

CMPT-884 Midterm

Question 1 (4 marks). Consider a transaction database:

$t_1 = 1, 2, 3, 5$
 $t_2 = 1, 2, 3, 4, 5$
 $t_3 = 1, 2, 3, 7$
 $t_4 = 1, 3, 6$
 $t_5 = 1, 2, 4, 5, 6$

Suppose the minimum support is 60%. Find all frequent itemsets. Indicate each candidate set C_k , candidates pruned by the pruning step, and L_k .

C_1 =all items

$L_1 = 1, 2, 3, 5$

$C_2 = \langle 1, 2 \rangle, \langle 1, 3 \rangle, \langle 1, 5 \rangle, \langle 2, 3 \rangle, \langle 2, 5 \rangle, \langle 3, 5 \rangle$ ($\langle 3, 5 \rangle$ pruned)

$L_2 = \langle 1, 2 \rangle, \langle 1, 3 \rangle, \langle 1, 5 \rangle, \langle 2, 3 \rangle, \langle 2, 5 \rangle$

$C_3 = \langle 1, 2, 3 \rangle, \langle 1, 3, 5 \rangle, \langle 1, 2, 5 \rangle, \langle 2, 3, 5 \rangle$ ($\langle 1, 3, 5 \rangle, \langle 2, 3, 5 \rangle$ pruned)

$L_3 = \langle 1, 2, 3 \rangle, \langle 1, 2, 5 \rangle$

$C_4 = \langle 1, 2, 3, 5 \rangle$ (pruned)

$L_4 = 0$

Question 2 (6 marks). Suppose that you are asked to classify a collection of documents (or emails). Each document is represented by a set of keywords. You are given a training set in which each document is labeled by a class. Though the universe of keywords is very large, say a few thousands, the size of a document is usually small. Based on what you have learnt so far,

1. Briefly describe, in a few lines of English, a method for classifying documents
 2. What issues related to classification are or are not addressed by this method?
1. (1) For each training document, add its class as a new item to the document. (2) Mine all association rules of the form $LHS \rightarrow class$, where LHS contains only keywords and $class$ is an item corresponding to a class. Some minimum support and confidence must be specified. (3) For a new document, search for all association rules $LHS \rightarrow class$ with LHS contained in the document. If some rules are found, the rule with the highest confidence is used to classify the document. If no rule is found, some default class is used.
 2. This method uses the minimum support to prune overfitting rules. Since minimum support is specified by the user, this is not a reliable pruning.

Question 3 (5 marks). Suppose that you are asked to mine frequent itemsets of the form $A_{i_1} = a_{i_1} \wedge \dots \wedge A_{i_k} = a_{i_k}$ from a table T with A_i being attributes.

1. Describe a method of using data cube operator for mining frequent itemsets
 2. Compare this method with Apriori.
1. (1) Add a new measure C to T , with the value of C initialized to 1 in every row. (2) Mine data cube with C being the aggregated measure. (3) For every data cube X generated, the aggregation of C is the support of X ; if the support is above the minimum support, return X , otherwise, discard X .

2. This method will generate all possible itemsets X , whereas Apriori generates some itemsets. Also, the above generates itemsets from large size to small size. Apriori does the opposite.

Question 4 (5 marks).

1. Compare MOLAP (array-based OLAP) and ROLAP.
 2. What is the similarity and difference between attribute-oriented induction and computing data cubes?
 3. Describe a classifier that is 100% accurate on given the training set (assuming that the data does not contain any tuples that agree on all attributes except the class).
-
1. MOLAP treats a relation as an array accessed by regular array index. ROLAP treats a relation as a relation. An array can be accessed in any combination of index directly, which is needed in generated data cube, but a relation can only be scanned sequentially. However, many cells in the array do not correspond to any row in the relation, thus, compression is necessary for saving space and time.
 2. AOI rolls up the data in one way, but data cube rolls up the data in all possible ways.
 3. Create a rule $a_1, \dots, a_k \rightarrow c$ for each tuple $(a_1, \dots, a_k, c)$, where c is the class.