BIRLA INSTITUTE OF TECHNOLOGY, MESRA, RANCHI
(END SEMESTER EXAMINATION)

CLASS: BE **SEMESTER : VII/ADD**
BRANCH: IT **SESSION : MO/18**

SUBJECT: IT7021-DATA MINING CONCEPTS AND TECHNIQUES

TIME: 03:00 HRS. **FULL MARKS: 60**

INSTRUCTIONS:
1. The question paper contains 7 questions each of 12 marks and total 84 marks.
2. Candidates may attempt any 5 questions maximum of 60 marks.
3. The missing data, if any, may be assumed suitably.
***4*. Before attempting the question paper, be sure that you have got the correct question paper.**
5. Tables/Data hand book/Graph paper etc. to be supplied to the candidates in the examination hall.

Q.1(a) With reference to a data mining engine, what do you understand by coupling? [2]
(b) What is evolution analysis. Give some examples where this kind of data mining is useful. [4]
(c) Explain the architecture of a typical data mining system in terms of data mining primitives. [6]

Q.2(a) State the 3-4-5 rule. [2]
(b) Consider the dataset given below. If the weight is to be discretized into two values, which is a better split point between split point A and split point B. (Point A is between the 3rd and 4th tuple, point B is between 5th and 6th tuple.) [4]

Row #	Weight	Gender	
1	42	F	
2	45	F	
3	56	M	← A
4	58	F	
5	63	M	
6	69	F	← B
7	71	M	
8	80	M	

(c) Prove that $\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N}\left[\sum x_i^2 - \frac{1}{N}(\sum x_i)^2\right]$. What advantage does the right-hand side method of computing the variance provide over the left hand side?. [6]

Q.3(a) Explain the slice and dice operations with reference to a multidimensional data cube. [2]
(b) How is a ROLAP data mining engine different from a MOLAP data mining engine in terms of storage and access. If you need to store a lot of materialized cubes which one would you prefer? [4]
(c) With reference to a data warehousing system what is a lattice of cuboids. How does the existence of concept hierarchies in a dimension affect the size of the lattice? [6]

Q.4(a) How is classification different from prediction? [2]
(b) Why is accuracy not always the best method of measuring the effectiveness of a classifier. What other parameters could be used instead of accuracy? [4]
(c) Explain what you understand by tree pruning. [6]

Q.5(a) What is a "closed frequent itemset" and why is it useful? [2]
(b) Consider the transactions shown in the table below. List the frequent itemsets present in the data using the FP Tree algorithm. You can assume support is 30%. [4]

Transaction #	Items
1	Apple, Juice, Ball, Chicken
2	Apple, Juice, Ball
3	Apple, Juice
4	Apple, Pear
5	Milk, Juice, Ball, Chicken
6	Milk, Juice, Ball
7	Milk, Juice
8	Milk, Pear

(c) What are null transactions? Suggest at least two different methods used in Association rule mining to mitigate the effects of null transactions. [6]

PTO

Q.6(a) Prove that if f(x) is the activation function f(x)=1/[1-exp(-x)], then df(x)/dx = f(x)[1-f(x)] [2]

(b) How is bootstrapping different from random sampling? What proportion of tuples are likely to make it to the train set in bootstrap .632 algorithm. Explain. [4]

(c) Consider the table given below. What is the drop in the Information Gain value for the dataset when considering only the attribute "age" compared to the original data set. [$\log_{10}2 = 0.30$] [6]

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	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

Q.7(a) Justify that the running time of the K-Means algorithm is O(n), where n is the number of instances. [2]

(b) The distance between 6 data points is given below. If the initial medoiods are chosen as points "C" and "F", what is the final set of clusters obtained after completing a single iteration of PAM. [4]

	A	B	C	D	E	F
A	0	2	3	1	4	3
B		0	4	3	2	4
C			0	2	4	3
D				0	2	1
E					0	3
F						0

(c) Explain the working of ROCK. Assume that two clusters exist made up of the following transactions: [6]
Cluster 1 = {{a,b,c}, {a,b,d}, {a,b,e}, {a,c,d},{a,c,e}, {a,d,e}, {b,c,d}, {b,c,e}, {b,d,e},{c,d,e}}.
Cluster 2 = {{a,b,f}, {a,b,g},{a,f,g},{b,f,g}}. If we use a value of Θ = 0.5, then what is the distance (in terms of links) of {a,b,f} from the transactions {a,b,g} and {a,b,c}?

********28.11.18******M**