TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO
BÀI TẬP THỰC HÀNH TUẦN 3
Khai thác dữ liệu Web

Giảng viên hướng dẫn : Khoa Phó Ngọc Đăng

Sinh viên thực hiện : Vũ Mạnh Hùng

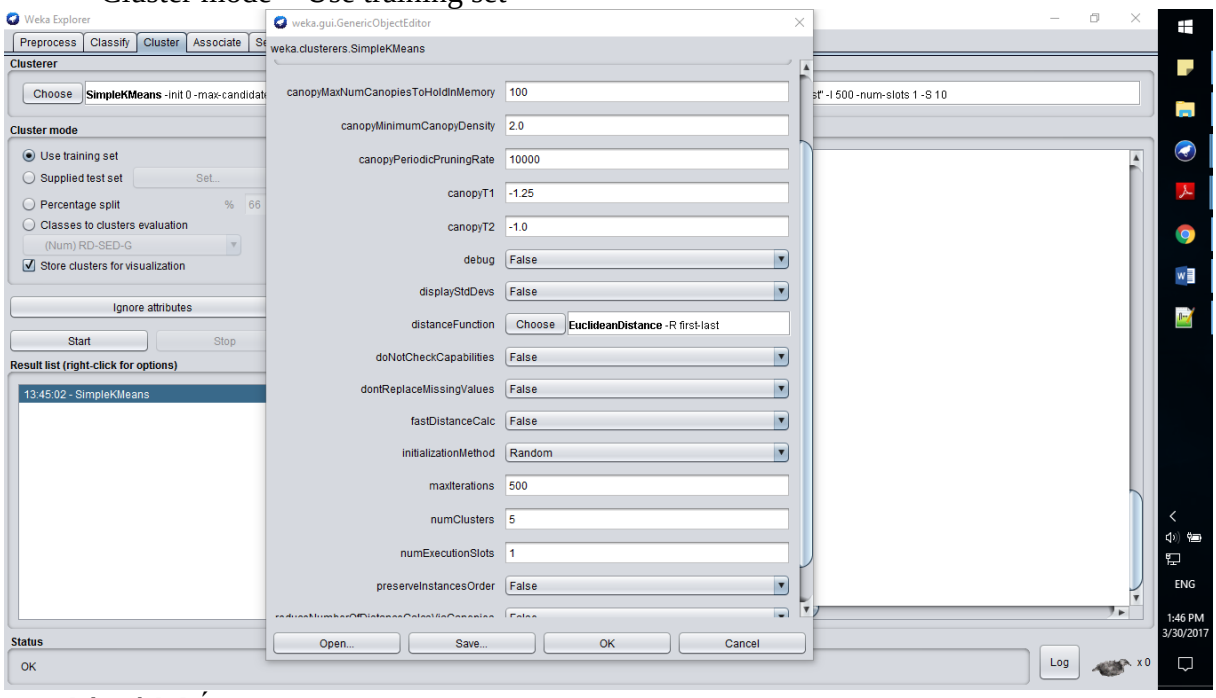
Mã số sinh viên : 1461390

Lớp : 15CK3

Ca : 1 – C6

TP HCM, tháng 4 năm 2017

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?
     Số vòng lặp lại 12.
   - Có dữ liệu thiếu hay không? Nếu có, dữ liệu thiếu được xử lý như thế nào?
     Có dữ liệu bị thiếu. Dữ liệu được xử lý theo "dontReplaceMissingValues -- Replace missing values globally with mean/mode" thay thế bằng giá trị trung bình
   - Cluster centroids là gì?
     Là điểm trung tâm của nhóm.
   - Số mẫu của mỗi nhóm (cluster) là bao nhiêu?

Nhóm	Số mẫu
Cluster0	109
Cluster1	13
Cluster2	181
Cluster3	7
Cluster4	217

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.

