Bài tập 02: Khai thác tập phổ biến và luật kết hợp

Mining Frequent Itemsets and Association Rules

Môn học: Khai thác dữ liệu và ứng dụng.

Lớp CQ2017 – học kỳ I / 2019 - 2020

Nội dung bài tập

Trong bài tập này, các bạn sẽ khai thác luật kết hợp trên dữ liệu với nhiều thuộc tính.

Ghi chú: *Tài liệu này biên soạn dựa trên một số bài tập của CE Department- Sharif University of Technology.*

Chuẩn bị

Tải tập dữ liệu **churn.txt** (attached file).

Yêu cầu bài tập

Sử dụng module association rule mining của Weka để khai thác các luật kết hợp từ tập dữ liệu trên. **Lưu ý:** do cách thể hiện dữ liệu của Weka, trong quá trình khai thác đôi khi Weka sẽ ngưng hoạt động do tràn bộ nhớ.

Chạy thử nghiệm nhiều lần trên tập dữ liệu với các tham số khác nhau cho đến khi đạt được các luật kết hợp thể hiện tốt.

Hướng dẫn cụ thể:

- **Code**: sử dụng Weka khai thác các luật kết hợp cũng như chuẩn bị dữ liệu và trình bày kết quả. Tự viết lại các hàm cần thiết.

- **Data**:
    - Có thể thực hiện trên tập con của tập dữ liệu nếu Weka không thể xử lý toàn bộ tập dữ liệu. **Lưu ý:** nếu các bạn lấy tập con ngẫu nhiên, để kết quả của mình và các bạn giống nhau các bạn nên khai báo giá trị random seed.
    - Tiền xử lý dữ liệu (nêu rõ mục đích, tác dụng của các biện pháp tiền xử lý).
    - Định nghĩa các khái niệm phân cấp (concept hierarchies) (Ví dụ: số nhà -> đường -> thành phố -> quốc gia) dựa trên các thuộc tính nhằm giúp các bạn phân tích dữ liệu ở các mức độ tổng quát khác nhau. Các bạn có thể đọc thêm về các khái niệm phân cấp ở slide này:

      https://www.slideshare.net/mauliktogadiya/data-mining-65738602.
- **Experiments**: Khai thác luật kết hợp dựa trên các hệ số khác nhau (confidence, support,...). Phân tích tập luật thu được, thử nghiệm lại với "góc nhìn" khác của dữ liệu bằng cách tổng quát hóa (hay cụ thể hóa) dữ liệu dựa trên các khái niệm phân cấp hoặc lựa chọn các phần khác nhau của dữ liệu.
- **Results**: Giả định rằng bạn là người sử dụng, mục tiêu của bạn là muốn thu được các tập luật nhằm phục vụ cho việc ra quyết định, hiểu rõ hơn về dữ liệu, hay gia tăng lợi nhuận của công ty. Khai thác các luật cho đến khi bạn đạt được một tập luật thỏa mãn mục tiêu đó.

Yêu cầu báo cáo

Báo cáo của các bạn nên được chia thành các phần như sau:

- **Data**: mô tả tập dữ liệu, các khái niệm phân cấp.
- **Code**: hướng dẫn và mô tả về đoạn code mà bạn đã viết (hay sử dụng). **Lưu ý:** cần ghi rõ nguồn (nếu bạn sao chép đoạn code ở đâu đó).
- **Experiments:**
    - Nêu rõ mục đích việc phân tích dữ liệu. Ví dụ:
        - Để "hiểu" dữ liệu rõ hơn? Vậy bạn muốn biết gì từ dữ liệu?
        - Phục vụ cho việc ra quyết định? Vậy bạn cần dựa vào dữ liệu để quyết định điều gì?
        - Phục vụ các tác vụ phân lớp, phân tích đặc tính dữ liệu,... Giải thích?
    - Với mỗi thử nghiệm:

        - Thể hiện dữ liệu: bạn dùng đã dùng dữ liệu gì cho việc thử nghiệm?
        - Bạn sử dụng phương pháp tiền xử lý nào? Tại sao?
        - Tham số của hệ thống, hệ số, độ đo sử dụng. Ý nghĩa?
        - Bạn sử dụng phương pháp hậu xử lý nào? Tại sao?
        - Phân tích kết quả thử nghiệm và ý nghĩa của chúng.
- **Tóm tắt kết quả:**
    - Cách đánh giá kết quả? Bạn dựa trên tiêu chí nào để đánh giá kết quả của bạn.
    - Tập luật tốt nhất mà bạn đạt được? Mô tả.
    - Điểm mạnh và điểm yếu trong bài tập này của bạn.

Đánh giá

	Tiêu chí đánh giá	Tỷ lệ điểm
1	Hoàn thành toàn bộ yêu cầu bài tập.	70%
2	Báo cáo trình bày tốt, giải thích rõ ràng.	30%

Quy định

Tuân thủ quy định thực hành. Báo cáo viết dựa trên các yêu cầu trên. Trang đầu ghi rõ thông tin nhóm.

Tổ chức bài nộp như sau:

- Thư mục có tên <ID nhóm> chứa:
    - Báo cáo: 1 file pdf, 1 file word nội dung như nhau.
    - Dữ liệu thu được trong quá trình thực nghiệm (nếu có).
    - Mã nguồn của chương trình.
- Nén thư mục lại với định dạng zip hoặc rar. Nộp bài tập trên moodle.

Nếu có câu hỏi nào khác, gửi email cho mình với địa chỉ duc082014@gmail.com.

Tài liệu tham khảo

[1] http://ce.sharif.edu/courses/85-86/1/ce925/assignments/files/assignDir2/ProjectDefinition1.pdf.