BÀI TẬP THỰC HÀNH
GOM NHÓM

Mục tiêu:

Sinh viên sử dụng WEKA Explorer GUI để chạy thuật toán gom nhóm **K-Means** nhằm xác định mức độ ô nhiễm của một nhà máy xử lý nước. Khác với phân lớp, các thuật toán gom nhóm dựa vào độ tương đồng của các mẫu không biết trước nhãn.

Quy định

- Hình thức: thư mục bài làm có tên là **<MSSV>**, bao gồm:
    - [Document]: thư mục chứa file *.doc báo cáo trả lời các câu hỏi. Lưu ý: ghi họ tên và mssv vào đầu bài báo cáo

Đề bài

Cơ sở dữ liệu: sử dụng cơ sở dữ liệu water-treatment.arff

- Dataset được lấy từ trang web: http://archive.ics.uci.edu/ml/datasets.html
- 527 mẫu, 39 thuộc tính

Thuật toán sử dụng: SimpleKMeans

Câu hỏi:

1. Sử dụng thuật toán SimpleKMeans của Weka để gom nhóm dữ liệu trên với
    - k = 5
    - seed = mặc định
    - distance = Euclidean Distance
    - maximum number of iterations = 500
    - Cluster mode = Use training set
2. Phân tích kết quả:
    - Thuật toán chạy bao nhiêu vòng lặp?
    - Có dữ liệu thiếu hay không? Nếu có, dữ liệu thiếu được xử lý như thế nào?
    - Cluster centroids là gì?
    - Số mẫu của mỗi nhóm (cluster) là bao nhiêu?
3. Một trong những thử thách của thuật toán KMeans là tìm ra số lượng cluster tối ưu để giảm sai số. Hãy chạy lại thuật toán với những tham số seed và numClusters khác nhau, so sánh tỷ lệ lỗi.
4. Dán vào bài làm hình ảnh minh họa kết quả gom nhóm của Weka. Giải thích đồ thị bạn nhìn thấy (trục tung/hoành biểu diễn gì? Màu sắc biểu diễn gì? Phân bố điểm biểu diễn gì?...)

   Tips: Chọn Store Clusters for Visualization trước khi Start, sau đó click phải vào dòng SimpleKMeans trong Result list, chọn Visualize cluster assigments.
5. Giải thích 4 cách đánh giá mô hình gom nhóm trong weka:
    - Use training set
    - Supplied test set
    - Percentage split
    - Classes to clusters evaluation

Tài liệu tham khảo

[1] Textbook: J. Han and M. Kamber: Data Mining, Concepts and Techniques, Second Edition - Chapter 7: Cluster Analysis (7.2 and 7.4.1)