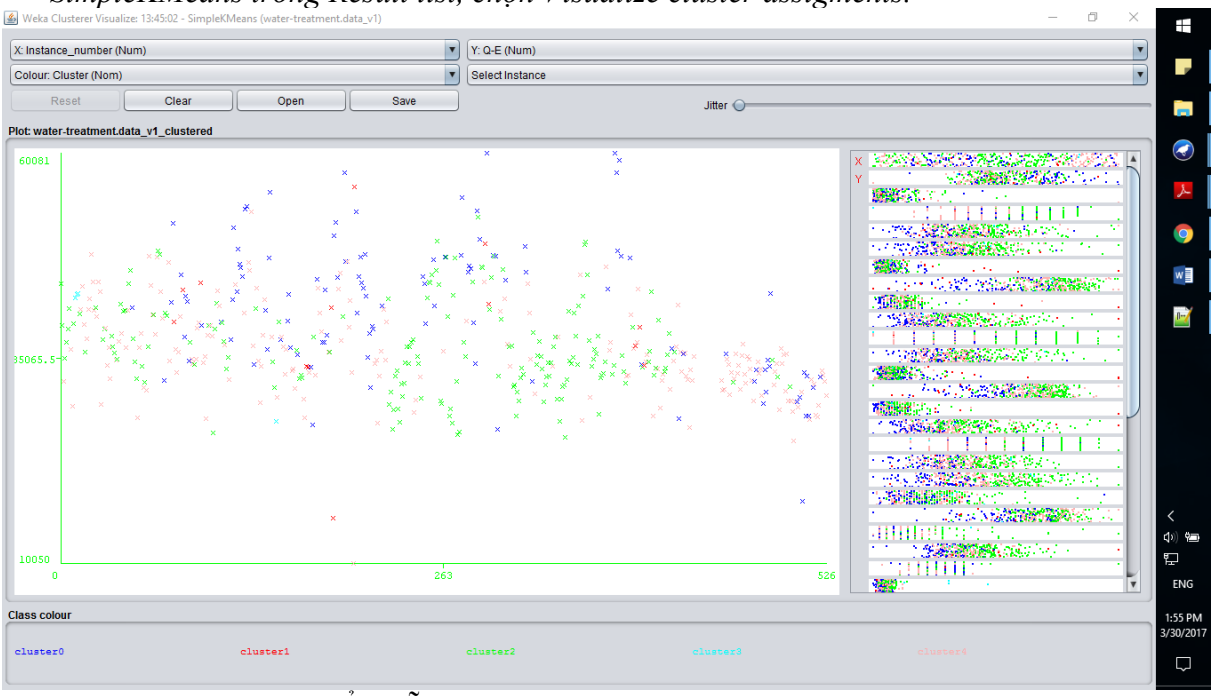
STT	seed	numClusters	Tỉ lệ lỗi
	10	2	252.245578305058
	10	5	210.1827110888923
	10	10	172.22598506010226
	10	15	155.1231900051867
	2	10	169.90316656156858

	5	10	169.9937706361494
	15	10	171.73345199005456
	20	10	173.22242104945857

Như vậy:

- Khi giữ nguyên seed, tăng numClusters thì tỉ lệ lỗi giảm nhanh
- Khi giữ nguyên numClusters, tăng seed thì tỉ lệ lỗi tăng chậm
- numClusters tác động nhiều đến tỉ lệ lỗi hơn so với seed

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.



- Trục tung/hoành biểu diễn gì?
  Trục tung biểu diễn: Instance_number(Num)
  Trục hoành biểu diễn: Q-E (Num)
- Màu sắc biểu diễn gì?
  Biểu diễn các nhóm
  - Màu lam biểu diễn cluster0
  - Màu đỏ biểu diễn cluster1
  - Màu xanh nõn chuối biểu diễn cluster2
  - Màu xanh lơ biểu diễn cluster 3
  - Màu hồng biểu diễn cluster4
- Phân bố điểm biểu diễn gì?
  Biểu diễn sự phân bố của các nhóm.

5. Giải thích 4 cách đánh giá mô hình gom nhóm trong weka:
   - Use training set: Sử dụng tập huấn luyện làm tập kiểm thử
   - Supplied test set: Chỉ định tập dữ liệu mới làm tập kiểm thử
   - Percentage split: Chia tập dữ liệu ban đầu thành 2 tập con, tập huấn luyện và tập kiểm thử theo tỉ lệ %

- Classes to clusters evaluation: Tương tự như “Use training set” nhưng có sử dụng thuộc tính gom nhóm để đối chiếu kết quả gom nhóm.