TRƯỜNG ĐH CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC MÁY TÍNH

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc Lập – Tự do – Hạnh phúc

THỰC HÀNH 1

CLUSTERING ALGORITHMS

Môn học

MÁY HỌC TRONG THỊ GIÁC MÁY TÍNH

Lớp: CS332.I11.KHTN

GVLT: thầy Lê Đình Duy

GVTH: thầy Mai Tiến Dũng

SVTH : Trần Quang Đạt - 14520156

TP.Hồ Chí Minh, tháng 10, năm 2017

MỤC LỤC

I. GIỚI THIỆU:

1. Thuật toán Clustering là gì?

- là một trong những kỹ thuật khai phá dữ liệu. Bài toán phân cụm là 1 nhánh ứng dụng chính của lĩnh vực Unsupervised Learning (Học không giám sát), trong đó dữ liệu được mô tả trong bài toán không được dán nhãn (tức là không có đầu ra).Phân nhóm là cách nhóm các đối tượng thành các nhóm sao cho các đối tượng trong cùng một nhóm gần nhau hơn và các đối tượng của hai nhóm khác nhau khác nhau rất nhiều.
- Phân cụm ảnh thường được sử dụng để xác định vị trí các đối tượng, đường biên (đường thẳng, cong ,...). Hay nói cách khác phân vùng ảnh là một quá trình dán nhãn cho mỗi điểm ảnh trong một bức ảnh, các điểm ảnh trong cùng một nhãn sẽ có những đặc tính giống nhau về màu sắc, cường độ hoặc kết cấu của ảnh.
- *Input*: tập dữ liệu bất kì; *Output*: là tập dữ liệu đã được phân cụm, và các dữ liệu trong cùng một cụm là có tính chất giống nhau.

2. Ứng dụng của các thuật toán Clustering trong cuộc sống?

- Phân vùng ảnh được áp dụng trong nhiều lãnh vực trong cuộc sống. Kỹ thuật này là bước tiền xử lý quan trọng trong hầu hết các hệ thống xử lý ảnh, kết phân vùng tốt sẽ giúp cho quá trình xử lý về sau đạt hiệu quả cao hơn nhằm tiết kiệm về chi phí tính toán, thời gian cũng như tăng độ chính xác của các ứng dụng áp dụng nó.
- Một vài ứng dụng cụ thể trong phân vùng ảnh: Lĩnh vực hình ảnh y tế, Nhận dạng đối tượng: phát hiện đi bộ, phát hiện dừng xe,...
- Một số nhiệm vụ nhận dạng: nhận dạng khuôn mặt, nhận dạng vân tay, nhận dạng mắt..., Hệ thống giám sát giao thông, Camera giám sát an ninh.

-

II. THUẬT GIẢI VÀ PHƯƠNG PHÁP THỰC HIỆN:

1. Kmeans clustering:

- Nhận vào tập dữ liệu, chọn k(k là số lượng các cụm được xác định trước), k điểm là k tâm của k cụm đó. Tìm khoảng cách của các điểm đến k tâm là nhỏ nhất
- Phương pháp thuật hiện:
  - Chọn ngẫu nhiên k tâm cho k cụm. Mỗi tâm đại diện cho mỗi cụm.
  - Tính khoảng cách (Euclid) từ các đối tượng xung quanh đến k tâm.
  - Khoảng cách từ các đối tượng đến tâm nào gần nhất thì nhóm vào tâm đó.
  - Xác định lại tâm mới cho k cụm.
  - Lặp lại việc tính khoảng cách cho đến khi k tâm của k cụm không thay đổi.

2. Spectral clustering:

- Spectral clustering cho phép tận dụng giá trị đặc trưng của ma trận tương tự với dữ liệu lớn để thực hiện giảm chiều trước khi chia thành các kích thước nhỏ hơn. Xử lý phân nhóm tương tự như phân vùng đồ thị. Nhóm chỉ sử dụng các vector riêng của ma trận được lấy từ dữ liệu có sẵn.
- Phương pháp thực hiện:
  - Phân dataset thành k cụm.
  - Tạo ma trận tương đồng W với độ đo Euclid. Tạo ma trận D.
  - Tìm k eigenvector tương ứng.
  - Phân hoạch các nhóm dựa trên eigen vector.

