CẢI TIẾN THUẬT TOÁN K-MEANS VÀ ỨNG DỤNG HỖ TRỢ SINH VIÊN CHỌN CHUYÊN NGÀNH THEO HỌC CHẾ TÍN CHỈ

Nguyễn Văn Lễ*, Mạnh Thiên Lý
Nguyễn Thị Định, Nguyễn Thị Thanh Thủy
Trường Đại học Công nghiệp Thực phẩm TP.HCM
*Email: *lecntp@gmail.com*
Ngày nhận bài: 27/4/2018; Ngày chấp nhận đăng: 05/6/2018

TÓM TẮT

K-Means là thuật toán được ứng dụng rất hiệu quả trong nhiều bài toán phân cụm dữ liệu. Nhóm tác giả áp dụng thuật toán này để phân cụm chuyên ngành trên tập dữ liệu điểm số, tuy nhiên thuật toán kém hiệu quả trong một số trường hợp nên độ chính xác không cao. Vì vậy, trong bài báo này, nhóm tác giả đề xuất phương pháp phân cụm trên tập dữ liệu nhóm điểm đặc trưng cho mỗi chuyên ngành. Ngoài ra, cải tiến thuật toán K-Means để loại bỏ phần tử nhiễu nhằm giảm thời gian tính toán của thuật toán. Kết quả phân cụm sẽ hỗ trợ sinh viên Khoa Công nghệ Thông tin, Trường Đại học Công nghiệp thực phẩm Thành phố Hồ Chí Minh lựa chọn chuyên ngành phù hợp.

Từ khóa: K-Means, phân cụm dữ liệu, chọn chuyên ngành, khoảng cách Euclid, trọng tâm.

1. GIỚI THIỆU

1.1. Phân cụm

Phân cụm dữ liệu là phương pháp xử lý thông tin nhằm khám phá mối liên hệ giữa các mẫu dữ liệu bằng cách tổ chức chúng thành các cụm tương tự. Tất cả các dạng dữ liệu được biểu diễn bởi các đặc trưng đó là vectơ n-chiều. Để phân cụm dữ liệu cần thực hiện các bước cơ bản sau:

Chọn đặc trưng: Các đặc trưng lựa chọn phải hợp lý để có thể mã hoá nhiều nhất các thông tin liên quan đến công việc quan tâm.

Chọn độ đo gần nhất: Một độ đo chỉ ra mức độ tương tự hay không tương tự giữa hai vectơ đặc trưng.

Tiêu chuẩn phân cụm: Tiêu chuẩn phân cụm có thể được biểu diễn bởi hàm chi phí hoặc một vài quy tắc khác.

Công nhận kết quả: Sau khi có kết quả phân cụm, cần kiểm tra tính đúng đắn của nó.

Giải thích kết quả: Bằng kết quả thực nghiệm cần phân tích để đưa ra kết luận đúng đắn.

Một số phương pháp phân cụm điển hình: Phân cụm phân hoạch, phân cụm phân cấp, phân cụm dựa trên mật độ, phân cụm dựa trên lưới, phân cụm dựa trên mô hình, phân cụm có ràng buộc. Tác giả chọn phương pháp phân cụm phân hoạch nhằm gom cụm các sinh viên theo chuyên ngành phù hợp dựa trên điểm số một số học phần đã tích lũy được trong thời gian học tập.

1.2. Thuật toán K-Means

1.2.1. Giới thiệu

K-Means là thuật toán rất quan trọng và được sử dụng phổ biến trong kỹ thuật phân cụm dữ liệu [1]. Thuật toán này tìm cách phân cụm các đối tượng đã cho vào k cụm (k là số cụm được xác định trước, k nguyên dương) sao cho tổng bình phương khoảng cách giữa các đối tượng đến tâm cụm (centroid) là nhỏ nhất. Về nguyên lý, có n đối tượng, mỗi đối tượng có m thuộc tính, các đối tượng được phân chia thành k cụm dựa trên các thuộc tính của đối tượng bằng việc áp dụng thuật toán K-means. Bài toán này xem mỗi thuộc tính của đối tượng (đối tượng có m thuộc tính) như một tọa độ của không gian m chiều và biểu diễn đối tượng như một điểm trong không gian m chiều, đó là:

$$a_i = (x_{i1}, x_{i2}, \dots, x_{im}) \qquad (1)$$

Trong đó: a_i $(i = 1..n)$: Đối tượng thứ i

x_{ij} $(i = 1..n, j = 1..m)$: Thuộc tính thứ j của đối tượng i.

