

TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HỒ CHÍ MINH

KHOA ĐÀO TẠO CHẤT LƯỢNG CAO

NGÀNH CÔNG NGHỆ THÔNG TIN

🙡🙣



BÁO CÁO

Môn: Đồ án 2

Đề tài: Phân cụm dữ liệu với thuật toán Affinity Propagation (tuyên truyền quan hệ/ái lực)

GVHD: TS. Trần Nhật Quang

SVTH: Nguyễn Thiên Quốc 16110191

Phan Thanh Nam 16110162

Lớp: 16110CL3

TP. Hồ Chí Minh, tháng 05 năm 2019

**TÌM HIỂU VỀ ĐẠO VĂN**

1. **Khái niệm đạo văn**

Đạo văn được định nghĩa trong cuốn từ điển American Heritage như sau: "Đạo văn là sử dụng trái phép hoặc sao chép quá mức từ ngữ, ý tưởng của người khác và coi như đó là của mình".[1.1] Vì vậy, đạo văn không đơn giản chỉ là sao chép y nguyên từ ngữ mà còn bao gồm ý nghĩa của cả tác phẩm. Không thể lấy cái cớ đó là từ đồng nghĩa hay không có sự giống nhau về mặt từ ngữ để biện minh cho hành vi đạo văn.

Những trường hợp sau cũng bị coi là đạo văn: [1]

* Tải một tờ báo trên mạng.
* Thuê người khác viết bài.
* Cố tình biến ý tưởng của người khác thành của mình.

1. **Các cách để tránh lỗi đạo văn**

* Tìm hiểu kỹ vấn đề đang muốn nói tới.
* Diễn đạt lại nhiều lần bằng các cách khác nhau.
* Trích dẫn đoạn văn và nguồn của bài viết.
* Khi còn nghi ngờ, hãy nêu tên tác giả của ý tưởng đó.
* Hiểu rõ một số kiến thức cơ bản về vấn đề bản quyền.
* Nắm được những gì *không cần* trích dẫn.

1. **Cam kết**

Chúng tôi xin cam đoan đồ án này do chính chúng tôi thực hiện. Chúng tôi không sao chép, sử dụng bất kỳ tài liệu, mã nguồn của người khác mà không ghi rõ nguồn gốc. Chúng tôi xin chịu hoàn toàn trách nhiệm nếu vi phạm. Nếu đạo văn, đạo code sẽ bị phạt tuỳ theo mức độ vi phạm.

Sinh viên 1 Sinh viên 2

Phan Thanh Nam Nguyễn Thiên Quốc

**NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN**

…………………………………………………………………………………………...

…………………………………………………………………………………………...…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

Giảng viên hướng dẫn

*(Kí và ghi họ tên)*

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**NHẬN XÉT CỦA GIẢNG VIÊN PHẢN BIỆN**

…………………………………………………………………………………………...

…………………………………………………………………………………………...…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

…………………………………………………………………………………………...

Giảng viên phản biện

*(Kí và ghi họ tên)*

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**LỜI NÓI ĐẦU**

Trên thực tế không có sự thành công nào mà không gắn liền với những sự hỗ trợ, thời gian cũng như công sức đầu tư. Trong vòng khoảng 3 tháng vừa qua quả thật là quãng thời gian gian nan nhưng cũng không kém phần quý báu. Qua 3 tháng chúng em đã học được nhiều điều, tìm hiểu được thêm nhiều điều mới lạ. Cũng nhờ có công ơn của thầy Trần Nhật Quang mà chúng em mới có thể hoàn thành được đồ án này, người đã chỉ dạy, hỗ trợ và cung cấp tài liệu cần thiết. Chúc thầy nhiều sức khoẻ, nhiệt huyết để tiếp tục giảng dạy chúng em cũng như các thế hệ sau nữa!

Với khoảng thời gian giới hạn, kiến thức hạn chế cùng với nhiều bỡ khác nên sự sai sót là điều không thể tránh khỏi. Xin quý Thầy Cô xem xét và có những ý kiến đóng góp để chúng em có thể hoàn thiện hơn với những bài tập sau này. Chúng em xin chân thành cám ơn!

Mục Lục

[DANH MỤC CÁC HÌNH 1](#_Toc9516282)

[DANH MỤC CÁC BẢNG 2](#_Toc9516283)

[NỘI DUNG 3](#_Toc9516284)

[I. Mô tả project 3](#_Toc9516285)

[1. Đặt vấn đề 3](#_Toc9516286)

[2. Giới thiệu thuật toán 3](#_Toc9516287)

[3. Phương pháp đánh giá 5](#_Toc9516288)

[4. Ứng dụng 6](#_Toc9516289)

[II. Cài đặt và kiểm thử 7](#_Toc9516290)

[1. Phân cụm dữ liệu giả lập 7](#_Toc9516291)

[2. Phân cụm khách hàng bằng thuật toán AP 11](#_Toc9516292)

[III. Phân công công việc 14](#_Toc9516293)

[IV. Kết luận 14](#_Toc9516294)

[TÀI LIỆU THAM KHẢO 15](#_Toc9516295)

