Ref. No.: EX/IT/T/412A/2018

B.E. INFORMATION TECHNOLOGY 4TH YEAR 1ST SEMESTER EXAMINATION-2018

Subject: Data Mining

Time: 3 Hours **Full Marks: 100**

Note: Attempt *Q.1* and any *five* from the rest.

Q.1 Answer any *ten* questions.

a) What are the steps of different approaches of clustering high-dimensional data?
b) Explain: supervised vs. unsupervised learning and Euclidian distance vs. Manhattan distance.
c) Write the definitions with suitable example(s) of normal/suspected outliers and dendogram.
d) What is k-nearest-neighbor graph approach? How can it be used in data mining?
e) Describe the role of MDL in subspace clustering.
f) Write the definitions with suitable example(s) of closed frequent itemset and maximal frequent itemset.
g) Explain different types of data visualization techniques.
h) Define with suitable example(s): nominal attributes and ordinal attributes.
i) What are the time-complexities of k-means, k-medoids, CLARA and CALARANS algorithms?
j) Explain the terms dense-unit and minimum-cover-with-maximal-region with respect to subspace.
k) What are various valued-attributes (with suitable examples) used as splitting attribute in decision tree?
l) Explain with suitable example(s): normalized schema and galaxy schema as data warehouse model.

2x10

Q.2

a) The following table consists of eight transactions with min_sup = 30% and min_conf = 60%.

TID	Items
1	E, A, D, B
2	D, A, C, E, B
3	C, A, B, E
4	B, A, D
5	D
6	D, B
7	A, D, E
8	B, C

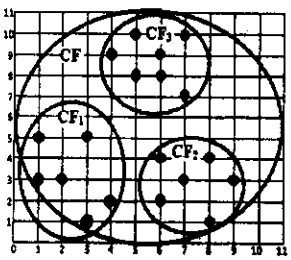
i) Find all frequent itemsets using *Apriori* and *FP-growth*, respectively. Compare the efficiency of the two mining processes.
ii) List all the strong association rules (with support s =30% and confidence c=60%) matching the following *metarule*, where X is a variable representing customers, matching denotes variables representing items (e.g., A, B etc.):

$$\forall x \in transaction,\ buys(X, item_1) \wedge buys(X, item_2) \Rightarrow buys(X, item_3) \quad [s, c]$$

b) Explain the drawbacks (with suitable examples) of:
i) divisive method, ii) k-means algorithm, and iii) k-medoids algorithm.
c) Explain with suitable example(s): traditional clustering algorithms are inefficient on high-dimensional data.

(3x2+1+2)+(1+2x2) +2

Q.3

a) 

The following figure represents three disjoint clusters and their merger C_1, C_2, C_3 and C respectively in 2-D co-ordinate systems.

i) Find out the values of CF_1, CF_2, CF_3 and CF based on 2-D systems of C_1, C_2, C_3 and C respectively where CF indicates clustering feature of a cluster.
ii) Construct the CF-tree using C_1, C_2, C_3 and C.

b) Suppose that a data warehouse consists of the three dimensions *time*, *doctor*, and *patient*, and the two measures *count* and *charge*, where *charge* is the fee that a doctor charges a patient for a visit.
i) Draw the schema diagrams for the above data warehouse using model of *fact constellation.*
ii) Starting with the base cuboid [*day, doctor, patient*], what specific *OLAP operations* should be performed in order to list the total fee collected by each doctor in 2012?
iii) To obtain the same list, write an SQL query assuming the data are stored in a relational database with the schema *fee (day, month, year, doctor, hospital, patient, count, charge).*
c) Find out the outliers w.r.t. the data set, B={2.37, 2.16, 14.82, 1.73, 41.04, 0.23, 1.32, 2.91, 39.41, 0.11, 27.44, 4.51, 60.05, 0.51, 4.50, 0.18, 14.68, 4.66, 1.30, 2.06, 49.09, 1.19} using boxplots.

(3+2)+(4+2+2)+ 3

Q.4

a) The following table consists of twelve objects (i.e., points) in 2-D systems.

Objects	x	y
X1	2	3
X2	10	9
X3	6	7
X4	12	10
X5	4	5
X6	11	10
X7	9	8
X8	7	11
X9	9	9
X10	3	4
X11	10	12
X12	8	10

Find best possible *two* clusters formed by the objects using the algorithms:
i) k-means and
ii) k-medoids

b) Write the definitions with appropriate expressions of:
i) relative interconnectivity and ii) relative closeness.

c) What are the challenges and their solutions of clustering high-dimensional data?

d) Describe various methods/tests/techniques with suitable example(s) used in four levels of data preprocessing.

(4+4)+3+3+2

Q.5

a) The following table consists of training data from an employee database.

department	age(yrs.)	salary	status
sales	31...35	46K...50K	senior
sales	26...30	26K...30K	junior
sales	31...35	31K...35K	junior
systems	21...25	46K...50K	junior
systems	31...35	66K...70K	senior
systems	26...30	46K...50K	junior
systems	41...45	66K...70K	senior
marketing	36...40	46K...50K	senior
marketing	31...35	41K...45K	junior
marketing	26...30	46K...50K	junior
marketing	41...45	66K...70K	senior
secretary	21...25	46K...50K	junior
secretary	46...50	36K...40K	senior
secretary	26...30	26K...30K	junior

The data should be generalized (discretized) as "age: 20-30=low_aged; 31-40=middle_aged; 41-60=high_aged" and "salary: 20K-40K=low; 41K-60K=medium; 61K-80K=high".

Assuming *status* as class label attribute, select appropriate splitting attribute and draw the decision tree accordingly using:
i) Information gain and
ii) Gain ratio

b) Write the definitions of coverage and accuracy. Find their values using the above database, (in Q.a) and satisfying the rule R, where R: (department=sales ∨ department = systems)∧(age=35...45)∧(salary=35K...50K)=>(status=junior) .

c) How is antimonotonicity-property used in apiori-like algorithms?

(5+5)+(2+2)+2

Q.6

a) The following contingency table summarizes supermarket transaction data. Suppose that the association rule *HomePC=>Laptop* is mined with minimum support threshold of 40% and a minimum confidence threshold of 66%.

	$\overline{\text{HomePC}}$	HomePC
$\overline{\text{Laptop}}$	1000	2500
Laptop	4000	4500

i) Check whether the association rule is misleading strong or not?
ii) If it is misleading then what measure(s) will you take and how *(in different ways)* to filter the misleading strong association rule?

b) Using the employee database given in Q.5(a) and assuming *status* as class label attribute, find out Gini Indexes for measuring attribute selection.

c) What is/are the problem(s) and solution(s) of overfitting the data in making decision tree? Explain with suitable examples.

d) Describe the following with suitable examples: working principles of *FP-growth* from *FP-tree*.

(2+2+3)+4+3+2

Q.7 Write short notes on any *four*.

a) Single-linkage vs. complete-linkage algorithm (with suitable examples).
b) A density, grid, and subspace based clustering algorithm.
c) Data mining for financial data analysis, science and technology and intrusion detection and prevention.
d) Mining sequence pattern of time series and biological data.
e) Hierarchical clustering algorithm: BIRCH vs. CHAMELEON.

4x4