1.2.2. Khoảng cách Euclid

Phương pháp phân cụm dữ liệu thực hiện dựa trên khoảng cách Euclid là khoảng cách nhỏ nhất từ đối tượng đến phần tử trọng tâm của các cụm. Phần tử trọng tâm của cụm được xác định bằng giá trị trung bình các phần tử trong cụm.

$a_i = (x_{i1}, x_{i2}, \dots, x_{im}); i = 1..n$ là đối tượng thứ i cần phân cụm

$c_j = (x_{j1}, x_{j2}, \dots, x_{jm}); j = 1..k$ là phần tử trọng tâm cụm j

Khoảng cách Euclid từ đối tượng a_i đến phần tử trọng tâm nhóm j; c_j được tính toán dựa trên công thức:

$$\partial_{ij} = \sqrt{\sum_{s=1}^{m}(x_{is} - x_{js})^2} \qquad (2)$$

Trong đó: ∂_{ij}: Khoảng cách Euclid từ a_i đến c_j

x_{is}: Thuộc tính thứ s của đối tượng a_i

x_{js}: Thuộc tính thứ s của phần tử trọng tâm c_j

1.2.3. Phần tử trọng tâm

K phần tử trọng tâm ban đầu được chọn ngẫu nhiên, sau mỗi lần gom các đối tượng vào các cụm, phần tử trọng tâm được tính toán lại:

$Cluster_i = \{a_1, a_2, \dots, a_t\}$ – cụm thứ i;

$i = 1..k$; k số cluster

$j = 1..m$; m số thuộc tính

t: Số phần tử hiện có của nhóm thứ i

x_{sj}: Thuộc tính thứ j của phần tử s; $s = 1..t$

c_{ij}: Toạ độ thứ j của phần tử trung tâm cụm i;

$$c_{ij} = \frac{\sum_{s=1}^{t} x_{sj}}{t} \qquad (3)$$

Thuật toán K-Means [2-4]

Input: Số cụm k và các trọng tâm cụm $\{m_j\}$; $k_j = 1$

Output: Các cụm $C[i]$ $(1 \leq i \leq k)$ và hàm tiêu chuẩn E đạt giá trị tối thiểu.
Begin

Bước 1: Khởi tạo

Chọn k trọng tâm $\{m_j\}$ $(1 \leq j \leq k)$, ban đầu trong không gian R_d (d là số chiều của dữ liệu). Việc lựa chọn này có thể là ngẫu nhiên hoặc theo kinh nghiệm.

Bước 2: Tính khoảng cách

Đối với mỗi điểm X_i $(1 \leq i \leq n)$, tính khoảng cách của nó tới mỗi trọng tâm $\{m_j\}$ $(1 \leq j \leq k)$. Sau đó tìm trọng tâm gần nhất đối với mỗi điểm.

Bước 3: Cập nhật lại trọng tâm

Đối với mỗi $1 \leq j \leq k$, cập nhật trọng tâm cụm m_j bằng cách xác định trung bình cộng các vectơ đối tượng dữ liệu.

Điều kiện dừng: Lặp lại các bước 2 và 3 cho đến khi các trọng tâm của cụm không thay đổi.

End

2. GIẢI PHÁP PHÂN CỤM CHUYÊN NGÀNH

2.1. Chuyên ngành và chọn chuyên ngành

Chương trình đào tạo hệ đại học chính quy ngành Công nghệ thông tin được chia làm 4 chuyên ngành gồm: Hệ thống thông tin, Công nghệ phần mềm, Mạng máy tính và Thương mại điện tử [5]. Mỗi chuyên ngành gồm những học phần chuyên sâu thể hiện khối kiến thức đặc thù của chuyên ngành đó. Hàng năm, Khoa Công nghệ Thông tin thường tổ chức buổi giới thiệu chuyên ngành cho sinh viên năm thứ 3. Qua đó, sinh viên sẽ xác định được chuyên ngành nào phù hợp và tiến hành đăng ký môn học theo chuyên ngành đã chọn. Tuy nhiên, có thể thấy việc lựa chọn chuyên ngành như trên phần lớn là cảm tính, theo sở thích của sinh viên mà chưa có căn cứ cụ thể dẫn đến việc chọn chuyên ngành không phù hợp gây ảnh hưởng lớn đến kết quả học tập của sinh viên. Nhóm tác giả đưa ra giải pháp và thực nghiệm giải quyết việc định hướng chuyên ngành cho sinh viên một cách tự động căn cứ vào kết quả học tập những học phần có kiến thức bổ trợ cho từng chuyên ngành. Các học phần này được chọn theo từng chuyên ngành như sau:

Chuyên ngành Hệ thống thông tin gồm: Cơ sở dữ liệu, Thực hành cơ sở dữ liệu, Hệ quản trị cơ sở dữ liệu, Thực hành hệ quản trị cơ sở dữ liệu, Phân tích thiết kế hệ thống thông tin, Thực hành phân tích thiết kế hệ thống thông tin.

