

DOI:10.22144/ctu.jsi.2017.019

ỨNG DỤNG GIẢI THUẬT GOM NHÓM DỮ LIỆU ĐỂ NHẬN DIỆN SỰ TƯƠNG ĐỒNG GIỮA CÁC GIỐNG LÚA

Lưu Tiến Đạo¹, Âu Tấn Tài², Vũ Anh Pháp³ và Trần Nguyễn Minh Thư²

¹Trung tâm Công nghệ Phần mềm, Trường Đại học Cần Thơ

²Khoa Công nghệ Thông tin và Truyền thông, Trường Đại học Cần Thơ

³Viện Nghiên cứu Phát triển Đồng bằng sông Cửu Long, Trường Đại học Cần Thơ

Thông tin chung:

Ngày nhận bài: 15/09/2017

Ngày nhận bài sửa: 10/10/2017

Ngày duyệt đăng: 20/10/2017

Title:

Applying clustering techniques for identifying similarities among rice varieties

Từ khóa:

Gom cụm dữ liệu, giống lúa, khai khoáng dữ liệu

Keywords:

Data mining, clustering, rice varieties

ABSTRACT

The Mekong Delta in southern Vietnam is facing climate change and sea level rise. A solution is to quickly and accurately create new high-quality rice varieties that boost yield and adapt well to biological and non-biological factors, especially well-adapt to current harsh conditions. Since 1976, Can Tho University has collected and stored most of traditional seasonal rice varieties of the Mekong Delta. At the moment, Mekong Delta Development and Research Institute of Can Tho University has stored more than 2,000 rice variety samples. They are valuable gene resources that can be used for preserving, exploiting, employing, and creating rice varieties. However, it is possible that there are similarities in these 2,000 samples for some rice varieties. In this paper, clustering techniques are used to create tools for rice variety experts to (i) identify similar samples and (ii) analyze their similarity coefficients.

TÓM TẮT

Vùng Đồng bằng sông Cửu Long (ĐBSCL) 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.

Trích dẫn: Lưu Tiến Đạo, Âu Tấn Tài, Vũ Anh Pháp và Trần Nguyễn Minh Thư, 2017. Ứng dụng giải thuật gom nhóm dữ liệu để nhận diện sự tương đồng giữa các giống lúa. Tạp chí Khoa học Trường Đại học Cần Thơ. Số chuyên đề: Công nghệ thông tin: 140-144.

1 GIỚI THIỆU

Việt Nam phát triển từ nền văn hóa lúa nước nên nguồn gen cây lúa rất đa dạng, phong phú. Từ những năm 1910, Việt Nam đã tiên phong trong bảo tồn các giống lúa địa phương. Trung tâm Thí nghiệm Lúa Cần Thơ, thành lập năm 1913 trực thuộc Cục Túc Mễ Đông Dương, đã sưu tập bảo tồn 800 giống lúa cổ truyền qua chương trình tuyển chọn giống lúa của Trung tâm. Ngay sau khi đất nước thống nhất, Đại học Cần Thơ từ năm 1976 đến nay đã sưu tập và lưu giữ hầu hết các giống lúa mùa cổ truyền của vùng ĐBSCL. Số lượng cụ thể là hơn 2.000 mẫu. Đây là nguồn gen quý giá phục vụ cho công tác chọn tạo giống.

