LAB04 – Clustering

Mục tiêu của bài tập

- Sử dụng công cụ WEKA để khảo sát thực nghiệm về hiệu quả của các giải thuật gom nhóm trên nhiều tập dữ liệu khác nhau
- Nâng cao năng lực lập trình thông qua việc tự cài đặt giải thuật gom nhóm cơ bản

Quy định

- Thời gian thực hiện: **2 tuần** (xem ngày cụ thể trên Moodle)
- Tổ chức thư mục bài làm: thư mục có tên là **⟨tên nhóm⟩** (nếu tên nhóm có dấu và khoảng trắng thì bỏ dấu và viết dính liền), bao gồm các tài liệu sau
    - Báo cáo viết trả lời các câu hỏi tự luận và hướng dẫn sử dụng chương trình tự cài đặt, định dạng **pdf**. Trang đầu tiên ghi rõ thông tin nhóm, tỉ lệ thực hiện bài tập của mỗi thành viên (nếu không ghi, mặc định tỉ lệ tương đương), những phần chưa thực hiện được (để giáo viên tránh chấm sót).
    - Dữ liệu thu được trong quá trình thực nghiệm (nếu có)
    - Mã nguồn của chương trình tự cài đặt (nếu có). Ngôn ngữ: **C++/Python**, không chấp nhận các ngôn ngữ khác.
- Nén thư mục bài làm theo định dạng **zip** hoặc **rar** và nộp theo link Moodle
- **Bài làm cần tuân thủ nghiêm ngặt đặc tả của đề bài, mọi sự khác biệt sẽ không được xét tính điểm.**
- Bài làm được đánh giá trên thang 25 rồi quy đổi về tỉ lệ tương ứng điểm thực hành.

Dữ liệu thực nghiệm

Tập dữ liệu sessions.csv chứa *100 lượt truy cập của người dùng* (user sessions) đến trang bán hàng trực tuyến ABC, trong đó giá trị "1" (hoặc "0") chỉ việc truy cập (hoặc không truy cập) đến một trang tương ứng trong tổng số *8 trang con* của ABC, bao gồm Home, Products, Search, Prod_A, Prod_B, Prod_C, Cart và Purchase. Người dùng có thể duyệt từ "Home" đến "Products" rồi đến trang của một sản phẩm. Người dùng cũng có thể tìm kiếm một sản phẩm cụ thể bằng "Search". Một lượt viếng thăm đến "Cart" chỉ việc đặt một sản phẩm vào giỏ hàng. Một lượt viếng thăm đến "Purchase" cho biết người dùng đã thanh toán các sản phẩm có trong giỏ hàng.

Nội dung thực hiện báo cáo với ứng dụng WEKA

Đọc tập dữ liệu **sessions.csv** vào WEKA Explorer tại tab Preprocess. Sử dụng giải thuật **SimpleKMeans** để phân chia user sessions thành các cụm và sử dụng chức năng Visualize cluster assignments để khảo sát các phân bố cụm thú vị.

1. (2.0đ) Gọi **k** là số lượng cụm cần phân chia. Thử nghiệm các giá trị k từ 3 đến 8, giữ nguyên các tham số khác. Sau đó, với mỗi giá trị k, ghi nhận độ lỗi SSE và các tâm cụm, đồng thời chụp màn hình WEKA Clusterer Output với nội dung tương ứng.
   Cần nộp lại các tập tin **k-sessions.arff** kết xuất từ WEKA Clusterer Visualize – Save tương ứng với các trường hợp thử nghiệm trên Không nộp tập tin (-2.0đ).

k	SSE	Cluster centroids								
			Home	Products	Search	Prod_A	Prod_B	Prod_C	Cart	Purchase
2		1								
		2								
3	…	…								

Từ kết quả ở Câu 1., hãy chọn một kết quả gom k cụm mà bạn cho là có thể giúp bạn trả lời được nhiều nhất và tốt nhất những câu hỏi dưới đây rồi trình bày lại nội dung trả lời.

2. (1.0đ) Giả sử bạn quan sát thấy một người dùng mới đã truy cập các trang là Home => Search => Prod_B. Bạn sẽ giới thiệu sản phẩm nào đến người này? Giải thích cụ thể theo số liệu tính toán về tính gần của mẫu vừa quan sát với các cụm đã có.
3. (1.0đ) Tương tự Câu 2., lần này người dùng truy cập các trang Products => Prod_C.
4. (2.0đ) Kết quả gom cụm mà bạn chọn có thể nhận diện được các hình mẫu người dùng dưới đây hay không? Nếu có, xu hướng thanh toán của những mẫu người này cao hay thấp? Dẫn chứng cụ thể bằng một số mẫu đại diện.
   - Người dùng thông thường (window shopper, xem nhiều sản phẩm),
   - Người dùng tập trung (biết rõ cần mua sản phẩm gì), và
   - Người dùng tìm kiếm (sử dụng chức năng search để tìm sản phẩm cần mua)
5. (2.0đ) Từ kết quả gom cụm mà bạn chọn, có cụm nào thể hiện sở thích mua hàng cụ thể của người dùng đối với sản phẩm đơn lẻ hay nhóm các sản phẩm hay không? Nếu có, nhận diện đặc điểm hành vi duyệt trang và xu hướng thanh toán của những người dùng trong nhóm này. Dẫn chứng cụ thể bằng một số mẫu đại diện.
6. (2.0đ) Giả sử rằng ABC đã đặt các banner quảng cáo lên một số trang nổi tiếng khác và những banner này trỏ trực tiếp đến trang của các sản phẩm A và B. Bạn có thể nhận diện cụm nào tương ứng với người dùng bị quảng cáo thu hút (tức là những người đến trực tiếp trang sản phẩm thay vì duyệt từ trang Home) hay không? Nếu có thể, cho biết chiến dịch quảng cáo cho sản phẩm nào thành công hơn.

