GANPAT UNIVERSITY
B. TECH SEMESTER - VII (IT) EXAMINATION
REGULAR EXAMINATION NOV/DEC-2012
IT 701: DATA MINING & DATA WAREHOUSING

Time: 3 Hours] **[Total Marks: 70**

Instructions:

1. **Attempt all questions.**
2. **Figures to the right indicate full marks**
3. **Each section should be written in a separate answer book**

Section 1

Q.1

(A) Given a set of 5-dimensional categorical samples: **(8)**
A = (1, 0, 1, 1, 0), B = (1, 1, 0, 1, 0), C = (0, 0, 1, 1, 0),
D = (0, 1, 0, 1, 0), E = (1, 0, 1, 0, 1), F = (0, 1, 1, 0, 0).
Apply agglomerative hierarchical clustering using:
a) Single-link similarity measure based on Rao's coefficient.
b) Complete-link similarity measure based on simple matching coefficient SMC.
c) Plot the dendograms for the solutions to part a) and b).

(B) Given 1-dimensional data set X={-5,0,23.0,17.6,9.23,1.11} normalize the data set using **(4)**
i)Min-Max Normalization[0,1] ii) Min-Max Normalization[-1,1]

OR

Q.1

(A) A database has five transactions. Let min sup = 60% and min conf = 80%. **(8)**
Find all frequent itemsets using Hash and FP-growth, respectively. Compare the efficiency of the two mining processes.

TID	items bought
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y }
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I ,E}

(B) A Database has four transactions. Let minimum support and confidence be 50%. **(4)**

Tid	Items bought
1	A,B,D
2	A,D
3	A,C
4	B,D,E,F

Find out the frequent item sets and strong association rules for the above table using Apriori Algorithm

Q.2

(A) What is Core Object? Explain A Density-Based Clustering Method Based on Connected Regions with Sufficiently High Density. **(6)**

(B) Explain Jaccard coefficient with the example. **(5)**

OR

Q.2 Suppose that the data mining task is to cluster the following eight points (with (x, y) representing location) into three clusters: **(6)**

(A) A1(2, 10), A2(2, 5), A3(8, 4), B1(5, 8), B2(7, 5), B3(6, 4), C1(1, 2), C2(4, 9):
The distance function is Euclidean distance. Suppose initially we assign A1, B1, and C1
as the center of each cluster, respectively. Use the k-means algorithm to show only
(a) The three cluster centers after the first round execution
(b) The final three clusters
(c) Illustrate the strength and weakness of k-means algorithm in comparison with hierarchical clustering schemes (such as AGNES).

(B) Explain Data Reduction techniques. **(5)**

Q.3

(A) Here 2 *2 contingency table with summarizing the transactions with respect to game and video purchases. Find out χ^2, Cosine, All-Conf, Lift.

	game	*game*	*Srow*
video	4,000	3,500	7,500
video	2,000	500	2,500
Scol	6,000	4,000	10,000

(B)

Attributes: RID, age, income, student, credit rating, Class: buys computer
1, youth, high, no, fair, no
2, youth, high, no, xcellent, no
3 ,middle aged, high, no, fair ,yes
4 ,senior ,medium, no, fair, yes
5 ,senior ,low, yes, fair, yes
6 ,senior, low, yes, excellent ,no
7, middle aged, low, yes, excellent, yes
8, youth, medium, no, fair, no
9, youth ,low, yes ,fair, yes
10, senior, medium, yes ,fair ,yes
11 ,youth, medium, yes, excellent, yes
12 ,middle aged ,medium no excellent, yes
13, middle aged ,high, yes ,fair ,yes
14, senior, medium ,no, excellent, no

(8)

Predicting a class label using naïve Bayesian classification.
Where X = (*age = youth, income = medium, student = yes, credit rating = fair*)

Section 2

Q.4

(A) Explain the Balanced Iterative Reducing and Clustering Using Hierarchies. Show how effective is BIRCH? Where, C1= (2, 5), (3, 2), (4, 3) and C2= (5, 2), (2, 3), (3, 4). Show CF1, CF2 and CF3. [8]

(B) What Is a Data Warehouse? Explain Subject-oriented, Integrated, Time-variant, Nonvolatile Data Warehouse. [4]

OR

Q.4

(A) Explain the architecture of Data Warehouse. [4]

(B) Explain DIM, Fact and Fact- less Fact Table. [4]

(C) Explain Stars, Snowflakes, and Fact Constellations Schemas for Multidimensional Databases [4]

Q.5

(A) What are the major issues in Data Mining? [6]

(B) Explain the different operations (or techniques) of OLAP with suitable example. [5]

OR

Q.5

(A) Explain smoothing, Aggregation, Generalization and Normalization with suitable Example.
Explain Interval-Scaled Variables, Categorical, Ordinal, and Ratio-Scaled Variables with suitable example. [6]

(B)

A sample data table containing variables of mixed type

object identifier	test-1 (categorical)	test-2 (ordinal)	test-3 (ratio-scaled)
1	code-A	excellent	445
2	code-B	fair	22
3	code-C	good	164
4	code-A	excellent	1,210

[5]

Q.6

(A) Explain Chameleon (A Hierarchical Clustering Algorithm Using Dynamic Modeling) Algorithm using Example. [4]

(B) Suppose that the data mining task is to cluster the following eight points (with (x, y) representing location) into three clusters: A1(2, 10), A2(2, 5), A3(8, 4), B1(5, 8), B2(7, 5), B3(6, 4), C1(1, 2), C2(4, 9):The distance function is Euclidean distance. Explain the k-medoids Algorithm. Suppose initially we assign A1, B1, and C1 as the center of each cluster, respectively. Use the k-medoids(PAM, a k-medoids algorithm for partitioning based on medoid or central objects) algorithm to show only (a) The three cluster centers after the first round execution. [8]

--------End of a Paper-----