No. of Printed Pages : **3**

Sl. No.

C5-R4 : DATA WAREHOUSING AND DATA MINING

NOTE :

1. Answer question 1 and any FOUR from questions 2 to 7.

2. Parts of the same question should be answered together and in the same sequence.

Time : 3 Hours **Total Marks : 100**

1. (a) Describe the types of situations that produce sparse or dense data cubes.

(b) Define Factless fact table with an example.

(c) Why is tree pruning useful in decision tree induction ? What is a drawback of using a separate set of tuples to evaluate pruning ?

(d) Would the cosine measure be the appropriate similarity measure to use with K-means clustering for time series data ? Why or why not ? If not, what similarity measure would be more appropriate ?

(e) What is the value ranges for the following normalization methods ?

(i) min-max normalization

(ii) z-score normalization

(iii) Normalization by decimal scaling

(f) Why concept hierarchies are useful in data mining ?

(g) Compare the advantages and disadvantages of eager classification (e.g., decision tree, Bayesian, neural network) versus lazy classification (e.g., k-nearest neighbor, case-based reasoning). **(7x4)**

2. (a) In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem.

(b) Describe the differences between the following approaches for the integration of a data mining system with a database or data warehouse system: no coupling, loose coupling, semi-tight coupling, and tight coupling. State which approach you think is the most popular, and why ? **(9+9)**

3. (a) Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8) :

(i) Compute the Euclidean distance between the two objects.

(ii) Compute the Manhattan distance between the two objects.

(iii) Compute the Minkowski distance between the two objects, using p=3.

(b) Why is it that BIRCH encounters difficulties in finding clusters of arbitrary shape but OPTICS does not ? Can you propose some modifications to BIRCH to help it find clusters of arbitrary shape ? **(9+9)**

4. A database has four transactions. Let min sup = 60% and min conf =80%.

cust_ID	*TID*	*items_bought (in the form of brand-item_category)*
01	T100	{King's-Crab, Sunset-Milk, Dairyland-Cheese, Best-Bread}
02	T200	{Best-Cheese, Dairyland-Milk, Goldenfarm-Apple, Tasty-Pie, Wonder-Bread}
01	T300	{Westcoast-Apple, Dairyland-Milk, Wonder-Bread, Tasty-Pie}
03	T400	{Wonder-Bread, Sunset-Milk, Dairyland-Cheese}

(i) At the granularity of item category (e.g., $item_i$ could be "Milk"), for the following rule template,

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

List the frequent k-itemset for the largest k, and all of the strong association rules (with their support s and confidence c) containing the frequent k-itemset for the largest k.

(ii) At the granularity of brand-item category (e.g., $item_i$ could be \Sunset-Milk"), for the following rule template,

$\forall X \in customer,\ buys(X,\ item_1) \wedge buys(X,\ item_2) \Rightarrow buys(X,\ item_3)$

List the frequent k-itemset for the largest k (but do not print any rules). **(9+9)**

5. 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) Enumerate three classes of schemas that are popularly used for modelling data warehouses.

(ii) Draw a schema diagram for the above data warehouse using one of the schema classes listed in 5(i).

(iii) 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 2021 ?

(iv) To obtain the same list, write a SQL query assuming the data is stored in a relational database with the schema fee (day, month, year, doctor, hospital, patient, count, charge).

(4.5+4.5+4.5+4.5)

6. Suppose that a restaurant chain would like to mine customers' consumption behavior relating to major sport events, such as "Every time there is a major sport event on TV, the sales of Kentucky Fried Chicken will go up 20% one hour before the match."

(i) For this problem, there are multiple sequences (each corresponding to one restaurant in the chain). However, each sequence is long and contains multiple occurrences of a (sequential) pattern. Thus this problem is different from the setting of sequential pattern mining problem. Analyze what are the differences in the two problem definitions and how such differences may influence the development of mining algorithms.

(ii) Develop a method for finding such patterns efficiently. **(9+9)**

7. Explain following concepts :

(i) Web Data Mining

(ii) CLARA and CLARANS

(iii) Fuzzy set theory **(6+6+6)**

- o O o -