# DANH MỤC CÁC HÌNH

[Hình 1. Dữ liệu 2 chiều, các cụm tách rời nhau 7](#_Toc9516316)

[Hình 2. Thuật toán phân 3 cụm, hội tụ sau 131 vòng lặp. V-measure = 1 8](#_Toc9516317)

[Hình 3. Kết quả của chương trình, chưa tối ưu hoá vòng lặp, t = 287(s) 8](#_Toc9516318)

[Hình 4. Kết quả của chương trình, sau khi đã tối ưu hoá vòng lặp, t = 7.5(s) 9](#_Toc9516319)

[Hình 5. Dữ liệu 2 chiều bị chồng lấp, 3 cụm dữ liệu 10](#_Toc9516320)

[Hình 6. Thuật toán phân 2 cụm dữ liệu, hội tụ sau 874 vòng lặp 10](#_Toc9516321)

[Hình 7. Kết quả của thuật toán, V-measure = 0.449 11](#_Toc9516322)

[Hình 8. Dataset đưa vào hệ thống 12](#_Toc9516323)

[Hình 9. Thuật toán chia làm 15 nhóm, hội tụ sau 71 vòng lặp 12](#_Toc9516324)

[Hình 10. Kết quả của thuật toán. V-measure: 0.915 13](#_Toc9516325)

# DANH MỤC CÁC BẢNG

[Bảng 1. Phân công công việc của các thành viên 14](#_Toc9516355)

# NỘI DUNG

## Mô tả project

### Đặt vấn đề

Ngày nay, nhu cầu thông tin trong nhiều khía cạnh của cuộc sống thực sự rất cao. Các thông tin cần phải được gửi một cách nhanh chóng và chính xác. Vì vậy quá trình trích xuất thông tin từ dữ liệu là rất quan trọng. Có nhiều cách trong việc khai thác thông tin từ dữ liệu, một trong số đó là *phân cụm*. Phân cụm thường được sử dụng để phân tích dữ liệu có dữ liệu lớn hoặc thậm chí rất lớn về số lượng và nhãn (label) trên dữ liệu không xác định. Do việc gán nhãn cho số lượng lớn dữ liệu là công việc tốn rất nhiều chi phí và thời gian, chúng ta cần phải tìm một cách tiếp cận khác, cần thiết để khai thác thông tin hữu ích từ dữ liệu. Phân cụm tập trung vào việc tìm kiếm các phương pháp để phân tích cụm hiệu quả và hiệu quả trong các cơ sở dữ liệu lớn. [2]

Trong phân cụm thường cần phải xác định trước một vài *centers* của cụm để đảm bảo rằng *tổng các lỗi bình phương* (sum of squared errors) giữa mỗi điểm dữ liệu và *centers* tiềm năng của nó là nhỏ trong quá trình phân cụm. Kỹ thuật cổ điển để phân cụm, chẳng hạn như phân cụm *K-means*, phân vùng dữ liệu thành các cụm *k* nhưng nhiều trường hợp *k* là ẩn số hoặc thuật toán rất nhạy cảm với tập *centers* của dữ liệu ban đầu, vì vậy thường cần phải chạy lại nhiều lần để có được kết quả khả quan.

Một cách tiếp cận mới trong phân cụm ra đời vào năm 2007 có tên là *Affinity Propagation* (AP) đã được giới thiệu để giải quyết vấn đề này. [3]

### Giới thiệu thuật toán

[3] Thuật toán AP được công bố vào năm 2007 bởi Brendan Frey và Delbert Dueck trong ngành khoa học. AP là một dạng của *Học không giám sát* (Unsupervised Learning), được sử dụng khi bạn có dữ liệu không được gán nhãn – tức là, dữ liệu không có hạng mục hoặc nhóm chưa xác định. Mục đích của thuât toán này là tìm các nhóm trong dữ liệu nhưng không giống như phân cụm *k-means*, AP không đòi hỏi số lượng cụm được xác định hoặc dự đoán trước khi chạy thuật toán.

Thuật toán trao đổi thông tin giữa các cặp điểm dữ liệu cho đến khi một tập các *dữ liệu mẫu* (exemplars) được xác định, với mỗi dữ liệu mẫu tương ứng với một *nhóm* (cluster). Thuật toán AP lấy một số thực *s(k,k)* cho mỗi điểm dữ liệu – quy vào là một *“preference”* (ưu tiên). Điểm dữ liệu với giá trị *s* lớn hơn sẽ có khả năng trở thành dữ liệu mẫu hơn. Số lượng nhóm sẽ ảnh hưởng bởi giá trị ưu tiên này và thủ tục trao đổi thông tin.

