Data Analytics 1
Assignment 2
BUC Algorithm and Attribute oriented induction

Release: 19 August 2024
Deadline: 29 August 2024 (11:55 pm)

The objective of the assignment is to get the exposure to extract interesting CUBEs from the given large data set based on the BUC algorithm and get an understanding of Attribute oriented induction

Assignment Tasks

Attribute-Oriented Induction (20 marks)

Extract characteristic rules using attribute-oriented induction. This task involves data generalization through attribute removal or attribute generalization.

BUC Algorithm Implementation (55 marks)

Implement the Bottom-Up Cube (BUC) algorithm. This task has two parts:

a. **In-Memory Implementation (25 marks):** Assume that all CUBEs can be supported by the main memory and write a straightforward implementation of BUC.

b. **Out-of-Memory Implementation (30 marks):** Assume that the program will run out of main memory and introduce paging in your implementation.

Performance Analysis (15 marks)

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.

Provide a brief analysis of the trends observed in these plots.

Optimization Technique (30 marks)

Propose and implement one optimization technique for the BUC algorithm (e.g., Apriori pruning, iceberg cubing). Describe the optimization and its impact on performance.

Comparison of BUC and AOI (15 marks)

Compare and contrast the Bottom-Up Cube (BUC) algorithm and Attribute-Oriented Induction (AOI). Your comparison should include:

a. The primary purposes and use cases of each technique.

b. The types of insights or patterns each method is best suited to discover.

c. The computational efficiency and scalability of each approach.

d. The interpretability of the results produced by each method.

e. Scenarios where one method might be preferable over the other.

Provide concrete examples from your implementations to support your comparison.

Grading Criteria (Total: 150 marks)

- Attribute-Oriented Induction implementation and explanation: 20 marks
- BUC Algorithm Implementation:
    - In-Memory Implementation: 25 marks
    - Out-of-Memory Implementation: 30 marks
- Performance Analysis and interpretation: 15 marks
- Optimization Technique implementation and analysis: 30 marks
- Comparison of BUC and AOI: 15 marks
- Code quality and documentation: 5 marks
- Report clarity and completeness: 10 marks

Submission Guidelines

- Submit your code along with a report detailing your implementation, analysis, and findings.
- Include all plots and their interpretations in the report.
- Clearly explain the optimization technique you chose and provide benchmarks comparing the optimized version with the original implementation.
- Ensure your code is well-commented and follows good programming practices.

Submission format:

- Submit a zip folder named `<assignment2_teamId>` containing files `<assignment2_teamId>.ipynb` and `<assignment2_teamId_report>`