Data Mining
Information Systems Department
Faculty of Computers and Artificial Intelligence
Cairo University

Assignment 1

Association Rules

Instructions:

1. Assignment should be done individually; copies or any other method of cheating will be graded to - 5.
2. Each student can solve one problem or both, and in the second case we will consider the higher mark.
3. Total grade is 5 marks.
4. No late submissions are allowed.
5. Discussion will be during the office hours.
6. Deadline will be on 15/4 until 11:55 pm
7. Your program should include a user friendly interface.
8. The interface should enable user to select the percentage of the data needed to be read from the input file e.g. if the file contains 100 records, and the user needs to read 70% of the file then the analysis should be done on 70 records only.
9. Using the programming language, you prefer, write a program with the following specifications:
    a. **Inputs**:
        i. Excel, text, or CSV file
        ii. Min support count
        iii. Min confidence (percentage value)
    b. Outputs
        i. The frequent item sets
        ii. The strong association rules

Problem 1

Description:

- Consider the sets of transactions for a Bakery in the attached file "Bakery.csv".
    - column transaction represent a set of items defined with their names at a specific time.
    - Each transaction has many items; each item is in a separate row.
- We need to know the association between items.

Requirements:

- Write a program in any programming language that implements one of the association algorithms (Apriori, FP-Growth or vertical data format) on this set of transactions.
- Minimum support & minimum confidence should be variable as per user input during runtime.
- Then generate all association rules which can be mined from the transactions.
- The final output of your program should show the frequent item sets and association rules with their confidence.

Problem 2

Description:

- Consider the attached file "Scores.xlsx".
- It contains the number of points collected by school members of 10000 students in Tennis, Basketball and Swimming in a specific competition.
- Each row represents the scores of a one student.
- The scores that should be considered are superior, top level and ranking. The range of scores are distributed as follows:
  - Ranking: 65-74
  - Top level: 75-84
  - Superior: 85-100
- We need to know the association between scores in one sport with reference to the two other sports.
  - For example, superior in swimming derives to top level in basketball and ranking in tennis.

Requirements:

- Write a program, using any programming language, that implements one of the association algorithms (Apriori, FP-Growth or vertical data format) on this dataset.
- Minimum support & minimum confidence should be variable as per user input during runtime.
- Then, generate all association rules which can be mined from the transactions and met the minimum support and confidence.
- The final output of your program should display:
  1. The frequent item sets.
  2. The association rules with their confidence.