Việc phân tích quan hệ di truyền không chỉ có ý nghĩa trong việc quản lý, bảo tồn các giống cây trồng bản địa mà còn có ý nghĩa trong công tác lai tạo giống chất lượng cao (Vũ Thị Thu Hiền, 2012) (Đoàn Thị Thùy Linh và Nguyễn Văn Khoa, 2013) (Đoàn Thanh Quỳnh và *ctv.*, 2016) (Trần Thị Lương và *ctv.*, 2013). Tác giả Vũ Thị Thu Hiền đã dùng phần mềm Excel cùng với phần mềm NTSYSpc để phân tích, đánh giá sự đa dạng di truyền của 41 giống lúa có nguồn gốc khác nhau dựa trên 14 tính trạng kiểu hình (Vũ Thị Thu Hiền, 2012). Bốn mươi một giống lúa này được phân thành 10 nhóm cách biệt di truyền với sự sai khác 0,08. Tác giả Đoàn Thị Thùy Linh và Nguyễn Văn Khoa (2013) đã sử dụng các đặc trưng cơ bản của cây lúa như chiều cao cây, số nhánh hữu hiệu, chiều dài và chiều rộng lá đòng, chiều dài bông, chiều dài và chiều rộng hạt gạo, khối lượng 1.000 hạt... để đánh giá sự khác biệt về hình thái của 50 mẫu giống lúa địa phương vùng Tây Bắc được thu thập và trồng khảo nghiệm tại huyện Thuận Châu, tỉnh Sơn La. Các số liệu phân tích thống kê bằng phần mềm Excel kết hợp với phần mềm NTSYSpc. Không chỉ ở cây lúa, việc phân tích đa dạng di truyền còn diễn ra ở nhiều loại cây trồng khác. Có thể liệt kê những công trình nghiên cứu gần đây ở Việt Nam như phân tích đa dạng di truyền của các mẫu giống đậu cô ve (Phạm Thị Ngọc và *ctv.*, 2016), ngô (Lê Thị Minh Thảo và *ctv.*, 2014) hay đậu nành rau (Nguyễn Lộc Hiền và *ctv.*, 2010). Trong nghiên cứu của Nguyễn Lộc Hiền và *ctv.* (2010), tác giả đã dùng phần mềm Statistica 5.0 để phân nhóm 22 giống đậu nành rau dựa trên 15 tính trạng hình thái - nông học.

Nhìn chung, phần lớn các nghiên cứu về đa dạng di truyền sử dụng phần mềm có tính phí (Excel, NTSYSpc (Rohlf, 1998) hoặc Statistica (Nisbet *et al.*, 2009)) và cần cài đặt trên từng máy. Điều này gây khó khăn cho các nhà nghiên cứu và đặc biệt là sinh viên. Vì vậy, trong bài báo này, chúng tôi trình bày phần mềm hỗ trợ các chuyên gia đánh giá hệ số tương đồng giữa các giống lúa. Phần mềm này được

thiết kế với mục đích cung cấp công cụ hỗ trợ gom nhóm các giống lúa dựa vào nhiều đặc điểm (hơn 60 thuộc tính cho mỗi giống) với số lượng giống khoảng vài ngàn và dễ sử dụng, phù hợp cho nhiều đối tượng, từ nhà nghiên cứu đến sinh viên ngành nông học. Phần mềm cho phép sử dụng ngay trên trình duyệt web, không cần tải và cài đặt phần mềm. Phần mềm được cài đặt các giải thuật gom nhóm và có các công cụ hỗ trợ nhà nghiên cứu đánh giá và phân tích kết quả gom nhóm. Phần mềm được thiết kế tuân theo chuẩn HTML5 và CSS3. Do đó, nó có khả năng tương thích với mọi thiết bị ở tất cả kích thước màn hình khác nhau. Người dùng có thể sử dụng phần mềm trên mọi thiết bị mà không lo lắng về độ phân giải hay sự co giãn giao diện bởi chúng hoàn toàn giống như đang truy cập web trên màn hình máy tính cá nhân.

2 GOM NHÓM GIỐNG LÚA

2.1 Hệ thống tiêu chuẩn đánh giá cây lúa

Năm 1996, IRRI đã ban hành phiên bản thứ tư của hệ thống tiêu chuẩn đánh giá cây lúa (Standard Evaluation System for Rice - SES) (IRRI, 1996). Hệ thống tiêu chuẩn này giúp các nhà nghiên cứu lúa trên thế giới có một tiếng nói chung trong công tác đánh giá đặc tính của cây lúa, tạo điều kiện thuận lợi cho việc thu thập, xử lý và phân tích các số liệu. Ví dụ hình dạng hạt gạo được xác định dựa trên tỷ số chiều dài/chiều rộng sau khi bóc vỏ trấu, trước khi xát.