3. DBSCAN clustering:

- DBSCAN clustering là vùng lân cận của mỗi đối tượng trong một cụm có số đối tượng lớn hơn ngưỡng tối thiểu.Hình dạng vùng lân cận phụ thuộc vào hàm khoảng cách giữa các đối tượng.
- Phương pháp thực hiện:
  - Tìm số lượng vùng lân cận cho mỗi điểm trong phạm vi eps, chọn các điểm có số lượng vùng lân cận >=minPTs làm core point
  - Tìm tất cả điểm có liên kết với core point.
  - Nhóm tất cả các điểm non core point vào cluster gần nhất trong phạm vi eps.

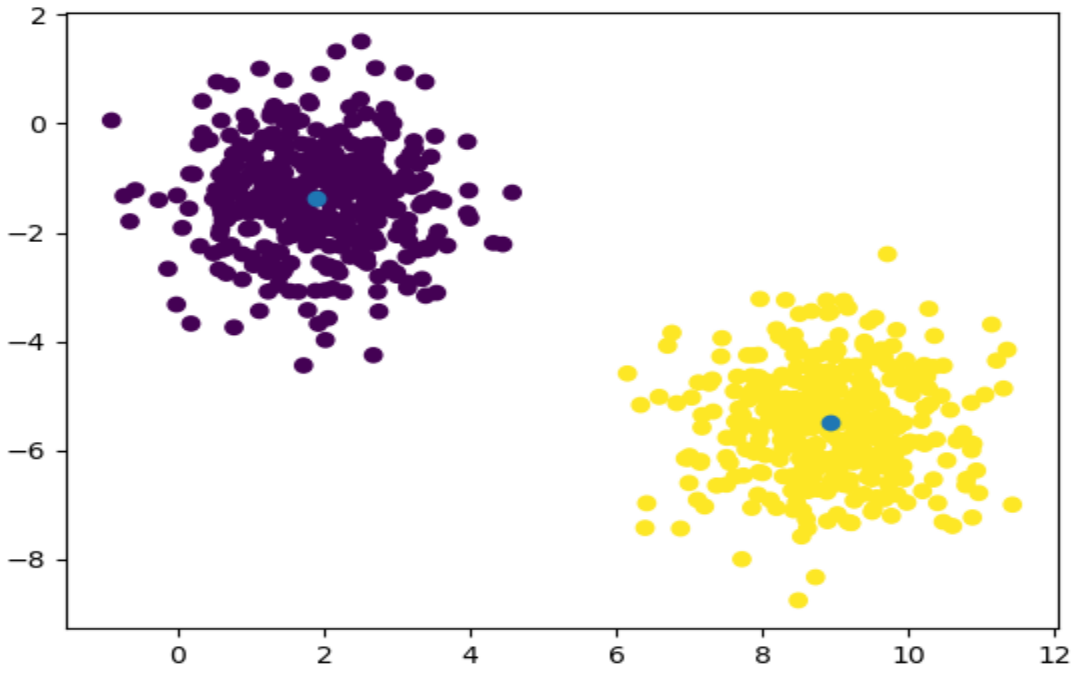
4. Agglomerative clustering:

- Là một thuật toán phân cụm đơn giản.
- Phương pháp thực hiện:
  - Chúng ta xem mỗi đối tượng là một cụm và tính khoảng cách giữa các cụm.
  - Nhóm 2 cụm gần nhất thành 1 cụm.
  - Lặp lại cho đến khi tất cả các đối tượng được nhóm vào 1cluster cuối cùng.
  - Chọn threshold để cắt dendrogram thành số cụm mong muốn.

