed type variable: 5 Marks

5. Explain about the k-means and k-medoids clustering. Illustrate the strength and weakness of k-means in comparison with the k-medoids algorithm. (10)

(i) k-Means Clustering:

- takes the input parameter, k, and partitions a set of n objects into k clusters
- Cluster similarity is measured in regard to the mean value of the objects in a cluster, which can be viewed as the cluster's centroid or center of gravity.
- Given k, the k-means algorithm is implemented in four steps:

i. Partition objects into k nonempty subsets
ii. Compute seed points as the centroids of the clusters of the current partition (the centroid is the center - mean point - of the cluster)
iii. Assign each object to the cluster with the nearest seed point
iv. Go back to Step 2, stop when no more new assignment

- Strength: Relatively efficient *O*(tkn), where t is no. of iterations, k is no. of clusters and n is no. of objects.
- Weakness:
  - Applicable only when mean is defined (then what about categorical data?)
  - Need to specify k (the number of clusters, in advance)
  - Unable to handle noisy data and outliers
  - Not suitable to discover clusters with non-convex shapes

(ii) K-Medoids Method (Representative Object-Based Technique):

- The k-means algorithm is sensitive to outliers because an object with an extremely large value may substantially distort the distribution of data
- Instead of taking the mean value of the object in a cluster as a reference point, medoids can be used, which is the most centrally located object in a cluster.

Scheme:

For K-Means clustering: 6 Marks

For k-Medoids clustering: 4 Marks

6. Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8): (10)
(a) Compute the Euclidean distance between the two objects.
(b) Compute the Manhattan distance between the two objects.

(a) Euclidean distance:

$$d(i,j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \ldots + |x_{ip} - x_{jp}|^2)}$$

$$d(i,j) = \sqrt{|(20-22)|^2 + |(0-1)|^2 + |(36-42)|^2 + |(8-10)|^2}$$

$$= \sqrt{(2)^2 + (1)^2 + (6)^2 + (2)^2} = \sqrt{4+1+36+4}$$

$$= \sqrt{45} = 6.71$$

(b) Manhattan Distance:

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \ldots + |x_{ip} - x_{jp}|$$

$$d(i,j) = |(20-22)| + |(0-1)| + |(36-42)| + |(8-10)|$$

$$= 2 + 1 + 6 + 2 = 11$$

Scheme:

Each distance measure carries 5 Marks (2X5 = 10 Marks)

7. Explain the BIRCH and ROCK clustering. (10)

(i) BIRCH: Balanced Iterative Reducing and Clustering Using Hierarchies

- Incrementally construct a CF (Clustering Feature) tree, a hierarchical data structure for multiphase clustering
- A clustering feature (CF) is a three-dimensional vector summarizing information about clusters of objects.
- Given n d-dimensional objects or points in a cluster, {xi}, then the CF of the cluster is defined as: CF = <n, LS, SS> where
  - n is the number of points in the cluster
  - LS is the linear sum of the n points
  - SS is the square sum of the data points
- Two phases:
  - Phase-1: scan DB to build an initial in-memory CF tree (a multi-level compression of the data that tries to preserve the inherent clustering structure of the data)
  - Phase-2: use an arbitrary clustering algorithm to cluster the leaf nodes of the CF-tree
  - Advantage: Scales linearly - finds a good clustering with a single scan and improves the quality with a few additional scans
  - Weakness: handles only numeric data, and sensitive to the order of the data record.

