

Assignment 3

CIS 453/553 Data Mining, Winter 2008

due 11:59 pm, Feb 15th

1. Explain why the following algorithm is more efficient than the method for generating association rules from frequent itemsets in section 5.2.2 (section 6.2.2 in 1st edition).

Algorithm:Rule_Generator. Given a set of frequent itemsets, output all of its strong rules.

Input:

ls, set of frequent itemsets;
min_conf, the minimum confidence threshold.

Output: Strong rules of itemsets in ls.

Method:

1) for each frequent itemset l of ls
2) rule_generator_helper(l, l, min_conf);

procedure rule_generator_helper

(s: current subset of l; l: original frequent itemset; min_conf)

(1) k = length(s);
(2) if (k>1) then {
(3) Generate all the (k-1)-subsets of s;
(4) for each (k-1)-subset x of s
(5) if (support_count(l)/support_count(x) >= min_conf) then {
(6) output the rule "x=>(l-x)";
(7) rule_generator_helper(x, l, min_conf);
(8) }
(9) //else do nothing
(10)}

2. A database has five transactions. Let min_sup = 60% and min_conf = 80%.

TID	items_sold
T100	A, B, C, D, E, F
T200	A, H, S, C, F, T
T300	B, U, V, F, W, D
T400	V, B, B, C, F, X
T500	G, B, C, D, E, F

(a) Find all frequent itemsets using Apriori and FP-growth, respectively. List the results of each step. Compare the efficiency of the two mining processes.

(b) List all of the strong association rules (with support s and confidence c) matching the following metarule, where X is a variable representing customers, and $item_i$ denotes variables representing items (e.g, A, B, C):

$$\forall x \in transaction, buys(X, item_1) \wedge buys(X, item_2) \Rightarrow buys(X, item_3)[s, c]$$

(c) Suppose A is "2% milk" and V is "whole milk". If we consider both of them as "milk" (M), can we get more rules than (b)? If yes, list them.

3. Are Max patterns also Closed patterns? Are Closed patterns also Max

patterns? Prove your conclusions.

4. Give a contingency table to show that items in a strong (i.e., support and confidence are high (e.g., >60%)) association rule may actually be negatively correlated.

5. Why $\text{avg}(X) \leq v$ is a convertible constraint? Why mining with convertible constraints (e.g., $\text{avg}(X) \leq 30$) is efficient?

To turn in by paper version: Ask Cheri or Star to put your answers to Dejing's mailbox or submit to Dejing during the class or his office hour.

To turn in by emails: Send them to dou@cs.uoregon.edu. Plain text is preferred, and a pdf file is better. If you are using Word, you should be able to convert your word file to a pdf file.