Data Analytics - I

Assignment 2

Topic: Attribute Oriented Induction and BUC Algorithm

Release date: 2nd Sept 2022 **Deadline**: 16th Sept 2022

Objective: This assignment has two objectives. The first objective is to gain the skill of applying attribute-oriented induction for the learning of concepts from the data. The second objective is to get the exposure to extract interesting CUBEs from the given large data set based on the BUC algorithm.

Part A: Attribute-oriented induction

Details of the data set are available at the following link [link].

1. Extract characteristic rules with attribute-oriented induction
2. Classification rules from the data set with respect to (i) property status= under construction and property status= active (ii) City= Technology city (Hyderabad, Bangalore, Chennai, Mumbai) and City= Other cities (Ahmedabad, Kolkata, Delhi, Lucknow) (iii) Price per unit area= Budget (range1 [100 - 10000]), Price per unit area= Luxury (range2 [10000 - 50000]), Price per unit area= Ultra Luxury (range3 [50000 - 100000]).
3. Produce the final summarization relation and corresponding logic formulas for the following conditions
    a. Properties within tech-emerging cities like Hyderabad, Bangalore, Chennai, and Mumbai that are active.
    b. Properties within Other cities like Ahmedabad, Kolkata, Delhi, and Lucknow that are under construction and price per unit area within the range of Luxury.

Part B: Extracting data CUBE with BUC algorithm

Details of the data set (different from the previous question) are available at the following link [link]. Conceive two-level or three-level hierarchies for relevant attributes.

https://dl.acm.org/doi/pdf/10.1145/304182.304214
Implement the Bottom-Up Cube (BUC) algorithm according to the given paper [1]. This question has two parts:

i) Assume that all CUBEs can be supported by the main memory and write a straight-forward /implementation of BUC.
ii) Assume that the program will run out of main memory and introduce paging in your implementation.

Finally, plot the following graphs over multiple runs of the algorithm while varying some parameters and keeping others constant.

a. A plot of minsup vs. runtime, keeping allotted memory fixed.
b. A plot of allotted memory vs. runtime, keeping minsup fixed.

You are also required to submit the progress report after a week stating the progress and issues that you have been facing.
[Deadline: 9th Sept 2022]

Submission Instructions

- Languages allowed: C, C++, Python.
- Submit the report with the name **<team-name>_report_A2.pdf**
- The progress report also carries marks
- Submit a zip file named **<team-name>_A2.zip** with all your work in it. Make separate folders for the two questions.
- For BUC, keep two files; one containing the implementation and another pdf file containing the outputs/graphs.

- ***Please do not plagiarize***. Discussing with your teammates and peers is encouraged, but copying any part of the code will be dealt with seriously and can lead to an ‘F’ grade in the course.

[1] Beyer, Kevin, and Raghu Ramakrishnan. "Bottom-up computation of sparse and iceberg cube." *Proceedings of the 1999 ACM SIGMOD international conference on Management of data*. 1999.