Data Mining Assignment 1

Submission Date & Time: 0900Hrs on 28/03/2016

Maximum Marks: 30

The goal of this assignment is to generate interesting association rules using hash tree.

Problem Description: Use hash tree and come up with interesting association rules by using apriori algorithm.

Data: Download dataset file http://tunedit.org/repo/UCI/vote.arff

This gives the votes of 435 U.S. congressmen on16 key issues gathered in the mid-1980s, and also includes their party affiliation as a binary attribute. This is a purely nominal dataset with some missing values (corresponding to abstentions). Apply association-rule mining on this data to seek interesting associations. More information on the data appears in the comments in the ARFF file.

Support and confidence values are not fixed, experiment with different values of support and confidence, and come up with

a) Generate frequent item sets
b) Best Rules generated.
   Example: Rules found out might be as follows
   Format LHS(item set (count))->RHS(item set (count)) confidence value
   1. adoption-of-the-budget-resolution=y physician-fee-freeze=n 219 ==> Class=democrat 219 conf:(1)
   2. adoption-of-the-budget-resolution=y physician-fee-freeze=n aid-to-nicaraguan-contras=y198 ==> Class=democrat 198 conf:(1)
   3. physician-fee-freeze=n aid-to-nicaraguan-contras=y 211 ==> Class=democrat 210 conf:(1)
   4. physician-fee-freeze=n education-spending=n 202 ==> Class=democrat 201 conf:(1)

After generating rules post your results in output file.

Programming languages: C, C++, JAVA
Report:
Report should contain following things:

1. ID and names of team members
2. Language used
3. What preprocessing was done to make it amenable for association rule mining
4. Compilation steps

5. Support and confidence value at which interesting rules are generated.
6. Number of rules generated

Submission Documents

1. Source code files, along with necessary files. The code should be read/write from files. You should not use stdin/stdout for input output purposes.
2. Report in pdf format
3. Output file as described above.

Remarks

1. All submission documents should be zipped together and submitted to CMS through one of the group member's account before deadline.
2. Although output files have to be given in submission, it should be reproducible when the code is executed again. Any discrepancies will result in loosing marks.
3. As said above, there should not be any IO from stdin/stdout. Your code should execute at one go after compilation. So please include necessary files in your submission folder. There will not be any attempt to debug your code by the evaluator.

Evaluation

Exact marks for evaluation will be disclosed later. But it will have following components:

1. Completion of code
2. Successful compilation and execution
3. Association rules generated.
4. Report

Please contact following teaching assistants for any queries:

1. B. Rajitha (p2015409@hyderabad.bits-pilani.ac.in)