Ví dụ: Cho *x1* đến *xn* là một tập hợp các điểm, không có giả định về cấu trúc, và cho *s* là một hàm để xác định tính đồng dạng giữa bất kì 2 điểm dữ liệu, sao cho *s(xi,xj) > s(xi,xk)* khi và chỉ khi xi đồng dạng với *xj* hơn với *xk*. Mục đích của AP là làm nhỏ *lỗi bình phương* (Euclidean Distance), áp dụng khoảng cách euclidean âm bình phương cho điểm *xi* và *xk* :[3]

Đường chéo của *s* (ví dụ: *s(i,i)*) đặc biệt quan trọng, vì nó đại diện cho *preference input*, nghĩa là khả năng một đầu vào cụ thể có khả năng trở thành một *exemplar* như thế nào. Khi nó được đặt thành cùng một giá trị cho tất cả các đầu vào, nó sẽ kiểm soát số lượng lớp mà thuật toán tạo ra. Một giá trị gần với *min(s)* có thể tạo ra ít lớp hơn, trong khi giá trị gần hoặc lớn hơn *max(s)* có thể tạo ra nhiều lớp. Nó thường được khởi tạo với số trung bình *median(s)* hoặc *min(s)*.

Thuật toán tiến hành bằng cách kiểm tra thay phiên 2 thủ tục trao đổi thông tin để cập nhật 2 ma trận: [3]

* Ma trận “responsibility” (trách nhiệm) **R** có các giá trị *r(i, k)* được gửi từ điểm dữ liệu *i* đến các điểm “ứng cử viên” (candidate) dữ liệu mẫu *k*. Nó phản ánh mức độ phù hợp của điểm *k* để làm dữ liệu mẫu cho điểm *i*.
* Ma trận “availability” (khả dụng) **A** có các giá trị *a(i, k)* được gửi từ điểm “ứng cử viên” dữ liệu mẫu *k* đến điểm *i*. Nó phản ánh về mức độ phù hợp của điểm *i* khi chọn điểm *k* làm dữ liệu mẫu.

Các lần lặp được thực hiện cho đến khi ranh giới cụm không thay đổi qua một số vòng lặp hoặc sau một số vòng lặp được xác định trước. Các exemplars được trích ra từ ma trận cuối cùng ‘responsibility + availability’ với những phần tử là số dương (ví dụ: *r(i,i) + a(i,i) > 0*).

Các bước mô tả tóm tắt của thuật toán:

**Đầu vào:** Dữ liệu *x* chưa được phân cụm (chưa có label)

**Đầu ra:** Các *exemplars* cùng với label vector cho từng điểm dữ liệu *y*.

Bước 1: Tạo 2 ma trận A, R = 0 ứng với 2 thủ tục trao đổi thông tin và ma trận S bằng cách tính khoảng cách euclidean âm bình phương cho mỗi điểm dữ liệu.

Bước 2: Tạo 2 biến *preference, damping*. Chèn số preference vào đường chéo của ma trận S tạo ở bước 1. [4] Damping là yếu tố để ổn định số và có thể làm chậm tốc độ hội tụ (learning rate). Theo **sklearn** chọn damping trong phạm vi từ 0.5 đến 1. Sau đó tạo 1 ma trận last\_exemplar = 0 để lưu lại giá trị exemplar mỗi vòng lặp cho việc so sánh.

Bước 3: Cập nhật 2 ma trận A, R.

Bước 4: Tính A + R và trích các exemplars. Nếu exemplars không thay đổi so với vòng lặp trước thì dừng thuật toán.

Bước 5: Lưu giá trị exemplars vào last\_exemplar. Quay lại bước 3.

### Phương pháp đánh giá

Có nhiều phương pháp để đánh giá chất lượng phân cụm, tuỳ vào các bài toán khác nhau mà ta sử dụng các phương pháp khác nhau. Các phương pháp thường được sử dụng: Adjusted mutual information, Silhouette Coefficient, V-Measure,.. Báo cáo này sẽ sử dụng V-Measure để đánh giá.

[10]Cho kiến thức về “ground truth” (sự thật nền tảng) phân chia lớp của các samples, có thể định nghĩa một số số liệu trực quan bằng cách sử dụng phân tích entropy có điều kiện.

Rosenberg và Hirschberg (2007) đã xác định hai mục tiêu mong muốn sau đây cho bất kỳ bài toán phân cụm:

* Homogeneity (tính đồng nhất): mỗi cụm chỉ chứa các thành viên của một lớp duy nhất.
* Completeness (tính đầy đủ): tất cả các thành viên của một lớp nhất định được gán cho cùng một cụm.

Chúng ta có thể biến những khái niệm đó thành điểm số homogeneity\_score và completeness\_score. Điểm số sẽ từ 0.0 đến 1.0 (cao hơn là tốt hơn).

Với *H(C|K)* là một entropy có điều kiện của các lớp được gán các cụm và được đưa ra bởi:

Và *H(C)* là một entropy của các lớp, đưa ra bởi:

Với: *n*: tổng số các samples.

*nc* và *nk* : số lượng samples tương ứng thuộc lớp *c* và cụm *k*.

*nc,k* : số lượng các samples của lớp *c* được gán vào cụm *k*.

Entropy có điều kiện của các cụm cho lớp *H(K|C)* và entropy của cụm *H(K)* được xác định theo cách đối xứng.

