MÁY HỌC TRONG THỊ GIÁC MÁY TÍNH

GVHD: LÊ ĐÌNH DUY – MAI TIẾN DŨNG

HW1: CLUSTERING

HỌ VÀ TÊN: NGUYỄN CAO MINH
LỚP: KHTN2014
MSSV: 14520529
GITHUB: https://github.com/bigredbug47/ML_HW1_Clustering

MỤC LỤC

I. BÀI TOÁN CLUSTERING

- Bài toán clustering (gom nhóm) có mục đích gom các dữ liệu có tính chất hoặc đặc trưng liên quan tương tự nhau thành một cụm. Một số thuật toán gom nhóm phổ biến như: KMean, Spectral, DBSCAN…
- **Input:** Tập dữ liệu chưa được phân loại (chưa dán nhãn).
- **Output:** Tập dữ liệu đã được phân chia thành các cụm (cluster).

II. CÁC THUẬT TOÁN CLUSTERING CƠ BẢN

a. KMEAN:

- Thuật toán KMean xuất phát từ ý tưởng phân tập dữ liệu ban đầu thành K cụm, với K là số cụm được cho ban đầu, sao cho mỗi điểm dữ liệu có tổng bình phương khoảng cách đến tâm cụm là nhỏ nhất.
- Các bước thực hiện:
  - Chọn ngẫu nhiên K điểm làm tâm. Mỗi cụm được đại diện bằng tâm của nó.
  - Tính khoảng cách của các điểm đến tâm của cụm.
  - Nhóm các điểm có khoảng cách tới tâm gần nó nhất thành 1 cụm.
  - Xác định lại tâm mới.
  - Lặp lại việc tính khoảng cách và xác định lại tâm mới. Thuật toán dừng lại khi tâm không còn bị thay đổi.

b. DBSCAN:

- Thuật toán sẽ gom các điểm chứa lần nhau và số lượng chứa sẽ có một ngưỡng được quy định, nếu thấp hơn ngưỡng, sẽ xem đó là dữ liệu nhiễu.

c. SPECTRAL:

- Ý tưởng thuật toán: Biến đổi các đối tượng dữ liệu thành dạng đồ thị tương đồng, sau đó sử dụng phân hoạch đồ thị với k chiều nhập vào.
- Các bước thực hiện:
  - Chọn 1 hàm số để xác định độ tương đồng.
  - Xây dựng đồ thị với mỗi node là một đối tượng với giá trị các cạnh là độ lớn của sự tương đồng giữa các node.
  - Sử dụng đồ thị Lapcian để tính toán trị riêng và vecto riêng.
  - Kết luận các nhóm hình thành dựa trên trị riêng tính được.

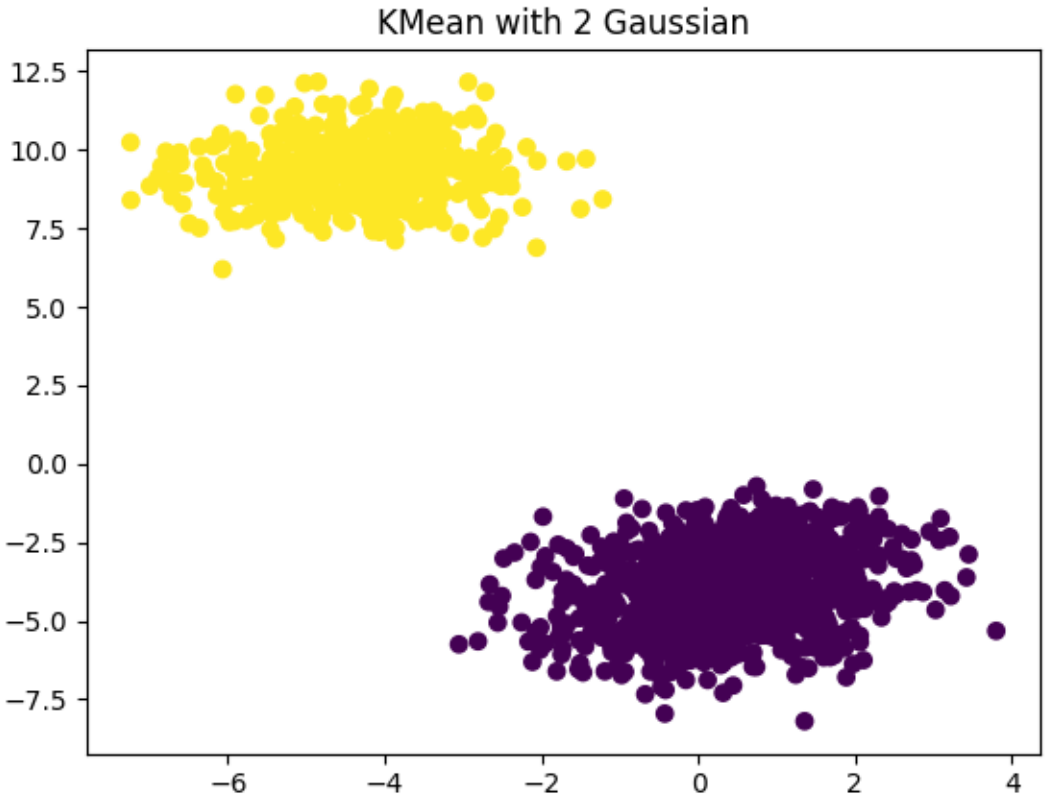
d. AGGLOMERATIVE:

- Thuật toán sẽ gom các nhóm lại với nhau (với mỗi điểm được xem là một nhóm) cho đến khi chỉ còn lại 1 nhóm duy nhất.

