CS412 "An Introduction to Data Warehousing and Data Mining" (Fall 20**)

Final Exam

(180 minutes, 100 marks, two-sheet reference, brief answers)

Name: NetID: Score:

1. [14] Data preprocessing.

   (a) [6] We have learned at least three correlation measures: (1) χ^2, (2) Pearson's correlation coefficient, and (3) Kulczynski measure.

      i. Explain what are the major differences among the three measures, and

      ii. give one example for each of the three cases that one is the most appropriate measure.

(b) [8] (Distance measures)

i. Give the name of the measure for the distance *between two objects* for each of the **4** *different kinds of data.*

ii. Give the names of **4** measures for the distance *between two clusters* for numerical data.

2. [14] Data Warehousing, OLAP and Data Cube Computation

(a) [7] Assume a base cuboid of **20** dimensions contains only two base cells:

$$(1)\ (a_1, a_2, b_3, b_4, \ldots, b_{19}, b_{20}), \text{ and } (2)\ (b_1, b_2, b_3, b_4, \ldots, b_{19}, b_{20}),$$

where $a_i \neq b_i$ (for any i). The measure of the cube is *count*.

i. How many **nonempty** aggregated (i.e., non-base) cells a complete cube will contain?

ii. how many **nonempty** aggregated cells an iceberg cube will contain, if the condition of the iceberg cube is "$count \geq 2$"?

iii. How many *closed cells* in the full cube? Note that a cell is *closed* if none of its descendant cells has the same measure (*i.e.*, count) value. For example, for a 3-dimensional cube, with two cells: "$a_1a_2a_3 : 3$", "$a_1 * a_3 : 3$", the first is closed but the second is not.

(b) [7] Suppose the **standard deviation** of n observations $x_1, x_2, \ldots, x_n$ is defined as

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{1}{n}[\sum x_i^2 - \frac{1}{n}(\sum x_i)^2]}. \qquad (1)$$

where $\bar{x}$ is the average (*i.e.*, mean) value of $x_1, \ldots, x_n$.

i. Is this measure a distributive, algebraic or holistic measure? and why?

ii. Outline an efficient algorithm that computes an *iceberg cube* with standard deviation as the measure, where the iceberg condition is $n \geq 100$ and $\sigma \geq 2$.

3. [12] Frequent pattern and association mining.

   (a) [6] Given a fixed *min_support* threshold, σ (*e.g.*, $\sigma = 0.5\%$), present an efficient incremental mining algorithm that can maximally use the previously mined information when a new set of transactions ΔTDB is added to the existing transaction database TDB.

(b) [6] Explain why both *Apriori* and *FPgrowth* algorithms may encounter difficulties at mining colossal patterns (the patterns of large size, *e.g.*, 100). A new algorithm based on *core pattern fusion* can mine such patterns efficiently. Explain why such an algorithm is efficient and effective at mining most of the colossal patterns.

4. [34] Classification and Prediction

(a) [6] All the following three methods may generate rules for induction: (1) *decision-tree induction*, (2) *sequential covering rule induction*, and (3) *associative classification*. Explain what are the major differences among them.

(b) [6] What are the major differences among the three methods for increasing the accuracy of a classifier: (1) *bagging*, (2) *boosting*, and (3) *ensemble*?

(c) [6] What are the major differences among the three methods for the evaluation of the accuracy of a classifier : (1) *hold-out method*, (2) *cross-validation*, and (3) *boostrap*?

(d) [8] Suppose you are requested to classify microarray data with 100 tissues and 1000 genes. Which of the following algorithms you would like to choose and which ones you do not think they will work? State your reasons.

(1) Decision-tree induction, (2) piece-wise linear regression, (3) SVM, (4) PatClass (pattern-based classification), (5) genetic algorithm, and (6) Bayesian Belief Network.

(e) [8] Given a training set of 50 million tuples with 25 attributes each taking 4 bytes space. One attribute is a class label with two distinct values, whereas for other attributes each has 30 distinct values. You have only a 1 GB main memory laptop. Outline an efficient method that constructs decision trees efficiently, and answer the following questions explicitly: (1) how many scans of the database does your algorithm take if the maximal depth of decision tree derived is 5? (2) what is the maximum memory space your algorithm will use in your tree induction?

5. [26] Clustering

(a) [8] Choose the best clustering algorithm for the following tasks (and reason on your choice using one sentence):

(1) clustering You-Tube videos based on their captions,

(2) clustering houses to find delivery centers in a city with rivers and bridges,

(3) distinguishing snakes hidden in the surrounding grass, and

(4) clustering shoppers based on their shopping time, the amount of money spent, and the categories of goods they usually buy.

(b) [6] Explain why BIRCH can handle large amount of data in clustering, and explain how such a methodology can be used to scale up SVM classification in large data sets.

(c) [6] What are the major difficulty to cluster a micro-array data set? Outline one efficient and effective method to cluster a micro-array data set.

(d) [6] Suppose a university database has multiple, interconnected relations: *Professor*, *Department*, *Student*, *Course*, and *Publications*. Outline an effective algorithm that may cluster *Professors* according to user's preference, *e.g.*, based on the research performance of the professors.