Bảng 1: Phân loại hình dạng hạt gạo theo IRRI

Thang điểm	Tỷ lệ dài/rộng	Hình dạng
1	>3.0	Thon dài
3	2.1 - 3.0	Trung bình
5	1.1 - 2.0	Bầu
7	< 1.1	Tròn

Bảng 2: Đánh giá mùi thơm theo IRRI

Cấp	Mùi thơm
0	Không thơm
1	Hơi thơm
2	Thơm

Dựa trên hệ thống tiêu chuẩn đánh giá cây lúa của IRRI, Viện Nghiên cứu Phát triển ĐBSCL Trường Đại học Cần Thơ đã sưu tập và lưu trữ hơn 2.000 mẫu giống lúa. Mỗi giống lúa có hơn 60 thuộc tính như độ cứng của thân, chiều dài bông, độ ngập sâu trong năm, chiều dài hạt lúa, màu gạo lúc, mùi thơm, dạng gạo,.... Các đặc tính này thuộc 5 nhóm: sinh thái địa lý, đặc tính sinh lý, điều kiện môi trường canh tác, đặc tính sinh hoá hạt gạo và đặc tính hình thái. Mỗi đặc tính đều đã được số hoá nên việc tính toán độ tương đồng giữa các phần tử bằng các giải thuật gom nhóm được thực hiện dễ dàng. Ví

đụ đối với mùi thơm, IRRI chia làm 3 cấp độ theo Bảng 2.

2.2 Các giải thuật gom nhóm

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 giải 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 d_i ($i=1, \dots, k$) 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 tùy 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.

3 THIẾT KẾ PHẦN MỀM

3.1 Thiết kế tổng thể

Phần mềm có các thành phần như sau (Hình 1):

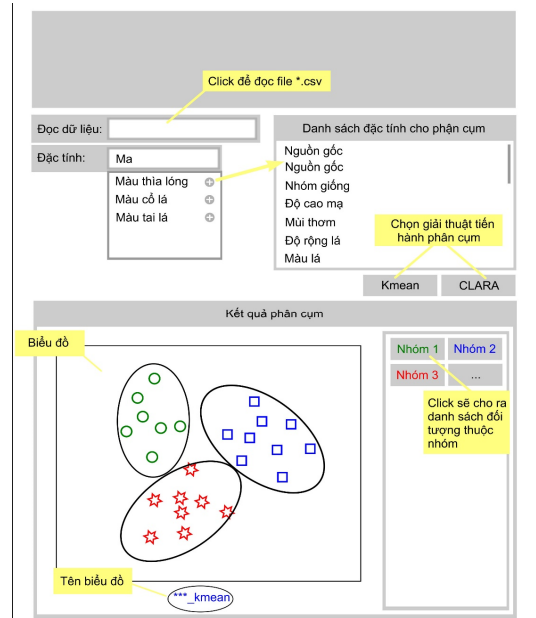
- Khung đọc dữ liệu: Dùng để nạp file dữ liệu với định dạng *.csv từ máy tính. Dữ liệu phải được xử lý theo mẫu định sẵn (Cột đầu tiên là tên các giống lúa, hàng đầu tiên là tên các thuộc tính).

- Khung đặc tính: Các đặc tính được truy xuất bằng cách lấy dữ liệu từ dòng thứ nhất của dữ liệu. Người dùng có thể tùy chọn các thuộc tính tùy vào mục đích nghiên cứu để tiến hành gom nhóm dữ liệu.

- Khung danh sách đặc tính: Hiện thị các đặc tính được chọn từ khung đặc tính, sẵn sàng cho bước thực hiện giải thuật gom nhóm.

- Khung lựa chọn giải thuật: Có thể tùy chọn 1 trong 2 cách gom nhóm: K-means và CLARA. Người dùng sẽ phải nhập số nhóm trước khi tiến hành gom nhóm dữ liệu.

- Khung kết quả gom nhóm: Hiện thị biểu đồ kết quả gom nhóm các giống lúa. Các nút chứa thông tin chi tiết của các giống lúa thuộc vào từng nhóm (nhấn vào để hiển thị).