III. CÀI ĐẶT THUẬT TOÁN

a. Bộ dữ liệu đơn giản tự phát sinh:

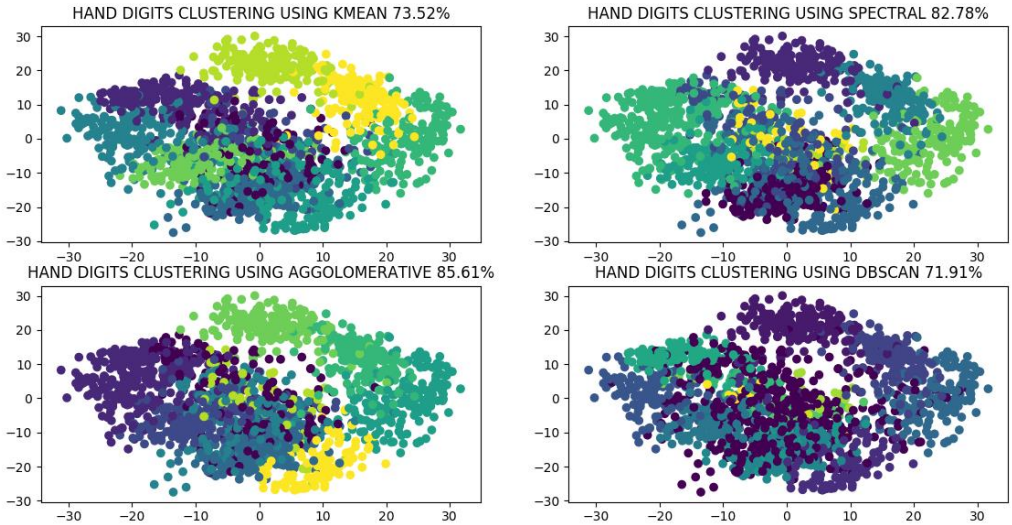
- Tạo bộ dữ liệu random với 1500 điểm, sử dụng hàm make_blobs từ thư viện sklearn.dataset.
- Sử dụng thuật toán KMean để gom thành 2 nhóm.



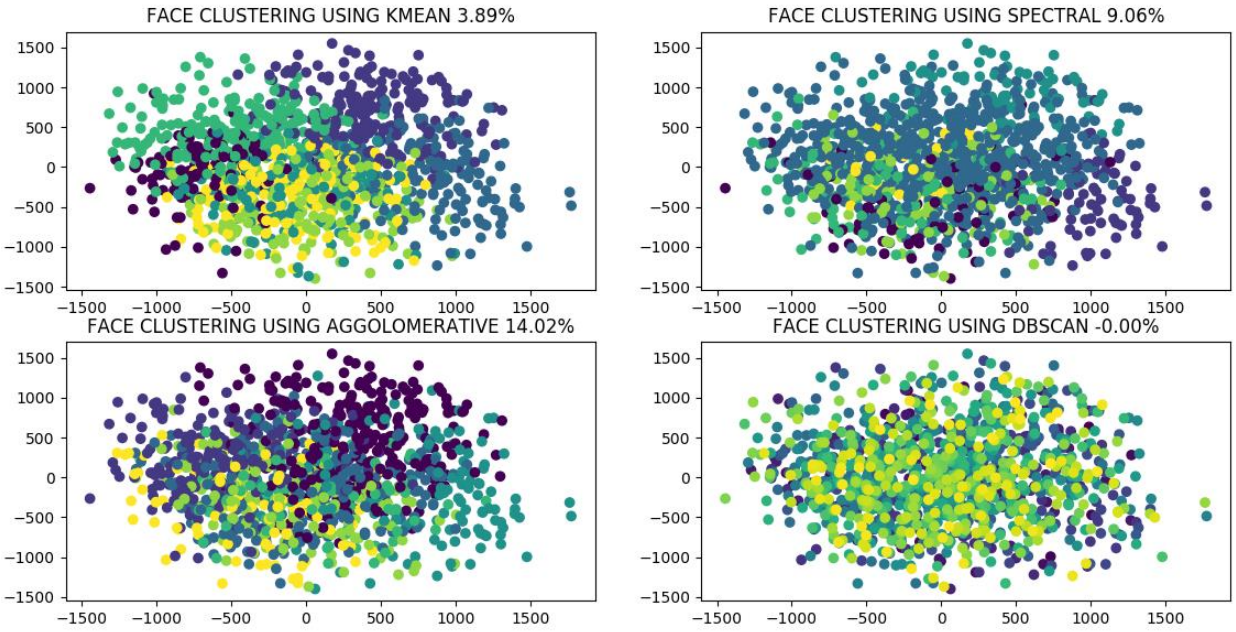
-

b. Sử dụng bộ dữ liệu chữ số viết tay với các thuật toán gom cụm cơ bản:

