M.E. COMPUTER SCIENCE AND ENGINEERING

FIRST YEAR SECOND SEMESTER

EXAMINATION 2018

Data Warehousing and Data Mining

Time : 3 hours

Full Marks : 100

Answer question no.1 and any four from the rest

All questions carry equal marks

1. Answer true or false stating reasons 2 x10 = 20

(a) Soft computing techniques are mainly used in a data mining system for data preprocessing.
(b) A heterogeneous database consists of a group of legacy databases that combines different kind of data systems.
(c) We can use classification techniques to subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached.
(d) A data warehouse is a department subset of a data mart.
(e) 'Summarization' provides the general characteristics of a target class of data.
(f) Predicting sales amounts of a new product based on advertising expenditure is a data mining problem.
(g) Robot path-finding problem can be solved by any suitable classification technique.
(h) Text data mining methods using keyword do not always provide satisfactory results.
(i) Pattern evaluation is a process that identifies only the interesting patterns.
(j) Data warehouses are designed for real time business operations.

2. (a) Describe with a neat diagram the architecture and operation of a typical data mining system. 12

(b) Describe how data mining can be useful in market analysis and management. 8

3. (a) Describe briefly with examples the difference between Predictive and Descriptive data mining. 10

(b) What is a data warehouse? Describe with an example how data are organized in a typical data warehouse. 2 + 8

4. (a) What is text data mining? Suppose you need to classify a given set of news articles on various faculties of games and sports where the name of each such faculty is the respective class title. Assume that the number of classes (from which the given set of news articles is derived) is not known *a priori*. Describe the following necessary steps to solve the problem. Use any suitable artificial neural network architecture as a soft computing tool for classification.

(i) Feature selection
(ii) Feature reduction
(iii) Analytical formulation of the problem
(iv) Algorithm for classification

3 + 3 + 3 + 3 + 8

5. Suppose that a movie production house wants to assess the possible profit of one of their movies which would be launched in near future. Assume that the said production house is having a database containing the following information regarding each of several hundred movies of the past.

(i) Rating of basic film genres such as action, adventure, animation, biography, comedy etc.
(ii) Rating of the Production House
(iii) Rating of popularity of leading actor/actress at the time of casting
(iv) Total profit/loss

Note that a movie has a value of rating for each genre where the said values would depend on the type of the movie. For instance, an action movie would have a very a high value of rating for 'action' genre, but would have a non-zero positive value for others geners also such as say 'comedy' genre depending upon the extent of comedy present in the movie.

(a) Formulate the problem as a data mining problem
(b) Describe the necessary steps involved in solving the problem
(c) Discuss the merits and demerits of the method you have suggested

4 + 12 + 4

6. (a) Explain briefly the significance of optimization techniques in solving problems of data mining. 4

(b) Consider a robot path-planning problem where each path is associated with a start node (S), target node (T) and an environment through which the robot can traverse. The environment is represented by a collection of orderly numbered grids as shown in the following figure. Shaded grids represent obstacles and the robot can move freely through blank grids. The objective is to find out the shortest distance collision-free path for any given start node and goal node

1 S	2	3	4	5	6
7	8	9	10	11	12
13	14	15	16	17	18
19	20	21	22	23	24
25	26	27	28	29	30
31	32	33	34	35	36 T

Suggest a method using Genetic Algorithms to solve the above-mentioned problem. 12

(c) Discuss the merits and demerits of the method you have suggested 4

7. Write short notes on any two of the following. 2 X 10

(a) OLTP vs OLAP

(b) Object-Relational Databases

(c) Data Cubes