Hình 1: Mô hình thiết kế tổng thể

3.2 Phương thức hoạt động

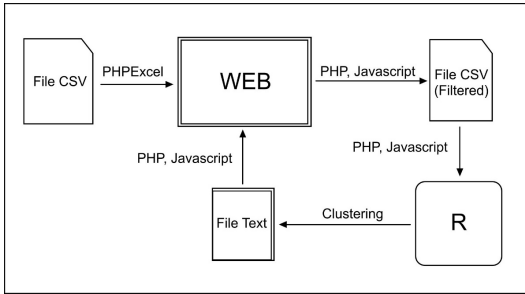
Phần mềm bao gồm 2 quá trình chính:

- Quá trình nạp dữ liệu:

Chương trình nạp dữ liệu đầu vào từ một file csv. Dữ liệu được duyệt qua, thông tin dòng đầu tiên (đặc tính) của bảng dữ liệu được trích xuất và hiển thị ra

dạng danh sách các đặc tính giống lúa để người dùng có thể chọn lọc ra tùy vào mục đích nghiên cứu.

Thông tin các đặc tính được chọn sẽ được đẩy vào một mảng để tiến hành tạo ra một file csv mới (chỉ bao gồm các cột được chọn). Các quá trình này được hỗ trợ bởi thư viện PHPExcel giúp đọc, ghi cũng như xuất các tập tin có định dạng *.xls, *.xlsx, *.csv,... Đồng thời, chương trình còn sử dụng kỹ thuật Ajax trên thư viện JQuery giúp câu lệnh không còn rườm rà nhưng vẫn hoạt động như mong muốn.



Hình 2: Phương thức hoạt động của phần mềm

– Quá trình gom nhóm:

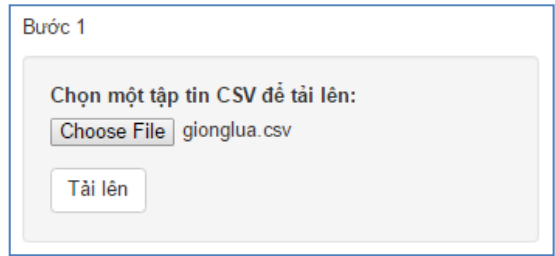
Tập tin dữ liệu được xuất ra từ bước trên được truyền cho R xử lý bằng giải thuật được chọn. Kết quả của quá trình gom nhóm, được xuất ra một tập tin ảnh png và tập tin text (txt). Tập tin ảnh là biểu đồ của kết quả gom nhóm được hiển thị trực tiếp trên nền web, tập tin text được tiếp tục xử lý bằng PHP, từng phần tử sẽ được đưa vào nhóm của nó (mảng). Quá trình được xác định bởi các chỉ số từ kết quả xuất ra bởi R (số nhóm được tạo ra dựa trên số nhóm người dùng nhập vào, truyền tham số truyền đến biến k trong mã lệnh R để thực thi giải thuật). Các phần tử thuộc mỗi nhóm sẽ tiếp tục được duyệt qua với tất cả phần tử, lọc ra thông tin từ các cột đặc tính còn lại và hiển thị thành bảng chứa đầy đủ thông tin các giống lúa đã chọn ở bước 2. Việc xét để lấy thông tin các giống lúa sẽ chạy số vòng lặp tương ứng với số các phần tử có trong mỗi nhóm, thay vì phải chạy qua hết các phần tử (độ phức tạp sẽ cao hơn).

4 KẾT QUẢ

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.

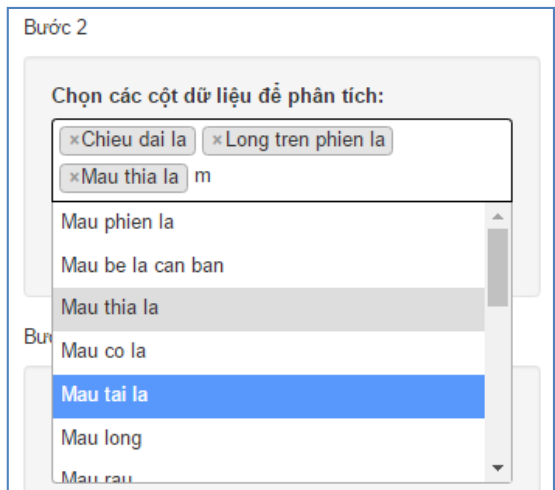
Để chạy gom nhóm cho một tập dữ liệu. Chúng ta thực hiện ba bước.

