## Bài tập về ứng dụng K-means đánh giá sự tương đồng của các giống lúa (dữ liệu của Viện Nghiên cứu Phát triển Đồng bằng sông Cửu Long, Trường Đại học Cần Thơ)

### Báo cáo tóm tắt bài toán

Vùng Đồng bằng sông Cửu Long (ĐBCSL) đang ứng phó với biến đổi khí

hậu, nước biển dâng. Vấn đề cấp bách đặt ra là cần tìm các giải pháp

chọn tạo nhanh và chính xác giống lúa mới, có năng suất, chất lượng cao,

chống chịu các tác nhân sinh học và phi sinh học, đặc biệt là thích ứng

với điều kiện khí hậu cực đoan đang diễn ra phức tạp. Từ năm 1976 đến

nay, Trường Đại học Cần Thơ đã sưu tập và lưu giữ hầu hết các giống

lúa mùa cổ truyền của vùng ĐBSCL. Hiện tại, Viện Nghiên cứu Phát triển

ĐBSCL - Trường Đại học Cần Thơ đã lưu giữ được khoảng 2.000 mẫu

giống lúa. Đây là nguồn tài nguyên gen quý giá phục vụ cho công tác bảo

tồn, khai thác, sử dụng và chọn tạo giống lúa. Tuy nhiên, trong 2.000 mẫu

giống lúa này có nhiều giống tương đồng với nhau do thu thập ở địa

phương khác nhau. Nghiên cứu này ứng dụng các giải thuật gom nhóm

dữ liệu (Clustering) để tạo ra phần mềm hỗ trợ cho các chuyên gia về

giống lúa (i) phát hiện ra các mẫu lúa giống nhau và (ii) đánh giá được

hệ số tương đồng giữa các giống lúa

### Ứng dụng của thuật toán k-means

Trong phiên bản đầu tiên hai giải thuật gom nhóm thông dụng là K-means và CLARA được cài đặt. Ý tưởng chính của thuật toán K-means là tìm cách gom nhóm các đối tượng (objects) đã cho vào k cụm (k là số các cụm được xác định trước, k nguyên dương) sao cho tổng bình phương khoảng cách giữa các đối tượng đến tâm nhóm (centroid) là nhỏ nhất (Macqueen, 1967).

Đối với giả thuật Kmeans hay CLARA, người dùng đều phải xác định số nhóm trước khi tiến hành tìm kiếm. Trong một số trường hợp các nhà nghiên cứu chỉ muốn tìm kiếm các giống lúa dựa vào một số lượng thuộc tính nhất định. Giải thuật xây dựng phần mềm cho phép người sử dụng chọn lựa các thuộc tính cần thiết trong tổng số các thuộc tính đã thu thập để tiến hành gom nhóm. Giá trị của các thuộc tính đều đã lượng hoá thành kiểu số, nên khoảng cách để đo độ tương đồng Euclidean được sử dụng

* Từ tập dữ liệu ban đầu gồm n phần tử và số cụm xác định là k.
* Chọn k đối tượng di (i=1, …, k) làm tâm của k cụm từ tập dữ liệu ban đầu.
* Đối với mỗi đối tượng không phải là tâm, tính khoảng cách (Euclidean, Mahattan, ...) từ nó đến trọng tâm của các cụm còn lại. Xác định trọng tâm gần nhất cho mỗi đối tượng tức là xác định nhóm cho mỗi đối tượng trong tập dữ liệu dựa vào khoảng cách tính được.
* Cập nhật lại trọng tâm cho mỗi cụm bằng cách tính trung bình cộng vector của các đối tượng dữ liệu trong mỗi cụm.
* Lặp lại các bước trên cho đến khi các trọng tâm của cụm không còn thay đổi.

CLARA (Clustering large applications) được phát triển bởi Kaufman vào năm 1990 là thuật toán mở rộng của thuật toán K-means (Kaufman et al., 2005). Thuật toán này nhằm giải quyết trường hợp giá trị của k và n là lớn. CLARA tiến hành trích mẫu cho tập dữ liệu có n phần tử. Thay vì lấy giá trị trung bình của các đối tượng trong một cụm như trong giải thuật K-means, CLARA lấy một đối tượng đại diện trong cụm, gọi là medoid. Nó tìm k cụm trong n đối tượng bằng cách trước tiên tìm một đối tượng đại diện (medoid) cho mỗi cụm. Tập các medoid ban đầu được lựa chọn tuỳ ý. Sau đó, nó lặp lại các thay thế một trong số các medoid bằng một trong số những cái không phải medoid miễn là tổng khoảng cách của kết quả phân cụm được cải thiện.

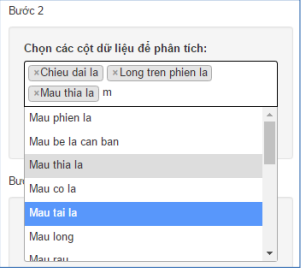
### Kết quả tính toán

Hiện nay, phần mềm này được tích hợp vào ngân hàng thông tin giống lúa vùng ĐBSCL đang triển khai trong mạng nội bộ của Viện Nghiên cứu Phát triển ĐBSCL để gom nhóm các mẫu giống lúa đang được bảo quản tại Viện. Để chạy gom nhóm cho một tập dữ liệu. Chúng ta thực hiện ba bước.

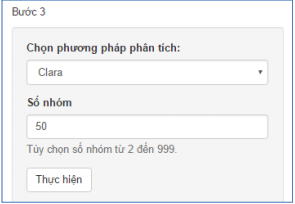
* Bước 1: Duyệt dữ liệu để tải lên web, thời gian tải lên phục thuộc vào kích thước của tập dữ liệu. Với tập dữ liệu bao gồm 1.000 giống lúa, quá trình đọc dữ liệu chỉ mất khoảng 8 giây



* Bước 2: Tìm và chọn các đặc tính bằng cách gõ từ khóa có trong tên đặc tính. Khuyến khích chọn các đặc tính thuộc cùng nhóm (cùng tiêu chí đánh giá) như vậy sẽ giúp việc gom nhóm này tương quan đến các tri thức về nông, sinh học và cho ra kết quả bám sát hơn với kinh nghiệm thực tiễn



* Bước 3: Chọn phương pháp tiến hành phân loại giống lúa (Kmeans và CLARA). Số nhóm cho phân loại sẽ được gợi ý cho người dùng bằng 5% số phần tử có trong dữ liệu (dữ liệu 1.000 giống sẽ là 50 nhóm). Có thể tùy chỉnh số nhóm nhưng phải nằm trong giới hạn từ 2 đến (n-1) nhóm



* Sau khi nhấn nút "Thực hiện" ở bước 3, phần mềm tiến hành gom nhóm giống lúa. Để giúp các nhà nghiên cứu về giống lúa đánh giá kết quả gom nhóm, chúng tôi đã xây dựng chức năng xem thông tin chi tiết các giống lúa trong mỗi nhóm