(ii) ROCK (RObust Clustering using linKs)

- A Hierarchical Clustering Algorithm for Categorical Attributes

- explores the concept of links (the number of common neighbors between two objects) for data with categorical attributes.
- Distance measures cannot lead to high-quality clusters when clustering categorical data
- ROCK takes a more global approach to clustering by considering the neighborhoods of individual pairs of points
  - If two similar points also have similar neighborhoods, then the two points likely belong to the same cluster and so can be merged
- two points, p_i and p_j, are neighbors if $sim(p_i, p_j) \geq \theta$, where
  - sim is a similarity function and
  - θ is a user-specified threshold
- If the number of links between two points is large, then it is more likely that they belong to the same cluster
- ROCK is more robust than standard clustering methods that focus only on point similarity
- Computational complexity: $O(n^2 + n\, m_m\, m_a + n^2 \log n)$ where m_m and m_a are the maximum and average number of neighbors, respectively, and n is the number of objects

Scheme:

For BIRCH clustering: 5 Marks

For ROCK clustering: 5 Marks

8. Explain the DBSCAN and OPTICS clustering. (10)

(i) DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

- The algorithm grows regions with sufficiently high density into clusters and discovers clusters of arbitrary shape in spatial databases with noise.
- Two parameters:
  - ε: Maximum radius of the neighbourhood
  - MinPts: Minimum number of points in an ε-neighbourhood of that point
- $N_\varepsilon(p)$: {q belongs to D | dist(p,q) <= ε}
- Directly density-reachable: A point p is directly density-reachable from a point q with respect to: ε, MinPts if
  - p belongs to $N_\varepsilon(q)$
  - core point condition: $|N_\varepsilon(q)| >= MinPt$
  - Density-reachable: A point p is density-reachable from a point q w.r.t ε, MinPts if there is a chain of points $p_1, \ldots, p_n$, $p_1 = q$, $p_n = p$ such that p_{i+1} is directly density-reachable from p_i

(ii) OPTICS: Ordering Points to Identify the Clustering Structure

- Disadvantage of DBSCAN: selecting parameter values, ε and MinPts, that will lead to the discovery of acceptable clusters
- OPTICS computes an augmented cluster ordering for automatic and interactive cluster analysis
- Produces a special order of the database with respect to its density-based clustering structure

- This cluster-ordering contains information equivalent to the density-based clustering corresponding to a broad range of parameter settings
- Good for both automatic and interactive cluster analysis, including finding intrinsic clustering structure
- Can be represented graphically or using visualization techniques
- two values need to be stored for each object:
  - The core-distance of an object p is the smallest ε′ value that makes {p} a core object. If p is not a core object, the core-distance of p is undefined
  - The reachability-distance of an object q with respect to another object p is the greater value of the core-distance of p and the Euclidean distance between p and q. If p is not a core object, the reachability-distance between p and q is undefined

Scheme:

For DBSCAN clustering: 5 Marks

For OPTICS clustering: 5 Marks

9. Explain the measures for quality and validity cluster analysis. (10)

- few methods to choose from for measuring the quality of a clustering.
- these methods can be categorized into two groups according to whether ground truth is available.
- Here, ground truth is the ideal clustering that is often built using human experts.

i. If ground truth is available, it can be used by extrinsic methods (also called supervised methods), which compare a clustering against the ground truth using certain clustering quality measure.

ii. If the ground truth is unavailable, we can use intrinsic methods (also called unsupervised methods), which evaluate the goodness of a clustering by considering how well the clusters are separated.

(i) Extrinsic Methods

- a measure $Q(C, C_g)$ on clustering quality is effective if it satisfies the following four essential criteria:

i. Cluster homogeneity: the purer, the better

ii. Cluster completeness: should assign objects belong to the same category in the ground truth to the same cluster

iii. Rag bag: putting a heterogeneous object into a pure cluster should be penalized more than putting it into a rag bag (i.e., “miscellaneous” or “other” category)

iv. Small cluster preservation: splitting a small category into pieces is more harmful than splitting a large category into pieces

- BCubed precision and recall metrics, which satisfy all four criteria
- The precision of an object indicates how many other objects in the same cluster belong to the same category as the object.
- The recall of an object reflects how many objects of the same category are assigned to the same cluster.

