Uva Wellassa University of Sri Lanka
Faculty of Science and Technology
Department of Computer Science and Technology
400 level 1st Semester Examination – May/July 2017
CST452-2 Data Warehousing and Data Mining

Uva Wellassa University

Instructions to candidates
Duration: Two (02) hours
Number of questions: Four (04)
Answer all the questions
Mark allocation: 100

1.

a. A data warehouse can be defined as "subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process". Briefly explain the four (04) characteristics of data stated in the definition.

(4 mark)

b. What are the differences between the three (03) main types of data warehouse usage: information processing, analytical processing and data mining?

(5 mark)

c. A data warehouse can be modeled by either a star schema or a snowflake schema.

i. Briefly describe the similarities and the differences of the two (02) models.

(3 mark)

ii. Suppose that a data warehouse consists of the four (04) dimensions; date, observer, location, and game, and the two (02) measures; count and charge. Draw a star schema diagram for this data warehouse.

(4 mark)

d. Suppose that the data for analysis includes the attribute 'age'. The age values are 35, 35, 36, 40, 13, 15, 16, 16, 19, 46, 52, 70, 20, 20, 21, 22, 33, 35, 35, 45, 22, 25, 25, 25, 25, 30, 33.

i. Find the mean, median and mode of the data.

(3 mark)

ii. Smooth the age values by bin means technique using bin depth of 9. Illustrate your steps.

(4 mark)

iii. Transform the age 35 using min-max normalization where the expected range is 0.0-1.0.

(2 mark)

2.

a. In real-world data, tuples with missing values for some attributes are a common occurrence. Briefly describe four (04) methods for handling this problem.

(4 mark)

b. Discuss two (02) issues need to be considered during data integration.

(4 mark)

24 43

c. Consider the association rule given in the following format.

$URL_1, URL_2 \rightarrow URL_3$ [**s** %, **c** %]

Where URL_1, URL_2 and URL_3 are web page addresses and **s** and **c** are parameters.

i. Briefly describe **s** and **c**.

(2 mark)

ii. "Frequent itemsets are identified before generating association rules". Do you agree with this statement? Justify your answer.

(3 mark)

d. The following data source represents the set of addresses of web pages in a particular website that is being accessed by a set of users. Each row of the data source represents the set of pages retrieved by the users. Assuming that the support and confidence thresholds are 40% and 70% respectively, find all the strong rules using Apriori algorithm.

User ID	Addresses of Web Pages
User1	URL_1, URL_2, URL_5
User2	URL_2, URL_4
User3	URL_2, URL_3
User4	URL_1, URL_2, URL_4
User5	URL_1, URL_3
User6	URL_2, URL_3
User7	URL_1, URL_3
User8	URL_1, URL_2, URL_3, URL_5
User9	URL_1, URL_2, URL_3
User10	URL_1, URL_2, URL_3, URL_4

(12 mark)

3.

a. Briefly describe the following approaches of clustering with examples in each case.

i. Partitioning approach

ii. Hierarchical approach

(4 mark)

b. Both k-means and k-medoids algorithms can perform effective clustering.

i. State Square Error Criterion used in the k-means algorithm and explain how it can be used to determine the best set of clusters for given value of k.

(2 mark)

ii. What is Swapping Cost in k-medoids algorithm? Explain the use of it to determine clusters in k-medoids.

(3 mark)

c. Discuss three (03) similarity measures that are used to determine the most similar pair of clusters in the agglomerative hierarchical clustering algorithm.

(6 mark)

d. Perform single link agglomerative hierarchical clustering for the following five (05) objects and show the results by drawing a dendrogram.

	P1	P2	P3	P4	P5
P1	0.00	0.10	0.41	0.55	0.35
P2	0.10	0.00	0.64	0.47	0.98
P3	0.41	0.64	0.00	0.44	0.85
P4	0.55	0.47	0.44	0.00	0.76
P5	0.35	0.98	0.85	0.76	0.00

(10 mark)

4. A medical researcher compiling data for a study has collected data about a set of cancer patients, all of whom suffered from blood cancer. During their course of treatment, each patient responded to one of two medications. A sample of the data collection is shown as below. Part of his job is to use decision tree classification on the sample to find out which drug might be appropriate for a future patient with the same illness.

Patient	Age	Blood Pressure	Sex	Drug
Amara	30-40	Low	F	Drug A
Malik	>40	Normal	M	Drug B
Prasad	<30	Normal	M	Drug B
Jeewa	<30	Normal	F	Drug A
Ramani	30-40	High	F	Drug A
Shiroshi	>40	High	F	Drug B
Laxmi	>40	High	F	Drug B
Dilshan	<30	Low	M	Drug B

a. "Two data sources called training and test will be used in classification". Discuss the validity of the statement by explaining why we need to do so.

(4 mark)

b. Briefly outline the major steps of decision tree classification.

(3 mark)

c. Explain how "Gain Ratio" can be used to find the splitting criterion for data set in classification process.

(3 mark)

d. Find the "Gain Ratio" of the attribute "Age".

(15 mark)