Assignment 8, Pattern-Based (Rule-Based) Classification

Due Date: 11:59pm, December 5, 2023

Purpose

- Understanding of pattern-based and rule-based classification
- Practice of finding classification association rules from large-scale data
- Practice of finding sub-dimensional classification rules from high-dimensional data

Description

Analyzing expression patterns of genes provides insights into cellular processes and disease mechanisms. We will investigate gene expression patterns related to two diseases, breast cancer and colon cancer, by monitoring expressions of 100 potential cancer genes for 100 patients having breast cancer or colon cancer. The gene expression is described as "up"-regulated or "down"-regulated for each gene. Find **the association rules for classification** using the **Apriori-like algorithm** and using **support and confidence as rule quality measures**. First, your program will have iterative increment of the size of gene-expression/disease sets to find all frequent gene-expression/disease sets with 30% minimum support (i.e., support threshold), for example, {gene13 down, gene59 up, ColonCancer}. Next, for each frequent gene-expression/disease set, your program will find all rules from a set of gene-expressions to a disease with 60% minimum confidence (i.e., confidence threshold), for example, {gene13 down, gene59 up} → {ColonCancer}. **Print all rules having at least two genes in the condition into an output file, each rule with its support and confidence per line**, for example, {gene13 down, gene59 up} → {ColonCancer}: 40% support, 75% confidence.

Data Set

Gene expression data ("up" or "down" regulated) for 100 samples (patients) and 100 genes are given in a tab-delimited text file. Each row represents each sample, and each column (from 2nd column to 101st column) represents each gene. The disease for each sample is shown on the last column (102nd column).

Submission

Submit your Python code, "assignment8.py", and the output file via LearnUs.

Note

- Compute support and confidence with all 100 samples. (Do not divide the data set into two subsets: breast cancer data and colon cancer data.)
- Coding competition: The fastest python code which runs correctly will have a 10% extra credit.