- Dùng bộ dữ liệu viết tay của sklearn.datasets với 10 nhóm khác nhau (ứng với 10 chữ số).
- Việc đánh gái kết quả của thuật toán dựa vào hàm scikit metrics, đưa ra tỉ lệ phần trăm đúng giữa label và kết quả sau khi clustering.
- Kết quả của các phương pháp clustering:
  - Kmean: 73%
  - Spectral: 83%
  - DBSCAN: 72%
  - Aggloremative: 85%
  - ⇨ Aggloremative Clutering đưa ra kết quả tốt nhất trong trường hợp này.

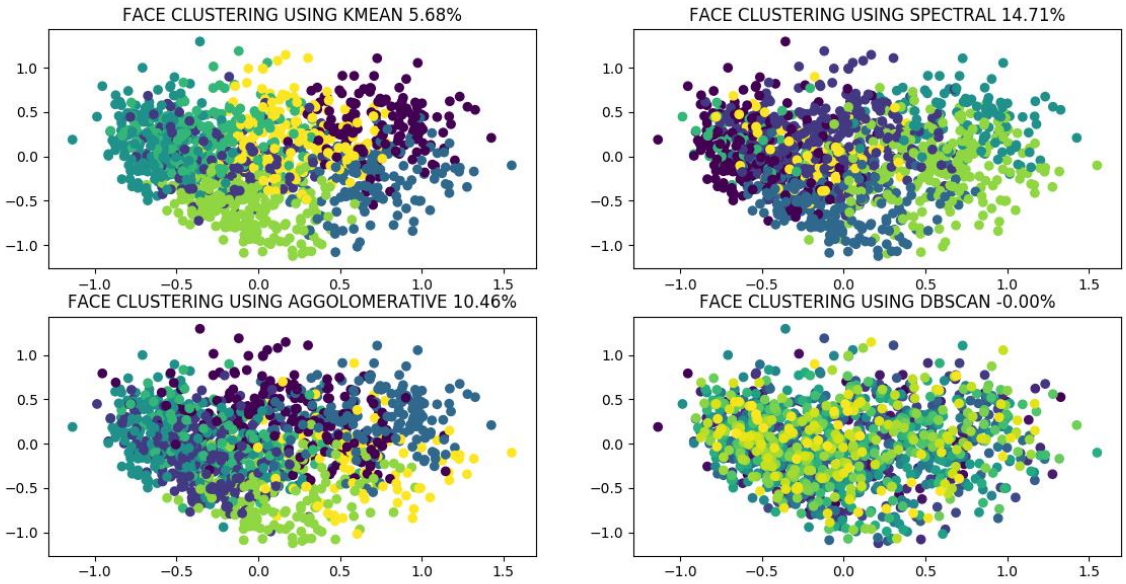

HAND DIGITS CLUSTERING USING KMEAN 73.52%
HAND DIGITS CLUSTERING USING SPECTRAL 82.78%
HAND DIGITS CLUSTERING USING AGGOLOMERATIVE 85.61%
HAND DIGITS CLUSTERING USING DBSCAN 71.91%

c. Bộ dữ liệu face với phương pháp rút trích feature LBP (local binary pattern):



- Ở trường hợp bộ dữ liệu hình ảnh face với cách rút trích feature LBP, ta thấy phương pháp DBSCAN không mang lại hiệu quả.
- Ta sử dụng phương pháp aggolomerative với hiệu quả tốt hơn (14.02%)

d. Bộ dữ face với phương pháp rút trích đặc trưng HOG (Histogram of Oriented Gradients):



- Với phương pháp rút trích đặc trưng bằng phương pháp HOG, ta thấy cách clustering Spectral mang lại hiệu quả hơn (14%) so với các cách khác.

IV. THAM KHẢO:

- M. B. O. B. Ulrike von Luxburg, “Consistency of Spectral Clustering”
- machinelearningcoban.com
- https://en.wikipedia.org/wiki/DBSCAN.
- http://www.improvedoutcomes.com/docs/WebSiteDocs/Clustering/Agglomerative_Hierarchical_Clustering_Overview.html
- http://scikitlearn.org/stable/modules/clustering.html#clustering.
- http://scikitimage.org/docs/stable/api/skimage.feature.html.
- https://en.wikipedia.org/wiki/Histogram_of_oriented_gradients.