V-Measure là một *trung bình điều hoà* (harmonic mean) của 2 điểm số trên:

### Ứng dụng

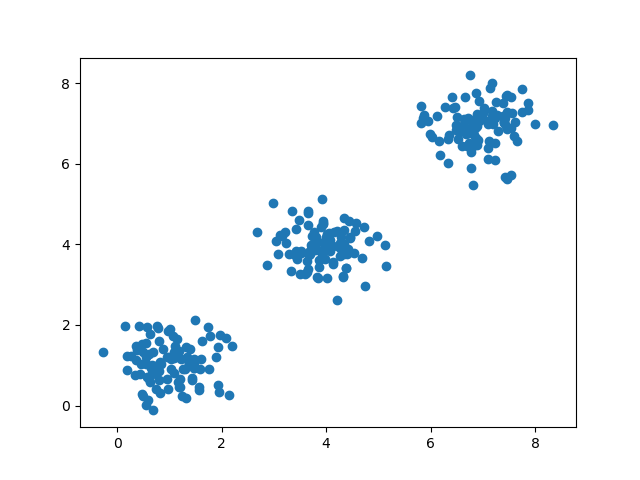
Các nhà phát minh của AP cho thấy nó tốt hơn cho computer vision (thị giác máy tính) và task liên quan đến computational biology (sinh học tính toán), ví dụ: phân loại các hình có mặt người,… so với thuật toán *k-means*[3], kể cả khi *k-means* được cho phép chạy lại ngẫu nhiên và khởi tạo bằng cách sử dụng *Phép phân tích thành phần chính* (PCA)[5].

[6]AP còn được sử dụng để phân tích lưu lượng mạng, phát hiện các điểm đáng chú ý trong hình ảnh, phân nhóm các hành động/chuyển động trong video, xác định mô hình trong luồng âm thanh, phân nhóm các phân tử sinh học hoạt động tương tự nhau, phân nhóm các đối tượng vật lý thiên văn,…