III. THỰC NGHIỆM VÀ ĐÁNH GIÁ:

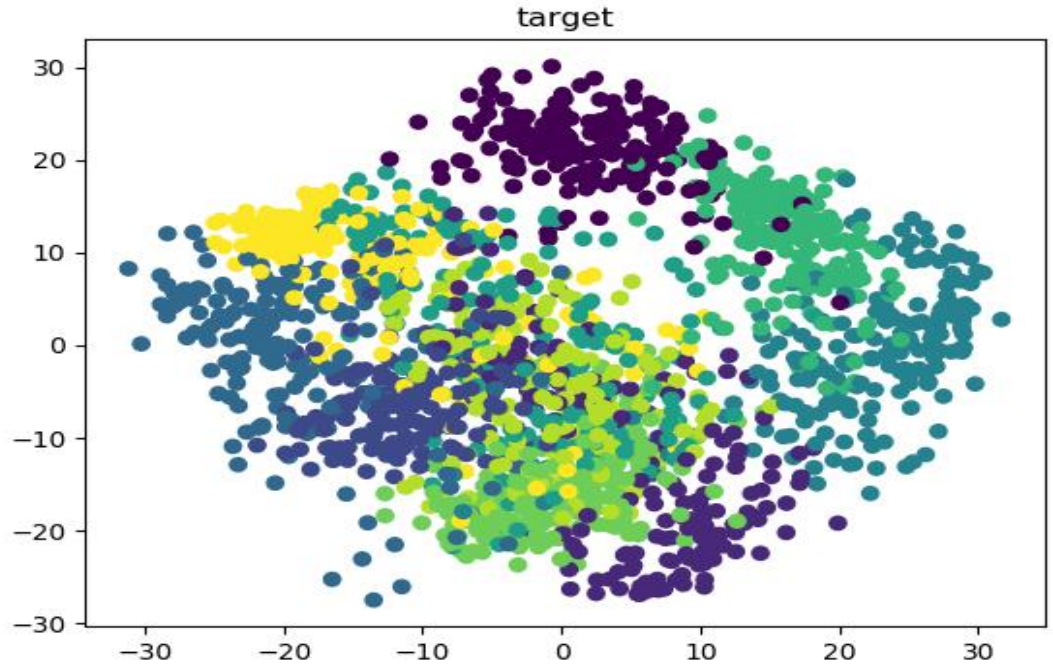
1. Bài 1: Thực hiện thuật toán Kmeans. Dữ liệu được sinh ngẫu nhiên trong chương trình gồm 2 Gaussians.

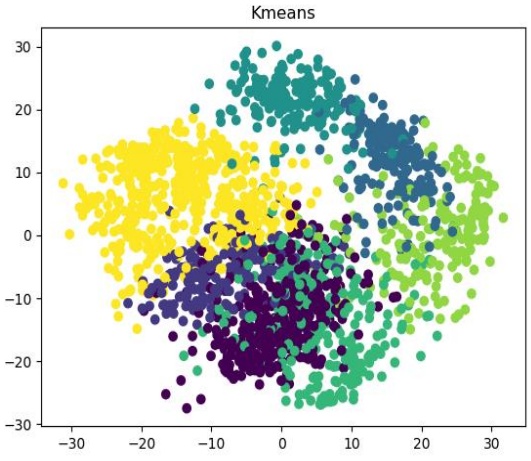
- Sử dụng hàm mat_blobs của thư viện sklearn để tạo ngẫu nhiên 2 Gaussians với 750 points. Và sử dụng hàm Kmeans có sẵn của thư viện sklearn.
- Đánh giá: có thể thấy việc phân cụm được thể hiện rất rõ rệt.
- Kết quả:

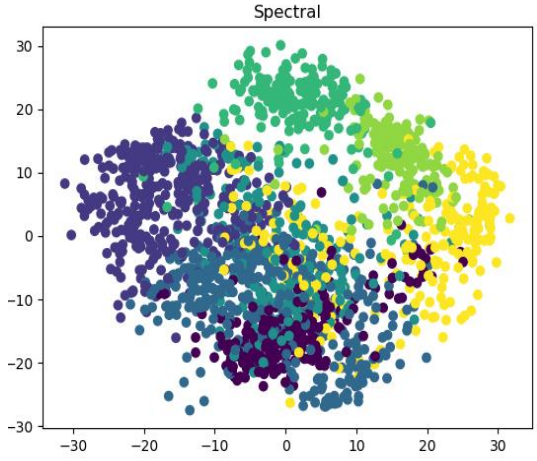


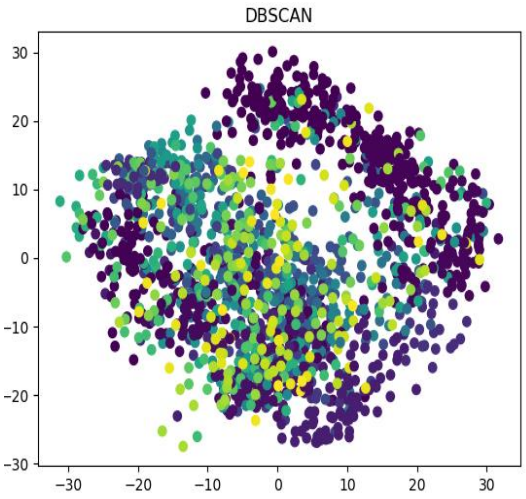
2. Bài 2: Sử dụng các phương pháp Kmeans, Spectral Clustering, DBSCAN, Agglomerative Clustering. Dùng tập dữ liệu hand-written digit.

