CS412 "An Introduction to Data Warehousing and Data Mining" (Fall 2009)
Final Exam

(Friday, Dec. 11, 2009, 180 minutes, 100 marks, two sheets of references, brief answers)

Name: NetID: Score:

1. [15] Data preprocessing.

   (a) [7] Data integration is essential in many applications. Suppose we are given a large data relation, Student, with a lot of tuples, and with attributes: (Student_Name, Major, University, Status, Office_Address). Present one effective method that can discover a set of different strings, such as "UIUC", and "University of Illinois at Urbana Champaign", essentially represent the same entity.

(b) [8] To compute some graph plots in multidimensional space, we often need to judge if a measure is algebraic, distributive or holistic. Judge which category each of the following measures belongs to and explain your judgment briefly.

(1) Boxplot

(2) Bottom-10 (among all the objects in the corresponding k-dimensional space)

(3) linear regression line

(4) confidence interval in the formula $\bar{x} \pm t_c \hat{\sigma_x}$

2. [15] Data Warehousing, OLAP and Data Cube Computation

(a) [8] Assume a base cuboid of 10 dimensions contains only four base cells: (i) $(a_1, a_2, a_3, a_4, \ldots, a_{10})$, (ii) $(b_1, b_2, b_3, a_4, \ldots, a_{10})$, (iii) $(c_1, c_2, a_3, a_4, \ldots, a_{10})$, and (iv) $(d_1, a_2, a_3, a_4, \ldots, a_{10})$, where no pair of these constants are equal. The measure of the cube is *count*.

(1) How many *nonempty* aggregate (i.e., nonbase) cells will a full cube contain?

(2) How many *nonempty* aggregate cells will an iceberg cube contain if the condition of the iceberg cube is "$count \geq 2$"?

(3) A *closed cube* is a data cube consisting of only closed cells. How many closed cells are in the full cube?

(b) [7] Databases are usually used to answer people's queries. When the expected answer set is large, it is often desirable to return only a small set of top-ranked answers. Suppose a user would like to get top-k ranked answers for Thankgiving online shopping, based on his/her own criteria of ranking. But the relevant dimension is pretty high (say over 30 dimensions). Design a data cube that may facilitate efficient processing of such queries.

3. [20] **Frequent pattern and association mining**

(a) [8] Suppose a WalMart manager is interested in only the *frequent patterns* (i.e., *itemsets*) that satisfy certain constraints. For the following cases, state the characteristics (*i.e.*, categories) of *every constraint* in each case and how to mine such patterns most efficiently.

i. The price difference between the most expensive item and the cheapest one in each pattern must be within $20.

ii. The sum of the price of all the items with profit over $10 in each pattern is at least $200.

iii. The average profit for those items priced over $50 in each pattern must be less than $10.

(b) [6] Explain why both *Apriori* and *FPgrowth* algorithms may encounter difficulties at mining colossal patterns (*i.e.*, the patterns of large size, *e.g.*, 100). Outline a method that may mine such patterns efficiently. Will such a mining method generate all the colossal patterns?

(c) [6] Frequent pattern mining often generates too many patterns. Outline two efficient methods that may generate less number but only interesting patterns.

4. [26] **Classification and Prediction**

(a) [5] What are the major differences among the three: (1) Naïve-Bayesian algorithm, (2) Bayesian Belief Network, and (3) Neural Network?

(b) [5] All the following three methods may generate rules for induction: (1) *decision-tree induction*, (2) *sequential covering rule induction*, and (3) *classification based on association (CBA)*. Explain what are the major differences among them. In a typical dataset, which one generates the most number of rules and which one generates the least?

(c) [6] Given a training set of 10 million tuples with 40 attributes each taking 4 bytes space plus one class label attribute. The class label attribute has four distinct values, whereas for other attributes each has 20 distinct values.

Some decision tree induction method uses the measure, Gini index, to measure the impurity of D, a data partition or set of training tuples, as

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2, \tag{1}$$

where p_i is the probability that a tuple in D belongs to class C_i and is estimated by $|C_{i,D}|/|D|$. Outline an efficient method that constructs such decision-tree classifier *efficiently*, and answer the following questions explicitly: (i) how many scans of the database does your algorithm take? (ii) what is the maximum memory space your algorithm will use in your induction in each scan?

(d) [5] 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*?

(e) [5] Give each situation that one of the following measure is most appropriate for measuring the quality of classification: (1) *accuracy*, (2) *F-measure*, and (3) *ROC curve*.

5. [24] **Clustering**

(a) [12] Outline the best clustering method for the following tasks (and briefly reason on why you make such a design):

(i) clustering a set of research papers based on their authors and their publication venues.

(ii) clustering a set of videos based on their image contents, captions, and where they reside on the web.

(iii) clustering UPS (or FeDex) customers for package delivery to minimize total transportation cost and have relatively even work load for each delivery employee, and

(iv) taking user's expectation expressed as a set of preferences, group students based on their academic records, research interests, and publication records.

(b) [6] Why subspace clustering is a good choice for high-dimensional data? Outline one efficient and effective subspace clustering method that can cluster a high-dimensional data set.

(c) [6] Why is it that BIRCH encounters difficulties to find clusters of arbitrary shape but OPTICS has no problem to do it? Propose some modifications to BIRCH so that it can help find clusters of arbitrary shape.