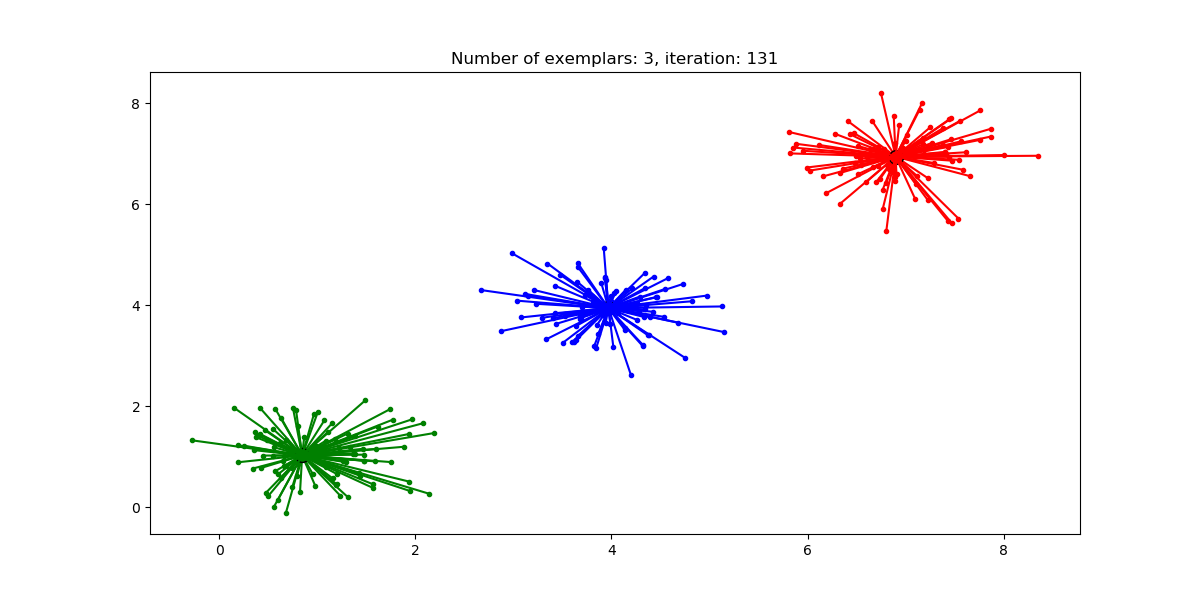
## Cài đặt và kiểm thử

### Phân cụm dữ liệu giả lập

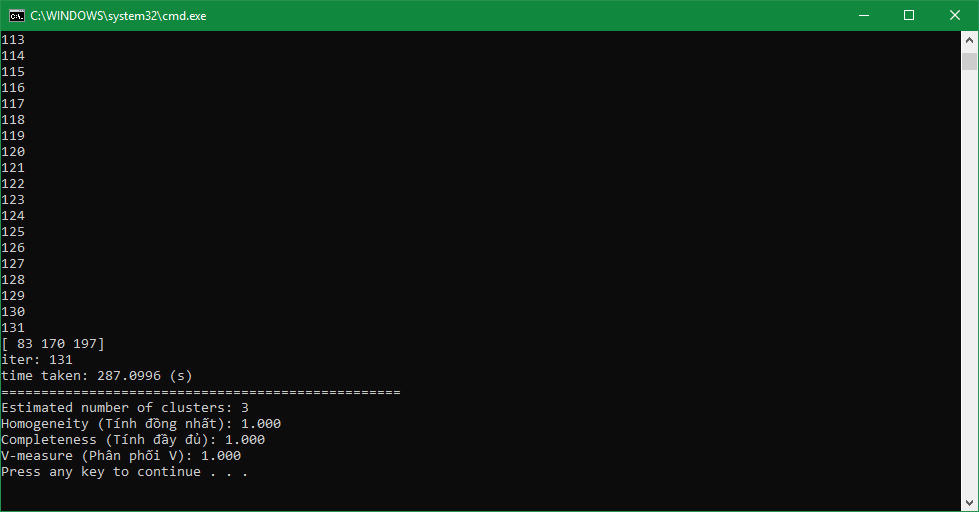
Với dữ liệu điểm 2 chiều (x,y) và những cụm điểm tách rời nhau, khi chạy với thuật toán AP thì ta có thể thấy kết quả chính xác 100% thông qua việc xác định các kiểm V-measure = 1. Nhưng có một nhược điểm của AP là việc tính toán cập nhật 2 ma trận A, R với số lượng dữ liệu nhiều như thế khiến cho độ phức tạp của bài toán là *O(n2)*. Rất mất nhiều thời gian. Chúng ta nên tối ưu hoá hàm bằng cách vector hoá vòng lặp. Thời gian chạy thuật toán sẽ được cải thiện rất nhiều!



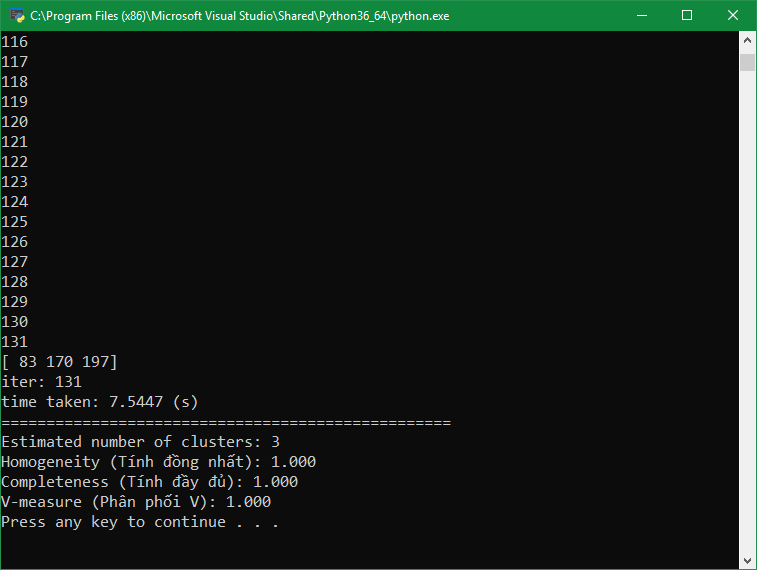
Hình 1. Dữ liệu 2 chiều, các cụm tách rời nhau



Hình 2. Thuật toán phân 3 cụm, hội tụ sau 131 vòng lặp. V-measure = 1



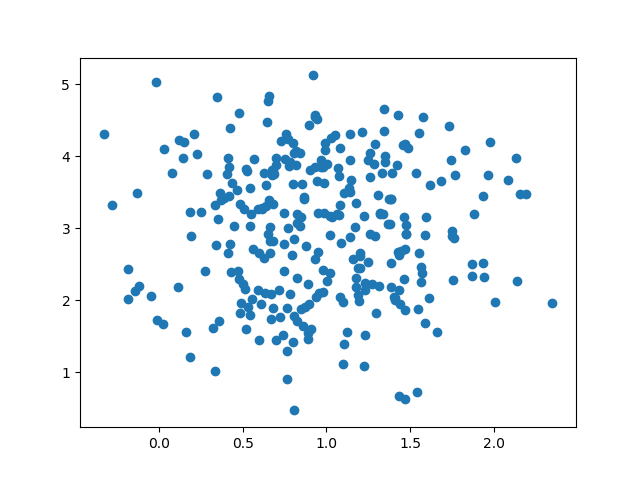
Hình 3. Kết quả của chương trình, chưa tối ưu hoá vòng lặp, t = 287(s)