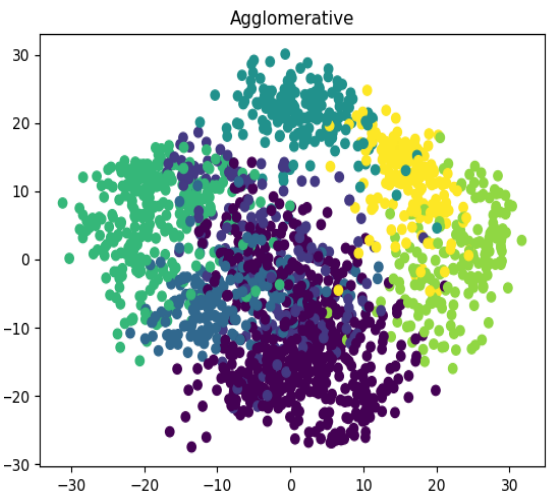
- Sử dụng hàm load_digits của sklearn để tạo bộ dữ liệu. Với số nhóm được chọn ở đây là 7. Sử dụng hàm metrics của thư viện sklearn để so sánh độ chính xác của các thuật toán.
- Kết quả:


target


Kmeans


Spectral

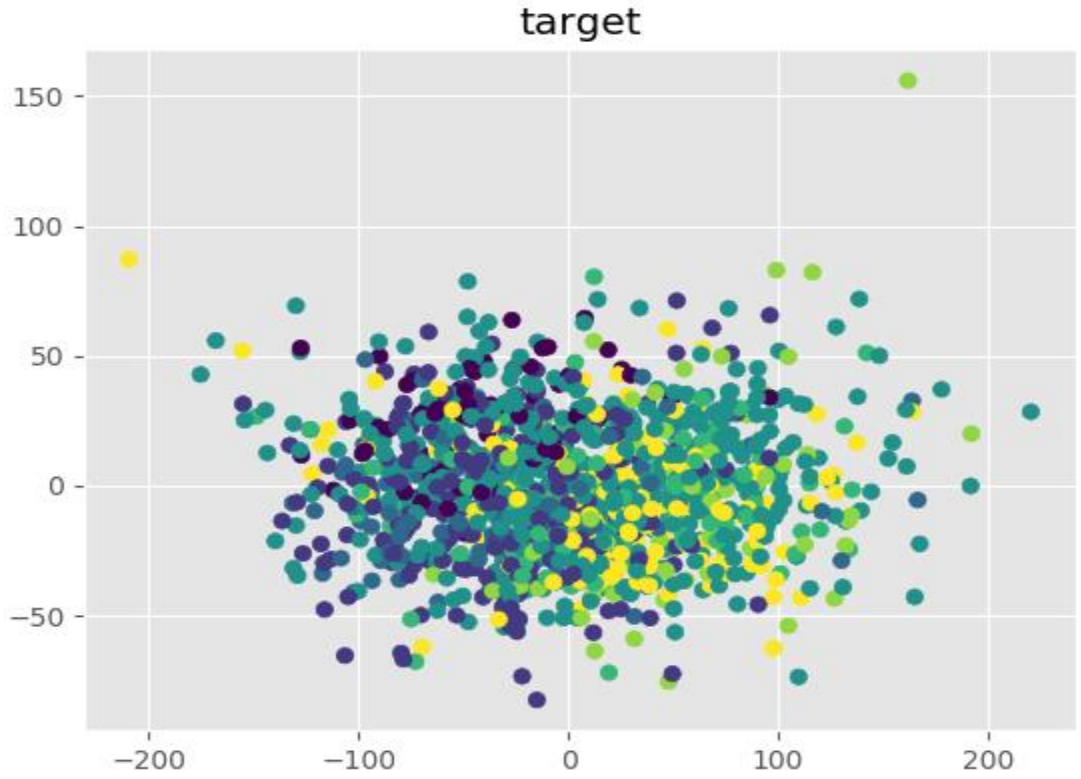

DBSCAN

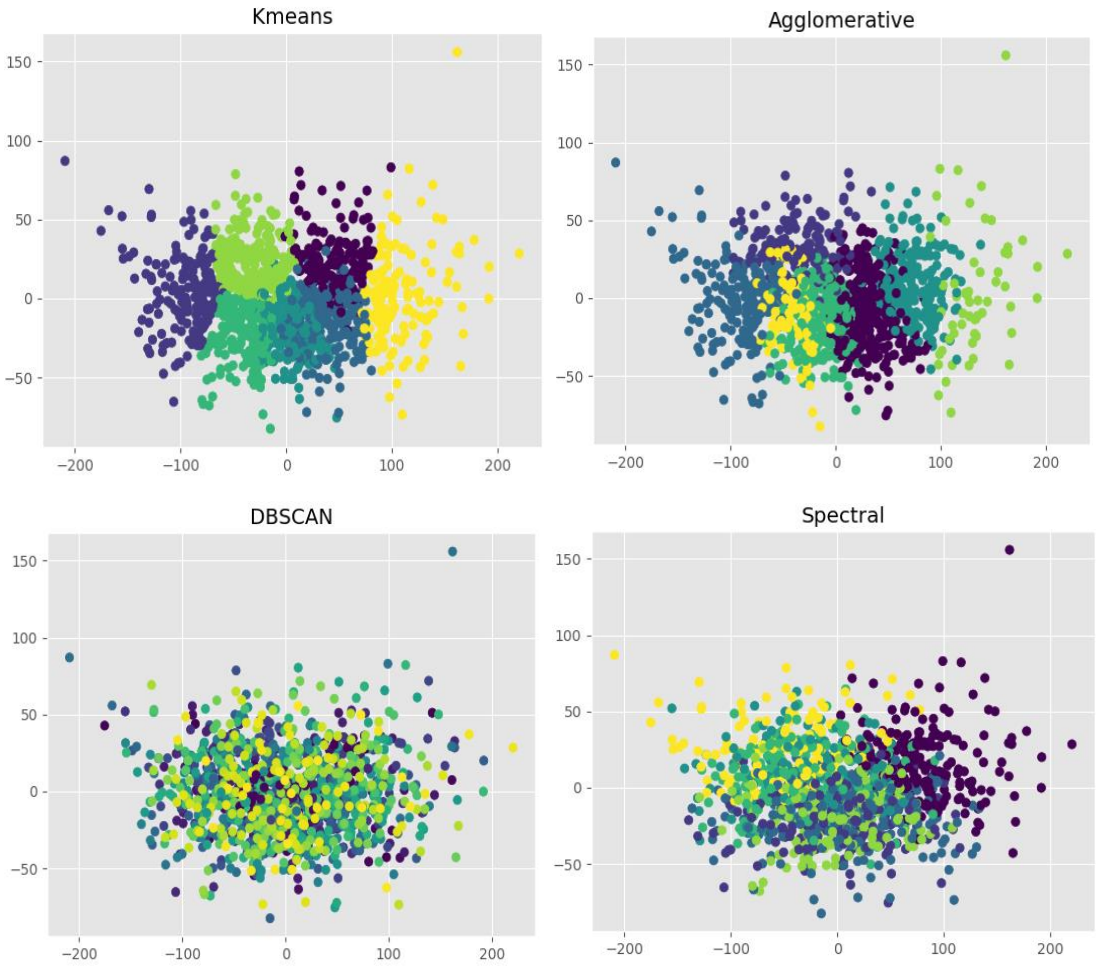

Agglomerative

- Đánh giá:
  - Với bộ dữ liệu này, các thuật toán cho các kết quả khá chính xác, gần giống với kết quả mong muốn.
  - Thuật toán Agglomerative mang lại kết quả tốt nhất(0,85), Kmens(0,74), Spectral(0,71), và DBSCAN(0,30)

3. Bài 3: Tương tự bài tập 2 - thay đổi tập dữ liệu là face.

- Sử dụng hàm fetch_lfw_people của thư viện sklearn để lấy tập dữ liệu về face của người. Trước khi lấy features thì giảm resize các ảnh. Dùng hàm local_binary_pattern của thư viện skimage để extract features từ tập ảnh. Tính toán histogram cho tập features đó. Xuất tập features cùng với tập target của tập ảnh ra file để tăng performance. Sử dụng các hàm đã viết ở bài 2 để phân cụm dữ liệu. Sử dụng hàm metrics của thư viện sklearn để so sánh độ chính xác của các thuật toán.
- Kết quả:





