IDNO		NAME	

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

Second Semester 2017-18

COMPREHENSIVE EXAMINATION

PART A (CLOSED BOOK)

CS F415 Data Mining

Date : 08-05-2018(FN) **Max.Duration : 90 minutes** **Max.Marks : 40 (20%)**

I. Multiple Choice questions. Choose only one most appropriate answer. [***50% Negative*** Marking] **[1*10=10M]**

1. Which of the following coefficient is used to evaluate hierarchical clustering?
   (A) Silhouette Coefficient
   (B) CoPhenetic Correlation Coefficient
   (C) Jaccard Coefficient
   (D) Correlation Coefficient
2. Which of the Clustering algorithm uses SNN similarity measure?
   (A) Jarvis-Patrick Clustering algorithm
   (B) Minimum Spanning Tree Clustering algorithm
   (C) Chameleon Clustering algorithm
   (D) BIRCH Clustering algorithm
3. Which of the following clustering algorithm is used for categorical data?
   (A) CURE
   (B) BIRCH
   (C) ROCK
   (D) OPUSSUM
4. Which of the following measure do not violate null addition property in association rule mining?
   (A) Interest Factor
   (B) Odds ratio
   (C) Jaccard
   (D) ϕ-Coefficient
5. The computation complexity of BIRCH algorithm is
   (A) $O(n^2)$
   (B) $O(n)$
   (C) $O(n\log n)$
   (D) $O(n^3)$
6. In decision tree algorithms applied to datasets with a large number of classes and numerical attributes, the attribute selection method to avoid is:
   (A) Information gain
   (B) Gain ratio
   (C) Gini index
   (D) Gain
7. In a transactional database, the lift measure of the items bred and rice is equal to 0.5. This means that
   (A) if consumers buy bred they are more likely to buy rice
   (B) if costumers buy bred they are less likely to buy rice
   (C) if costumers buy bred they can buy rice or not with the same probability
   (D) None of the above
8. Prediction differs from classification in:
   (A) not requiring a training phase
   (B) the type of the outcome value
   (C) using unlabeled data instead of labeled data
   (D) None of the above

9. Correlation analysis is used to:
   (A) extract association rules
   (B) define support and confidence values
   (C) eliminate misleading rules
   (D) None of the above
10. In high dimensional spaces, the distance between data points becomes meaningless because:
   (A) it becomes difficult to distinguish between the nearest and farthest neighbors
   (B) the nearest neighbor becomes unreachable
   (C) Sparsity of data becomes less
   (D) there are many uncorrelated features

II.

1. Expand the following acronyms. [6]

A. SNN	
B. DIANA	
C. DENCLUE	
D. LOF	
E. CURE	
F. DBSCAN	

2. State the key idea behind following clustering algorithms: [6]

A. Chameleon	
B. Jarvis-Patrick Clustering algorithm	
C. BIRCH	
D. CLIQUE	
E. PAM	
F. DENCLUE	

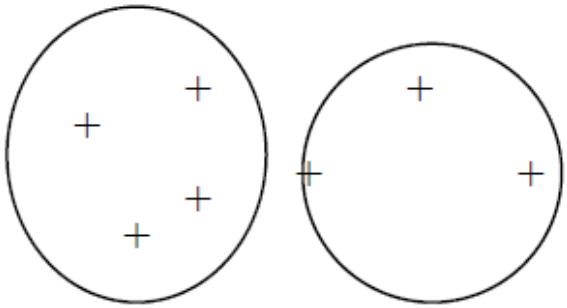
III. Write Short Answers

1. Give two examples of DIANA based clustering algorithms. [2]

IDNO		NAME	

2. Are the two clusters shown below well separated? Circle an answer: Yes No

Now in one or two sentences justify your answer. [2]



3. Explain with an example that confidence computing of an association rule does not require additional scans of the transaction data. [2]

4. Chameleon hierarchical clustering algorithm can handle data that contains clusters with widely different characteristics. Why? [2]

5. What are heavy-tailed distributions? In which data mining tasks, they pose a challenge? [2]

6. What is the difference between closed and maximal frequent item sets? How can these properties be used / exploited in frequent item set mining? [2]

7. The Fk−1 × Fk−1 method is a method for generating apriori candidates. Describe the method. With a short motivation, does this method eliminate the need for candidate pruning? [2]

8. In the figures below two bad clusterings based on K-means is shown. What is the main reason for the bad results, and what can be done to address the problems? [4]

In figure(a)	In figure(b)