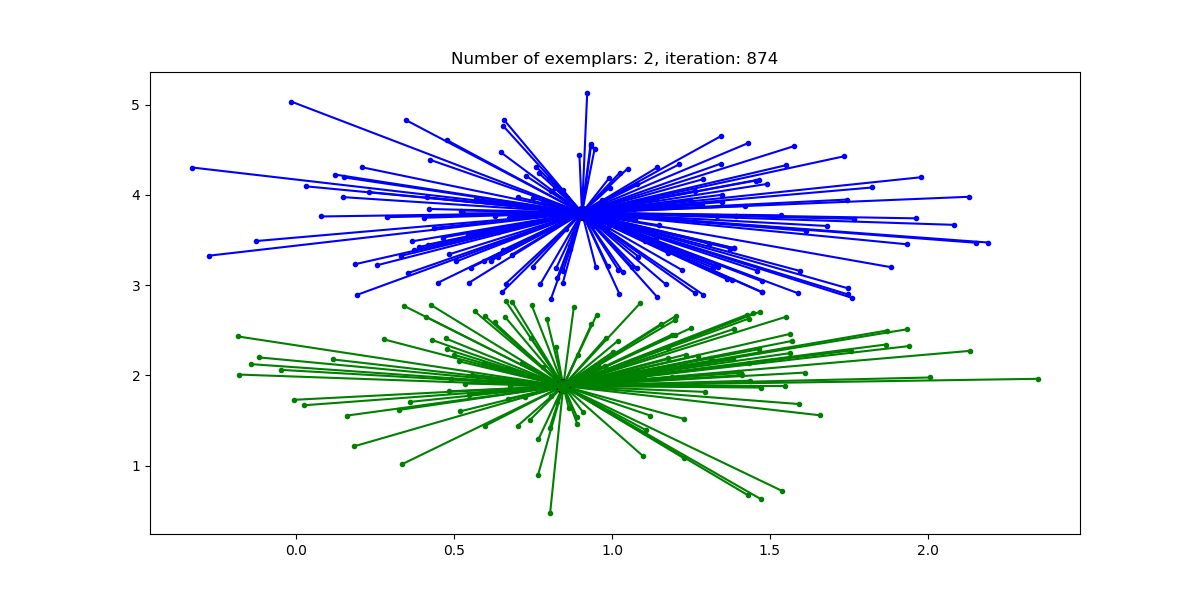
Hình 4. Kết quả của chương trình, sau khi đã tối ưu hoá vòng lặp, t = 7.5(s)

Không chỉ thực hiện với dữ liệu 2 và 3 chiều, mà thuật toán còn có thể chạy cho dữ liệu nhiều chiều với độ chính xác và tốc dộ tương tự các kết quả trên, tuy nhiên việc vẽ đồ thị là bất khả thi nên chúng ta sẽ đi tiếp.

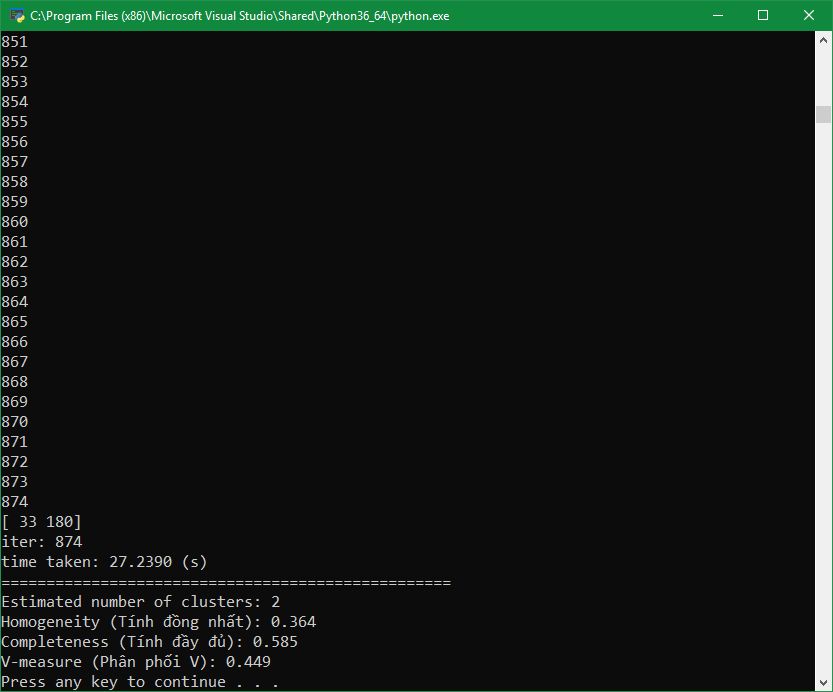
Đối với dữ liệu chồng lấp thì thuật toán sẽ không chạy hiệu quả. Độ hiệu quả của thuật toán sẽ tỉ lệ nghịch với sự chồng lấp của các cụm dữ liệu. Sự chồng lấp càng nhiều thì dộ hiệu quả càng thấp.



Hình 5. Dữ liệu 2 chiều bị chồng lấp, 3 cụm dữ liệu



Hình 6. Thuật toán phân 2 cụm dữ liệu, hội tụ sau 874 vòng lặp



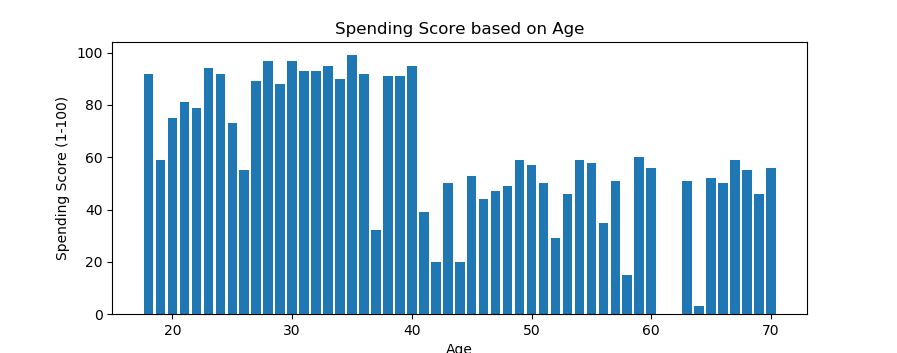
Hình 7. Kết quả của thuật toán, V-measure = 0.449