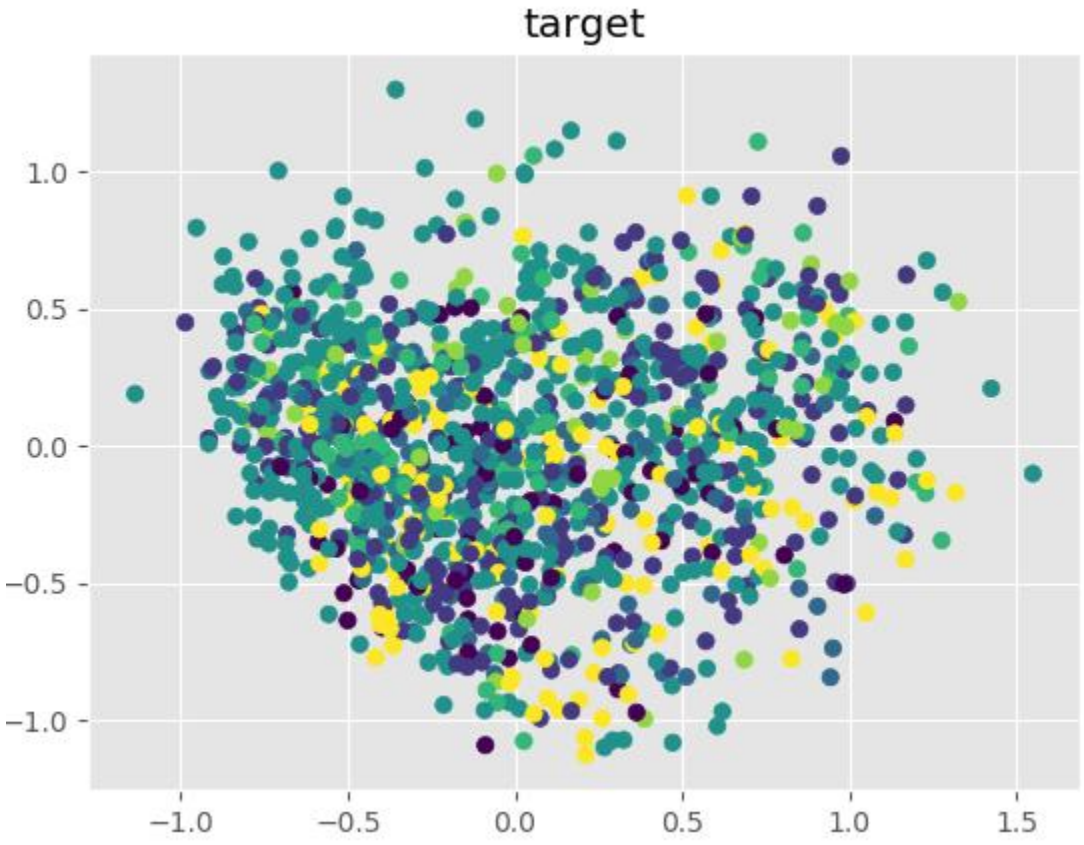
- Đánh giá:
  - Hầu hết các thuật toán đều cho kết quả không tốt. Kết quả rất thiếu chính xác với kết quả mong muốn.
  - Features LBP cho cluster dữ liệu đem lại kết quả thấp
  - Với bộ dữ liệu này Agglomerative(0,065), Kmeans(0,058), Spectral(0,051), DBSCAN(0,002).