Nội dung thực hiện cài đặt

Cài đặt chương trình đọc vào một tập dữ liệu bất kỳ có định dạng *.csv, thực hiện gom cụm bằng giải thuật k-means rồi xuất ra tập tin kết quả.

(1.0đ) Chương trình nhận dữ liệu đầu vào là **tập tin *.csv** có cấu trúc như sau

- Giả sử tập dữ liệu có **N thuộc tính số** và **M mẫu** tương ứng với các thuộc tính này. Dữ liệu được tổ chức thành bảng có M+1 dòng và N cột.
- Dòng đầu tiên chứa tên của N thuộc tính, phân cách nhau bằng dấu phẩy (","). Tên thuộc tính không có khoảng trắng và ký tự đặc biệt.
- M dòng tiếp theo, mỗi dòng gồm N giá trị, phân cách nhau bằng dấu phẩy (",").

(1.0đ) Chương trình phát sinh dữ liệu đầu ra là tập tin **model.txt** chứa thông tin tương tự như trong cửa số Clusterer output (tab Cluster – WEKA), bao gồm

- Giá trị Sum of squared errors
- Thông tin gom cụm: số lượng cụm và tâm cụm tương ứng

Ví dụ: xét tập dữ liệu iris (có trong thư mục data của WEKA)

Within cluster sum of squared errors: 16.237456311387238
Cluster centroids:

	Cluster#		
Attribute	0	1	2
	(50)	(50)	(50)
sepallength	5.936	5.006	6.588
sepalwidth	2.77	3.418	2.974
petallength	4.26	1.464	5.552
petalwidth	1.326	0.244	2.026

(1.0đ) Chương trình cũng phát sinh dữ liệu đầu ra khác là tập tin **assignments.csv** chứa thông tin tương tự như kết quả lưu từ WEKA Clusterer Visualize/Save, tức là kết quả gán cụm cho mỗi mẫu dữ liệu đầu vào. Tập tin có cấu trúc như sau

- Dữ liệu được tổ chức thành bảng có M+1 dòng và N+1 cột. Dữ liệu của M+1 dòng tương ứng với N cột đầu tiên được lấy từ tập tin dữ liệu đầu vào.
- Cột cuối cùng có tên là Cluster, giá trị tại mỗi dòng mang giá trị cụm được gán (ví dụ, 0, 1, 2, v.v) của mẫu tương ứng.

(2.0đ) Chương trình thực thi **giải thuật k-means** với cú pháp tham số dòng lệnh là

```
<ID nhom> <input> <output_model> <output_asgn> <k>
```

- <ID nhom>: tên của tập tin thực thi chương trình là ID của nhóm.

- <input>: tập tin dữ liệu đầu vào có định dạng *.csv
- <output_model>: tập tin đầu ra model.txt
- <output_asgn>: tập tin đầu ra assigments.csv
- <k>: số lượng cụm cần gom

Chương trình **xử lý tuần tự các mẫu theo thứ tự từ trên xuống**. Cần thể hiện ra màn hình console cho người dùng biết chương trình đang xử lý đến giai đoạn nào. Ví dụ: đang tính vòng lặp 1, đang tính vòng lặp 2, đang tính độ lỗi SSE, v.v.

Chương trình **xuất ra giá trị độ đo đánh giá thuộc tính theo chiến lược đã chọn** ra màn hình console trong quá trình tính toán.

(5.0đ) Chạy chương trình cài đặt với tập dữ liệu **sessions.csv**. Đối chiếu kết quả phát sinh được với kết quả của WEKA (đã thực hiện ở phần Nội dung thực hiện báo cáo viết) trên cùng giá trị k (từ 3 đến 8).

LƯU Ý: sinh viên phải TỰ CÀI ĐẶT giải thuật k-means.

Nội dung thực hiện báo cáo tìm hiểu giải thuật gom cụm

(5.0đ) Hãy tìm hiểu từ các tài liệu khoa học có uy tín (ví dụ, tìm kiếm trên Google Scholar và các nhà xuất bản thông dụng như IEEE, ACM, Springer, v.v.) một phiên bản cải tiến của giải thuật gom cụm k-means.

- Ý tưởng của giải thuật cải tiến
- Dẫn chứng mã giả và phân tích sơ bộ các bước chính trong giải thuật, cho ví dụ minh họa đơn giản theo mỗi bước
- Dẫn chứng dữ liệu thực nghiệm và số liệu thống kê từ bài báo được chọn để thấy rõ sự ưu việt của giải thuật cải tiến.

Tài liệu tham khảo

[1] Slide bài giảng lý thuyết lý thuyết

[2] Trang chủ của WEKA: http://www.cs.waikato.ac.nz/ml/weka/

[3] Textbook: J. Han and M. Kamber: Data Mining, Concepts and Techniques, Second Edition - Chapter 7: Cluster Analysis.

[4] I. H. Witten and E. Frank: Data mining, Practical Machine Learning Tools and Techniques