### Phân cụm khách hàng bằng thuật toán AP

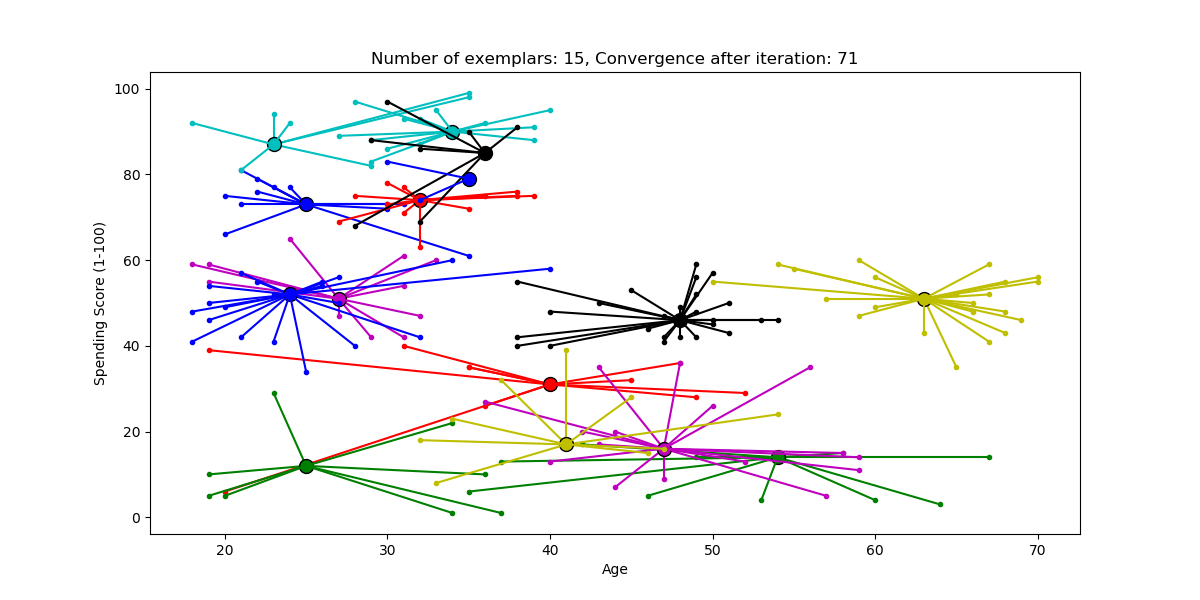
Dataset được sưu tầm trên trang kaggle[11]: Giả sử bạn sở hữu một cửa hàng, thông qua thẻ thành viên của khách hàng, bạn có một vài thông tin của khách hàng như: Customer ID, tuổi, giới tính, thu thập hàng tháng và điểm mua sắm tại cửa hàng. Bạn hãy phân loại khách hàng dựa vào những đặc điểm đó để có thể nắm được nhóm khách hàng hoặc lập kế hoạch chiến lược hợp lí dựa trên nhóm khách hàng đó.

Dataset gồm 200 khách hàng. Do dữ liệu này chưa được gán nhãn nên việc so sánh hiệu suất của thuật toán là điều không thể. Báo cáo này sẽ dựa vào label mà thuật toán học được và so sánh với label mà thuật toán AP của sklearn để so sánh kết quả đầu ra.