Chuyên ngành Công nghệ phần mềm gồm: Ngôn ngữ lập trình, Thực hành ngôn ngữ lập trình, Cấu trúc dữ liệu và giải thuật, Thực hành cấu trúc dữ liệu và giải thuật, Lập trình hướng đối tượng, Thực hành lập trình hướng đối tượng.

Chuyên ngành Mạng máy tính và truyền thông gồm: Kiến trúc máy tính, Hệ điều hành, Mạng máy tính, Thực hành mạng máy tính, Quản trị mạng, Thực hành quản trị mạng.

Chuyên ngành Thương mại điện tử gồm: Thiết kế web, Thực hành thiết kế web, Cơ sở dữ liệu, Thực hành cơ sở dữ liệu, Đồ họa máy tính, Thực hành đồ họa máy tính, Thương mại điện tử ngành CNTT.

2.2. Tiền xử lý dữ liệu

Dữ liệu thu thập ban đầu là các tập tin excel chứa thông tin điểm học tập của sinh viên đại học khóa 05 ngành Công nghệ thông tin. Mỗi tập tin chứa thông tin điểm số các môn trong một học kỳ được tổ chức thành bảng có nhiều cột và dòng, trong đó mỗi cột là một môn học, mỗi dòng là kết quả học tập của một sinh viên trong học kỳ đó.

Bảng 1. Bảng điểm tổng kết học kỳ của sinh viên [7]

STT	Mã sinh viên	Họ đệm	Tên	An toàn lao động	Cơ sở dữ liệu	Hàm phức và phép biến đổi	Hệ điều hành	Lập trình hướng đối tượng	Quy hoạch tuyến tính	Thiết kế Web	Thực hành cơ sở dữ liệu	Thực hành lập trình hướng đối tượng	Thực hành thiết kế Web	Toán kỹ thuật	Tư tưởng Hồ Chí Minh	Vật lý địa cương 3	Xác suất thống kê	Rèn luyện		Học lực				Ghi chú
				2	3	2	3	3	2	2	1	1	1	2				Điểm	Xếp loại	Điểm 10	Điểm 4	Điểm chữ	Xếp loại	
1	2001140003	Đặng Quốc	An	4.3	6.1	4.8	5.7			7.7				8			6.2	80	Tốt	5.44	1.71	D+	Trung bình yếu	Học tiếp
2	2001140009	Đồng Nguyễn Hoàng	Anh	4.6	4.1		6	5.5			6.3	6.5		4.5			5	87	Tốt	4.95	1.36	D	Yếu	Học tiếp
3	2001140006	Trần Ngọc	Anh	6.1	6.8		6.9	6.8	7.5	8	8.3	8	7.5	6.1				96	Xuất sắc	6.72	2.46	C+	Trung bình	Học tiếp
4	2001140014	Phạm Anh	Bằng	0.1	0		0	0		0				0				44	Yếu	0.01	0	F	Kém	Học tiếp
5	2001140015	Nguyễn Chí	Bảo	6	4.2	6.7	5.5			6.3	7.5	5.5	7.5	5.8				87	Tốt	5.79	1.95	C	Trung bình yếu	Học tiếp
6	2001140020	Trần Tuấn	Bảo							3.7			7.3		3.9	0	3.2	48	Yếu	2.25	0.15	F	Kém	Cảnh cáo h
7	2001140333	Trương Huỳnh Gia	Bảo	7.5	7.5		8.4	9.3		8.7	10	9.5	7.8		8.2			90	Xuất sắc	8.43	3.53	B+	Giỏi	Học tiếp
8	2001140364	Nguyễn Nam	Bình	3.3	2.5		7			8.3	0		0		0		1.4	54	Trung bình	3.69	1.08	F	Yếu	Học tiếp
9	2001140022	Huỳnh Kim	Chi															82	Tốt	0	0	F	Kém	Cảnh cáo h
10	2001140508	Phạm Văn	Chương				0									0		75	Khá	0	0	F	Kém	Học tiếp
11	2001140350	Trần Nguyên	Chương	5.8	6.8	6.4	7.5	7.3	7.7	7.9	8.7		9		6.4		8.1	90	Xuất sắc	7.22	2.79	B	Khá	Học tiếp
12	2001140026	Lê Quốc	Cường	6.5	5.2	6.3	7.2	7.8	6.5	7.4		7.5	8.2				6.6	87	Tốt	6.83	2.55	C+	Khá	Học tiếp
13	2001140374	Hồ Thành	Danh														4.3	72	Khá	2.25	0.1	F	Kém	Học tiếp
14	2001140511	Nguyễn Hữu	Đạt	5.4	5	7.7	6.2	7.8	7.4	7.2	5		6.6				6.6	90	Xuất sắc	6.78	2.58	C+	Khá	Học tiếp

