BÀI TẬP THỰC HÀNH 2
KHAI THÁC LUẬT KẾT HỢP

Mục tiêu:

Sinh viên học cách tìm luật kết hợp tin cậy trong tập dữ liệu weather.nominal bằng thuật toán Apriori của Weka và thuật toán FP-Growth (tự cài).

Quy định

- Làm nhóm tối đa 2 người/nhóm
- Thời hạn: xem trên Moodle
- Hình thức: thư mục bài làm có tên là MSSV1_MSSV2, bao gồm:
  - Document lưu ở dạng file *.doc(x) hoặc *.pdf: thông tin nhóm, những phần đã hoàn thành, những phần chưa hoàn thành, báo cáo trả lời các câu hỏi
  - Các file kết quả thu được theo yêu cầu trong bài

Đề bài

A. Lý thuyết

1. Hãy tìm hiểu trong tài liệu tham khảo và trình bày **chi tiết** một phương pháp cải tiến quá trình ***tìm luật kết hợp từ tập phổ biến*** (cải tiến Bước 2 trong qui trình khai thác luật kết hợp. Không phải trình bày cải tiến thuật toán tìm tập phổ biến). Giải thích vì sao nó hiệu quả hơn.

2. Cho CSDL sau và minsupp=50%, minconf=100%

TID	Items_bought
100	I, B, F, D, E, C, H, J
200	F, C, F, G, A, D, C
300	B, J, D, A, H
400	E, A, B, E, G

a) Sử dụng thuật toán Apriori để tìm tất cả các tập phổ biến. Sử dụng thuật toán Fp-Growth để tìm tất cả các tập phổ biến. So sánh kết quả. Liệt kê tập phổ biến tối đại, tập phổ biến đóng.

b) Tìm tất cả LKH có dạng **(item 1^ item 2 -> item 3)** thỏa mãn ngưỡng minsupp và minconf đã cho.

c) Ứng dụng cải tiến của câu 1 vào việc tìm các luật kết hợp ở câu b thỏa mãn ngưỡng minconf. So sánh hiệu quả về thời gian thực hiện với kết quả ở câu b).

B. Thực hành

Tập dữ liệu: plants, địa chỉ: http://archive.ics.uci.edu/ml/datasets/Plants
Đây là dữ liệu về sự phân bố của một số loài thực vật ở khu vực Mỹ và Canada.

Câu hỏi:

1. Hãy chuyển dữ liệu trong tập tin plants.data từ dạng giao dịch sang dạng nhị phân:
   - Mỗi dòng là một loài cây.
   - Cột đầu tiên là tên loài cây, các cột tiếp theo là các vùng phân bố.
   - Giá trị nhị phân gồm y và n. y đại diện cho sự xuất hiện của cây trong vùng phân bố và n là không xuất hiện.
   - *Lưu lại theo định dạng csv với tên* ***plants.cs***
2. Trả lời các câu hỏi sau:
   a. Có tất cả bao nhiêu loài cây.
   b. Có tất cả bao nhiêu vùng phân bố.
   c. Số loài cây trên mỗi vùng phân bố
   d. Vùng phân bố có ít loài cây nhất, cho biết số lượng và tỉ lệ %.
   e. Vùng phân bố có nhiều loài cây nhất, cho biết số lượng và tỉ lệ %.
   f. Trung bình một vùng phân bố có bao nhiêu loài cây.
3. Chúng ta chuẩn bị áp dụng giải thuật Apriori trên dữ liệu này, giả sử khi khai thác tập phổ biến và luật kết hợp ta chỉ quan tâm đến các vùng mà một loài cây có xuất hiện ở đó (các giá trị 'y') → dữ liệu được xem như dữ liệu giao dịch. Giải thuật Apriori trong Weka khi thực hiện sẽ bỏ qua các giá trị thiếu → chỉ cần loại các giá trị 'n' ra khỏi dữ liệu.
   - Hãy thay thế toàn bộ giá trị 'n' thành '?'.
   - Thuộc tính đầu tiên (tên loài cây) không cần thiết trong bài toán khai thác tập phổ biến, hãy xóa nó đi.
     Dữ liệu kết quả lưu thành định dạng arff với tên ***plants.arff***
4. Khai thác tập phổ biến:
   Sử dụng thuật toán **Apriori** trong Weka để khai thác tất cả tập hạng mục có độ phổ biến từ **0.1** trở lên.
   - *Trong báo cáo: Trình bày kết quả định lượng, ví dụ như sau:*

Kích thước	Số lượng
1 hạng mục	17
2 hạng mục	30
…	…
n hạng mục	…

*Danh sách tất cả các tập phổ biến thỏa yêu cầu được lưu trong tập tin **FI.doc** (SV copy đoạn chính giữa 2 dòng **Generated sets of large itemsets** và **Best rules found**)*

5. Khai thác luật kết hợp:
   - Với mỗi tập phổ biến có kích thước lớn nhất theo kết quả của câu 4:
     - Sử dụng thuật toán **FP-Growth** trong Weka để khai thác tất cả các luật kết hợp.
     - Có độ tin cậy (Confidence) từ **0.95** trở lên.
     - Chứa tất cả các hạng mục thuộc tập phổ biến đang xét.
   - *Trong báo cáo: Trình bày kết quả định lượng, ví dụ như sau:*

Tập hạng mục phổ biến	Số lượng luật
ak = y, ca =y, bd = y	2
…	…

*Danh sách tất cả các luật kết hợp thỏa yêu cầu được lưu trong tập tin **AR.doc**.*

Tài liệu tham khảo

[1] Slide lý thuyết bài 3

[2] Textbook: J. Han and M. Kamber: Data Mining, Concepts and Techniques, Second Edition - Chapter 5: Mining Frequent Patterns, Associations and Correlations (5.2.1 đến 5.2.4)

[3] J. Han, H. Pei, and Y. Yin. Mining Frequent Patterns without Candidate Generation. In: Proc. Conf. on the Management of Data (SIGMOD'00, Dallas, TX). ACM Press, New York, NY, USA 2000