Bước 1: Duyệt dữ liệu để tải lên web (Hình 3), 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.



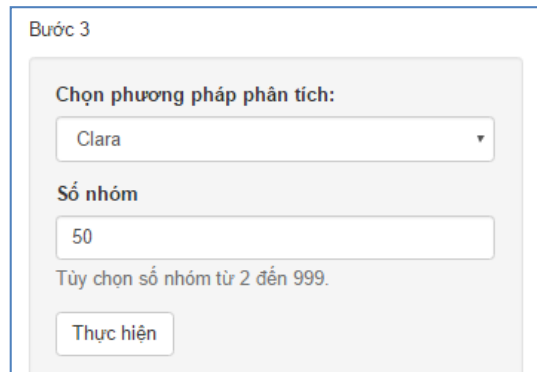
Hình 3: Nạp dữ liệu để gom nhóm

Bước 2: Tìm và chọn các đặc tính bằng cách gõ từ khóa có trong tên đặc tính (Hình 4). 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.



Hình 4: Tùy chọn các đặc tính tiến hành gom nhóm

Bước 3: Chọn phương pháp tiến hành phân loại giống lúa (Kmeans và CLARA) (Hình 5). 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.



Hình 5: Chọn phương pháp gom nhóm và số nhóm mong muốn

Sau khi nhấn nút "Thực hiện" ở bước 3, phần mềm tiến hành gom nhóm giống lúa và cho kết quả tương tự như Hình 6. 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 (Hình 7).

Nhóm 1	Nhóm 2	Nhóm 3	Nhóm 4	Nhóm 5	Nhóm 6	Nhóm 7	Nhóm 8	Nhóm 9	Nhóm 10	Nhóm 11	Nhóm 12	Nhóm 13
Nhóm 14	Nhóm 15	Nhóm 16	Nhóm 17	Nhóm 18	Nhóm 19	Nhóm 20	Nhóm 21	Nhóm 22	Nhóm 23	Nhóm 24	Nhóm 25	
Nhóm 26	Nhóm 27	Nhóm 28	Nhóm 29	Nhóm 30	Nhóm 31	Nhóm 32	Nhóm 33	Nhóm 34	Nhóm 35	Nhóm 36	Nhóm 37	
Nhóm 38	Nhóm 39	Nhóm 40	Nhóm 41	Nhóm 42	Nhóm 43	Nhóm 44	Nhóm 45	Nhóm 46	Nhóm 47	Nhóm 48	Nhóm 49	
Nhóm 50												

Hình 6: Kết quả gom nhóm

Nhóm 3 (Số lượng: 20)				
ID	Chiều dài lá	Long trên phiến lá	Màu thía lá	Màu tai lá
3	44	2	1	2
28	44	2	0.001	0.001
83	44	2	1	0.001
245	44	2	1	0.001
246	44	2	1	0.001
338	44	2	1	0.001
438	44	2	1	0.001
459	44	2	2	0.001
477	44	2	1	0.001
513	44	2	1	0.001
555	44	2	1	0.001
627	44	2	1	0.001
649	44	2	1	0.001
676	44	2	1	0.001
747	44	1	1	0.001
760	44	2	1	0.001
803	44	2	1	0.001
813	44	2	1	2
827	44	2	1	0.001
986	44	2	1	0.001

Hình 7: Thông tin chi tiết các giống lúa trong mỗi nhóm

5 KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Bài báo đã đề xuất một mô hình đánh giá sự tương đồng giữa các giống lúa dựa trên phương pháp gom nhóm dữ liệu. Chúng tôi tiến hành xây dựng phần mềm với 2 giải thuật K-means và CLARA. Phần mềm được xây dựng theo tiêu chí dễ sử dụng, có các công cụ trực quan hỗ trợ các nhà nghiên cứu đánh giá kết quả gom nhóm. Bên cạnh đó, người dùng cũng được cung cấp chức năng cho phép chọn lựa đặc tính để gom nhóm. Chúng tôi sẽ đưa thêm vào phần mềm nhiều giải thuật gom nhóm, ví dụ như các giải thuật cây phân cấp.