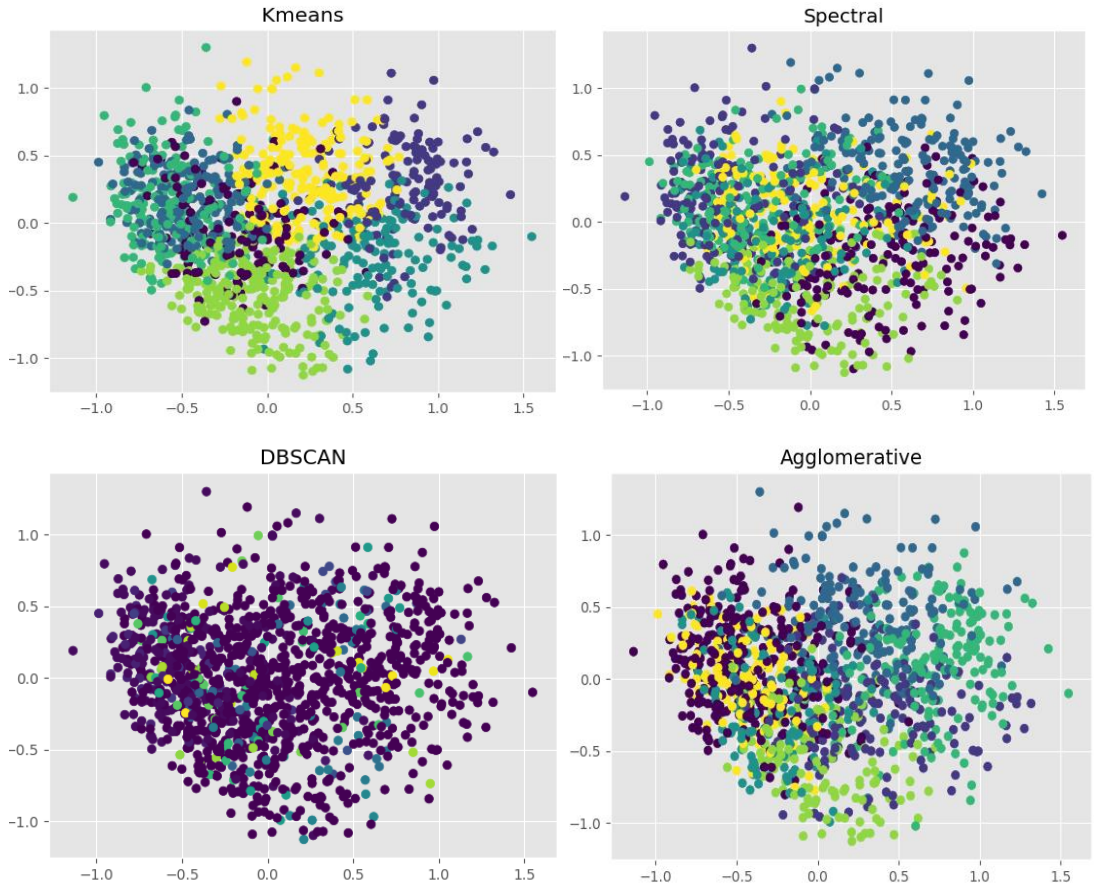
4. BÀi 4: Chọn một tập dữ liệu cho chính bạn - cùng một phương pháp rút trích đặc trưng tiên tiến.

- Ở đây sử dụng phương pháp rút trích đặc trưng HOG(Histogram Oriented Gradients.
- Vẫn tiếp tục sử dụng bộ dữ liệu face như bài 3.Sử dụng hàm HOG của thư viện skimage để extract features. Lưu tập features đã

extract và target của tập dữ liệu vào file. Sử dụng các hàm đã viết ở bài 2 để phân cụm dữ liệu. Sử dụng hàm metrics của thư viện sklearn để so sánh độ chính xác của các thuật toán.

- Các bước chính extract features trong thuật toán HOG:
  - Chuẩn hóa hình ảnh
  - Tính toán gradient của x và y
  - Tính toán histogram
  - Chuẩn hóa các block
  - Rút trích các vector đặc trưng
- Kết quả:





- Đánh giá:
  - Hầu hết các thuật toán đều cho kết quả không chính xác với kết quả mong muốn.
  - Nếu so với LBP thì dung HOG cũng chỉ cải thiện được một chút độ chính xác.
  - Với bộ dữ liệu và phương pháp này DBSCAN (0,17), Agglomerative(0,11), Kmeans(0,56), Spectral(0,54).

⇨ **Kmeans thuật toán tuy đơn giản nhất, nhưng lại khá hiệu quả và được sử dụng phổ biến.**

Tài Liệu Tham Khảo

- http://scikit-learn.org/stable/modules/clustering.html
- https://machinelearningcoban.com/2017/01/01/kmeans/
- https://en.wikipedia.org/wiki/Histogram_of_oriented_gradients
- http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_digits.html
- http://scikitimage.org/docs/dev/auto_examples/features_detection/plot_hog.html