Reg. No.										

Manipal Institute of Technology, Manipal

(A Constituent Institute of Manipal University)

II SEMESTER M.TECH (CSE / CSIS) END SEMESTER EXAMINATIONS, NOV/DEC 2016

ELECTIVE II: DATA MINING AND BUSINESS ANALYTICS [CSE 542]

REVISED CREDIT SYSTEM

Time: 3 Hours 14-5-2016 MAX. MARKS: 50

Instructions to Candidates:

- Answer **ANY FIVE FULL** questions.
- Missing data, if any, may be suitably assumed.

1A. Describe each of the following data mining functionalities: Characterization, Discrimination, Association, Classification, Clustering, and Outlier analysis with real-life examples. 3M

1B. Suppose a group of 12 sales price records has been sorted as follows:
5,10,11,13,15,35,50,55,72,92,204,215
i. Normalize by decimal scaling ii. Discretize into 3 bins by clustering
iii. Sample (size =12) with SRSWR 3M

1C. Discuss the different methods used in handling missing values and noisy data. 4M

2A. A database has 9 transactions. Let min_sup = 30% . Find all frequent itemsets using Apriori algorithm. List all association rules with confidence = 75% from any one frequent 3-Itemset with least support 4M

TID	1	2	3	4	5	6	7	8	9
Items Bought	M E X C	E P C Z	E X C	M E X P C Z	M X Z	E X Z	M X C	M E X Y C	M E Y C

2B. Illustrate mining frequent itemsets using the vertical data format for the following transactional data with support count, 2: 3M

TID	T10	T20	T30	T40	T50	T60	T70	T80	T90
List of Item IDs	I1,I2, I5	I2,I4	I2,I3	I1,I2, I4	11,I3	I2,I3	I1,I3	I1,I2, I3,I5	I1,I2, I3

2C. Strong rules are not necessarily interesting. Justify with an example. How could we address this problem by augmenting the support-confidence framework for association rules by correlation? 3M

3A. Give the basic algorithm for inducing a decision tree from training tuples. What is information gain? How it can be used as attribute selection measure in building the decision tree? 3M

3B. Explain Bagging and Adaboost ensemble methods for increasing the accuracy of Classifer. 3M

3C. Suppose that we want to select between two classifier models M1 and M2. The classification of these models (for P class) along with the probability for the data tuples is given below. Justify your selection using ROC curve. 4M

Classifier M1			Classifier M2		
Tuple #	**Class Label**	**Probability**	**Tuple #**	**Class Label**	**Probability**
1	P	.9	1	P	.9
2	P	.8	2	P	.55
3	N	.7	3	N	.6
4	P	.6	4	P	.8
5	P	.55	5	P	.7
6	N	.54	6	N	.54
7	N	.53	7	N	.5
8	N	.51	8	N	.51
9	P	.5	9	P	.53
10	N	.4	10	N	.4

4A. Briefly outline how to compute the *dissimilarity* between objects described by the following types of variables:
i. Numeric variables ii. Binary variables
iii. Nominal iv. ordinal variables v. Variables of mixed types 3M

4B. Outline clustering by k-means partioning algorithm. What are its limitations and how they are addressed? Cluster the following data points into two clusters using Manhattan distance. Let records with RID 3 and 6 are initial cluster centroids. 4M

RID	**Age**	**Years_of_Service**
1	30	5
2	50	25
3	50	15
4	25	5
5	30	10
6	55	25

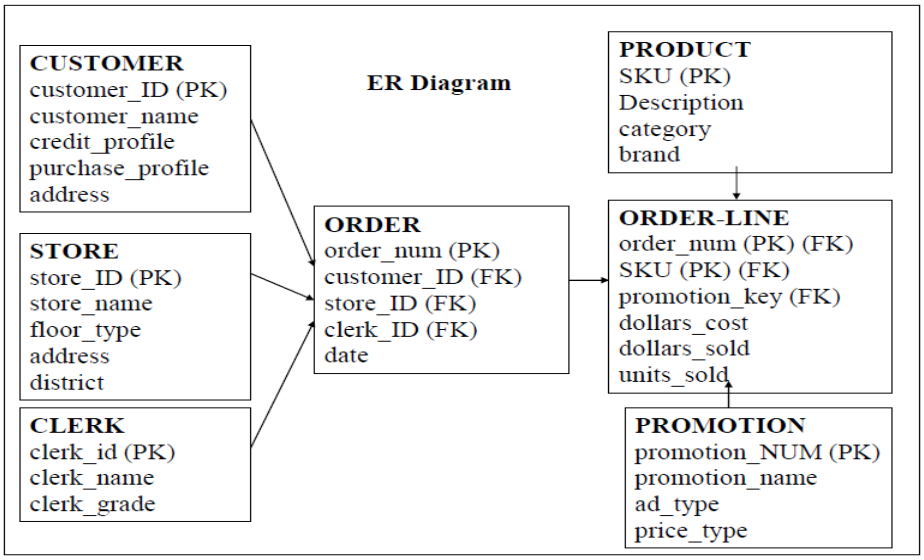
4C. Explain the supervised and unsupervised methods to measure cluster quality 3M

5A. What modifications in Hoeffding Tree algorithm leads to VFDT?. What is its Limitation? How it is handled by i. Concept-adapting VFDT and ii. Ensemble of Classifiers Algorithm? 3M

5B. Convert the following sequence DB into vertical format and find ID_lists for 1- ,2- using SPADE algorithm with minimum support count =2. 3M

SID	1	2	3	4
Sequence	<a(abc)(ac)dc>	<(ad)c(bc)a>	<(ab)dcb>	<(acd)cbc>

5C. Map the sales process in the ER Diagram into a star schema. Starting with the base cuboid [*Customer, Store, clerk, Product, Time, Promotion*], what specific *OLAP operations* should be performed
i. to list the total sales for each category in 2015
ii. to find total profit in each of the district in each of the category for TATA brand 4M



6A. What is BI? Explain its important features with an example. 3M

6B. With necessary examples, explain different types of facts and dimensions 4M

6C. What is balanced score card? What are its four perspectives? How the strategy maps depicts these perspectives in causal hierarchy? 3M

****