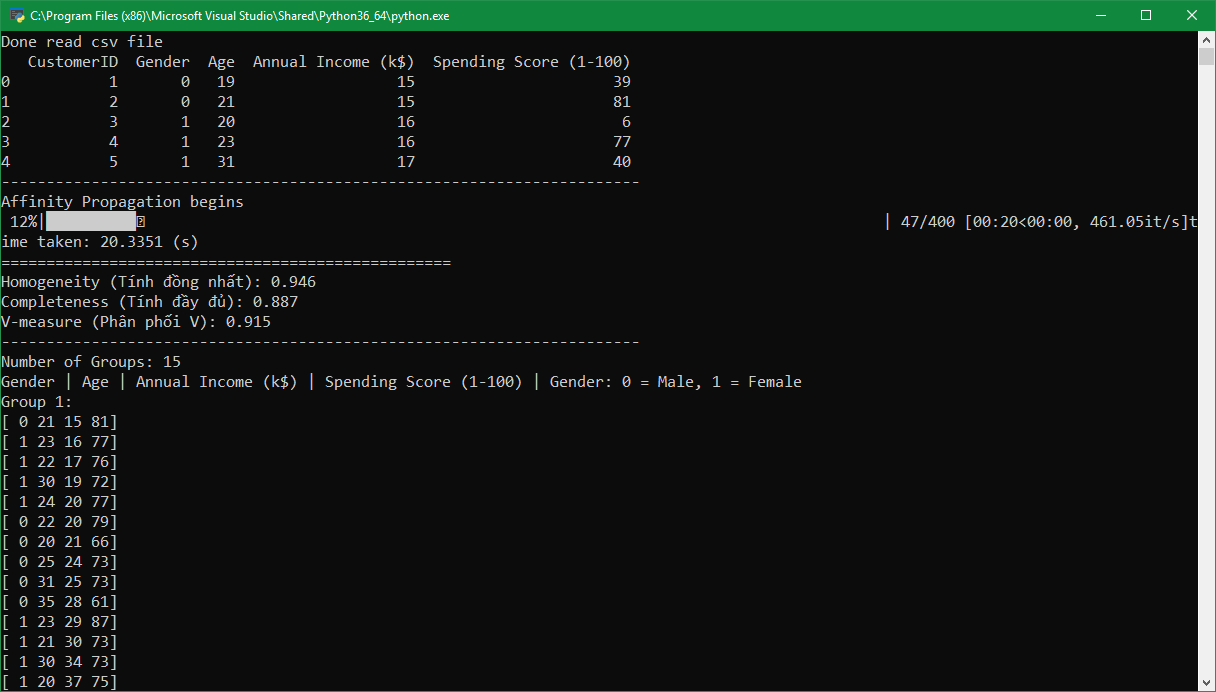




Hình 8. Dataset đưa vào hệ thống



Hình 9. Thuật toán chia làm 15 nhóm, hội tụ sau 71 vòng lặp



Hình 10. Kết quả của thuật toán. V-measure: 0.915

Ta có thể thấy điểm V-measure giữa label thuật toán học được và label của thuật toán AP sklearn là khá cao. Vậy thuật toán chạy đúng như kì vọng.

Bảng 1. Phân công công việc của các thành viên

## Phân công công việc

|  |  |  |
| --- | --- | --- |
| Tên sinh viên | Đóng góp | Mô tả khái quát công việc |
| Phan Thanh Nam | 40% | Tìm kiếm dữ liệu thực tế, góp ý xây dựng thuật toán, tìm kiếm tài liệu tham khảo. |
| Nguyễn Thiên Quốc | 60% | Xây dựng thuật toán, tổ chức code, nghiên cứu các hàm. |

## Kết luận

Chương trình chạy tốt, hoạt động nhanh và ổn định. Khó khăn lớn nhất của nhóm là tìm kiếm tài liệu tham khảo với vốn tiếng anh hạn hẹp. Tuy nhiên nhóm vẫn hoàn thành công việc nhưng khá tốn thời gian.

Đồ án được phát triển và thực hiện sau khi nhóm đã đọc và nghiên cứu kĩ cơ chế hoạt động của thuật toán cũng như học cách sử dụng ngôn ngữ Python để viết thuật toán.

Thuật toán AP sẽ khắc phục được nhược điểm của *k-means* khi phải chỉ ra số lượng cụm và khởi tạo tâm của từng cụm. Tuy nhiên, để AP chạy tốt cũng cần phải có hệ số *preference, damping* được chọn một cách hợp lý.

# TÀI LIỆU THAM KHẢO

1. Emily Listmann. (2007, March 3). How to Avoid Plagiarism. Retrieved from <https://www.wikihow.com/Avoid-Plagiarism>
   1. Media outlets reported allegations Biden plagiarized Kinnock, but not that he had previously credited him. (2008, August 23). Retrieved from <https://www.mediamatters.org/research/2008/08/23/media-outlets-reported-allegations-biden-plagia/144547>
2. Han, J., Kamber, M., & Pei, J. (2011). Data mining concepts and techniques third edition. Morgan Kaufmann.
3. Brendan J. Frey; Delbert Dueck (2007). "Clustering by passing messages between data points". Science. 315 (5814): 972–976.
4. sklearn.cluster.AffinityPropagation — scikit-learn 0.21.1 documentation. (n.d.). Retrieved May 22, 2019, from <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AffinityPropagation.html#sklearn.cluster.AffinityPropagation>
5. Delbert Dueck; Brendan J. Frey (2007). Non-metric affinity propagation for unsupervised image categorization. Int'l Conf. on Computer Vision.
6. FAQ for Affinity Propagation. (n.d.). Retrieved from <http://genes.toronto.edu/affinitypropagation/faq.html>
7. Refianti, Rina & Mutiara, Achmad & Syamsudduha, A.A.. (2016). Performance Evaluation of Affinity Propagation Approaches on Data Clustering. International Journal of Advanced Computer Science and Applications. 7. 10.14569/IJACSA.2016.070357.
8. Vink, R. (2018, May 18). Algorithm breakdown: Affinity propagation. Retrieved from <https://www.ritchievink.com/blog/2018/05/18/algorithm-breakdown-affinity-propagation/>
9. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
10. Rosenberg, A., & Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL).
11. Mall Customer Segmentation Data. (n.d.). Retrieved from <https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python>