(ii) Intrinsic Methods

- intrinsic methods evaluate a clustering by examining how well the clusters are separated and how compact the clusters are
- For a data set, D, of n objects, suppose D is partitioned into k clusters, $C_1,\ldots,C_k$.
- For each object $o \in D$, we calculate a(o) as the average distance between o and all other objects in the cluster to which o belongs.
- Similarly, b(o) is the minimum average distance from o to all clusters to which o does not belong.
- The silhouette coefficient of o is then defined as:

$$s(\boldsymbol{o}) = \frac{b(\boldsymbol{o}) - a(\boldsymbol{o})}{\max\{a(\boldsymbol{o}), b(\boldsymbol{o})\}}$$

- *a*(o) as the average distance between ***o*** and all other objects in the cluster to which o belongs
- b(o) is the minimum average distance from o to all clusters to which o does not belong.

Scheme:

Each method carries equal marks (2 x 5 = 10 Marks)

10. Data cubes and multidimensional databases contain categorical, ordinal, and numerical data in hierarchical or aggregate forms. Based on what you have learned about the clustering methods, which clustering method would you choose that finds clusters in large data cubes effectively and efficiently. Justify your answer. (10)

- We first need to pre-process and discretize existing data (such as ordinal and numerical data) to obtain a single dimensional discretization. We then can perform the multidimensional clustering in two steps:

i. The first step involves partitioning of the n-dimensional data space into non-overlapping rectangular units, identifying the dense units among them. This is done in 1-D for each dimension. We then can generate candidate dense units in k-dimensional space from the dense units found in (k ¡ 1)-dimensional space.

ii. In the second step, a minimal description for each cluster is generated. For each cluster, this determines the maximal region that covers the cluster of connected dense units. It then determines a minimal cover for each cluster.

- Using such a method, we can effectively find clusters from the data that are represented as a data cube.

Scheme: 10 Marks can be awarded for the above answer or any alternate answer with reference to the context and justification.

Scheme Of Evaluation
Internal Assessment Test 3 – May 2019

Sub:	Data Warehousing and Data Mining							**Code:**	17MCA442
Date:	14-05-2019	Duration:	90mins	Max Marks:	50	**Sem:**	IV	**Branch:**	MCA

Question #	Description	Marks Distribution	Max Marks
1.	**List some applications of clustering.**		10
	• Definition of clustering	2	
	• Applications	8	
2.	**Explain the requirements of clustering.**		10
	• description of the requirements	10	
3.	**Explain how to find the dissimilarity for interval-scaled variables and ratio-scaled variables.**		10
	• For Interval-scaled variables	5	
	• For ratio-scaled variables	5	
4.	**Explain how to find the dissimilarity for binary variables and variables of mixed type.**		10
	• For binary variables	5	
	• For variables of mixed type	5	
5.	**Explain about the k-means and k-medoids clustering. Illustrate the strength and weakness of k-means in comparison with the k-medoids algorithm.**		10
	• For k-means clustering	5	
	• For k-medoids clustering	5	
6.	**6. Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8): (a) Compute the Euclidean distance between the two objects. (b) Compute the Manhattan distance between the two objects.**		10
	• For Euclidean distance	5	
	• For Manhattan distance	5	
7.	**Explain the BIRCH and ROCK clustering.**		10
	• For BIRCH clustering	5	
	• For ROCK clustering	5	

8.	**Explain the DBSCAN and OPTICS clustering.**		**10**
	• For DBSCAN clustering	5	
	• For OPTICS clustering	5	
9.	**Explain the measures for quality and validity cluster analysis.**		**10**
	• For intrinsic method	5	
	• For extrinsic method	5	
10.	**Data cubes and multidimensional databases contain categorical, ordinal, and numerical data in hierarchical or aggregate forms. Based on what you have learned about the clustering methods, which clustering method would you choose that finds clusters in large data cubes effectively and efficiently. Justify your answer.**		**10**
	• Answer / justification	10	