LỜI CẢM ƠN

Bài báo này được thực hiện trong khuôn khổ đề tài nghiên cứu khoa học cấp trường, mã số đề tài: T2016-101. Các tác giả chân thành cảm ơn Trường Đại học Cần Thơ, Phòng Quản lý khoa học đã hỗ trợ để chúng tôi có thể thực hiện thành công đề tài.

TÀI LIỆU THAM KHẢO

Đoàn Thanh Quỳnh, Nguyễn Thị Hào, Vũ Thị Thu Hiền và Trần Văn Quang, 2016. Đánh giá đa dạng di truyền nguồn gen lúa nếp địa phương dựa trên kiểu hình và chỉ thị phân tử. Tạp chí Khoa học Nông nghiệp Việt Nam. Tập 14, số 4: 527–538.

Đoàn Thị Thùy Linh và Nguyễn Văn Khoa, 2013. Đa dạng di truyền một số mẫu giống lúa địa phương vùng Tây Bắc dựa trên đặc điểm hình thái. Hội nghị khoa học toàn quốc về sinh thái và tài nguyên sinh vật lần thứ 5, 18/10/2013, Hà Nội, Việt Nam. Nhà xuất bản Nông nghiệp. Hà Nội, 1132–1139.

Exeter Software. NTSYSpc, Numerical Taxonomy System, truy cập ngày 09/10/2017. Địa chỉ: <http://www.exetersoftware.com/cat/ntsyspc/ntsyspc.html>

IRRI, 1996. Standard Evaluation System for Rice. Genetic Resources Center. International Rice Research Institute. Philippines.

Kaufman, L., Rousseeuw, P.J., 2005. Finding groups in data : an introduction to cluster analysis, Wiley.

Lê Thị Minh Thảo, Nguyễn Thị Ánh, Trần Thanh Tân, Phạm Quang Tuấn và Vũ Văn Liết, 2014. Phân tích đa dạng di truyền dựa trên kiểu hình và chỉ thị phân tử SSR và đánh giá khả năng chịu hạn của các dòng ngô nếp tự phối - phục vụ phát triển ngô nếp cho các tỉnh miền núi phía Bắc. Tạp chí Khoa học và Phát triển. Tập 12, số 3: 285–297.

Macqueen, J.B., 1967. Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Math, Statistics, and Probability. University of California Press, pp. 281–297.

Nguyễn Lộc Hiền, Trần Thanh Xuyên, Trần Thị Bích Phương và Tadashi Yoshihashi, 2010. Sự đa dạng di truyền của các giống đậu nành rau Nhật Bản. Tạp chí Khoa học Trường Đại học Cần Thơ. 16a: 51–59.

Nisbet, R., Elder, J., and Miner, G, 2009. Handbook of Statistical Analysis and Data Mining Applications. Burlington, MA: Academic Press (Elsevier).

Phạm Thị Ngọc, Nguyễn Quốc Trung, Vũ Văn Liết, 2016. Phân tích đa dạng di truyền của các mẫu giống đậu cô ve bằng chỉ thị hình thái và chỉ thị phân tử SSR. Tạp chí Khoa học Nông nghiệp Việt Nam. Tập 14, số 12: 1874–1885.

Rohlf, F.J., 1998. NTSYS-pc: numerical taxonomy and multivariate analysis system, version 2.02e. Setauket: Applied Biostatistics Inc., Exeter Software.

Trần Thị Lương, Lưu Minh Cúc và Nguyễn Đức Thành, 2013. Phân tích quan hệ di truyền của một số giống lúa đặc sản, chất lượng, trồng phổ biến ở Việt Nam bằng chỉ thị phân tử SSR. Tạp chí Sinh học. 35(3): 348–356.

Vũ Thị Thu Hiền, 2012. Đa dạng di truyền dựa trên đặc điểm hình thái của các mẫu giống lúa có nguồn gốc khác nhau. Tạp chí Khoa học và Phát triển. Tập 10, số 6: 844–852.