Do mỗi tập tin excel chỉ chứa thông tin về điểm của một số môn học nên cần thực hiện tổng hợp dữ liệu từ nhiều tập tin, sau đó loại bỏ các môn học chung, chỉ giữ lại các môn học cơ sở ngành có kiến thức bổ trợ cho từng chuyên ngành (theo danh sách môn ở mục 2.1) nhằm phục vụ cho bài toán phân cụm chuyên ngành.

Bảng 2. Kết quả học tập thu được sau bước tiền xử lý

Mã SV	Họ Tên	Ngôn ngữ lập trình	Thực hành ngôn ngữ lập trình	Phân tích thiết kế hệ thống tin	Thực hành cơ sở tin học viễn thông	Thực hành hệ quản trị cơ sở dữ liệu	Thực hành lập trình mã nguồn mở	Thực hành phân tích thiết kế hệ thống thông tin	Thương mại điện tử ngành CNTT	Toán rời rạc	Cấu trúc dữ liệu và giải thuật	Kiến trúc máy tính	Thực hành cấu trúc dữ liệu và giải thuật	Cơ sở dữ liệu	Hệ điều hành	Lập trình hướng đối tượng	Thiết kế Web	Thực hành cơ sở dữ liệu	Thực hành lập trình hướng đối tượng	Thực hành thiết kế Web	Đồ hoa máy tính	Kỹ thuật lập trình	Lập trình Windows	Cơ sở tin học viễn thông	Hệ quản trị cơ sở dữ liệu	Lập trình mã nguồn
2001140003	Đặng Quốc An		7.8	6.5		7.5		6.5				6.1	6	6.1	5.7		7.7			8	8.6		8.3		5.6	
2001140009	Đồng Nguyễn Hoàng Anh	7.5	5			0	0		0			6.3		4.1	6	5.5		6.3	6.5		7.9		6.7		0	
2001140006	Trần Ngọc Anh	8.6	8.3	7.7			8.2	8.2	7.8		7.7	5.6	7	6.8	6.9	6.8	8	8.3	8	7.5			7.7		7.4	
2001140014	Phạm Anh Bằng	0.2	0							0.6	0	0	0	0	0	0	0									
2001140015	Nguyễn Chí Bảo	4.6	8.7	3.4	8.3	9.3		6	7			5.4	4	4.2	5.5		6.3	7.5	5.5	7.5			6.8	4.7	5.1	
2001140020	Trần Tuấn Bảo	4.4	6.9	3.1		6	6	8.2	7.7	0.4		4.2					3.7			7.3	7.8		0		5.4	
2001140333	Trương Huỳnh Gia Bảo	9.8	9.3	7.1		8.7		8	6.7	8.8	9.5	7.8	9.5	7.5	8.4	9.3	8.7	10	9.5	7.8			9.5		7.1	
2001140364	Nguyễn Nam Bình		0							0.3	0			2.5	7		8.3	0		0			0			
2001140022	Huỳnh Kim Chi	0	0							0																
2001140508	Phạm Văn Chương	6.3	0							4.2		3.7			0											
2001140350	Trần Nguyên Chương	7	6	6.8		7.7	7.5	5.3	6.4	5.5	6.3	5.4	8	6.8	7.5	7.3	7.9	8.7		9	6		5.4		6.5	
2001140026	Lê Quốc Cường	4.6		7.1		7.5		7.7	6.1	6.7		4.6		5.2	7.2	7.8	7.4		7.5	8.2			8.2		6.6	
2001140374	Hồ Thành Danh	0										5.2	6.5										0			
2001140511	Nguyễn Hữu Đạt	5.8	7.1	7.3		5.7		7.5	5.6		7.1	5.7	5.8	5	6.2	7.8	7.2	5		6.6			5.7		7.1	
2001140032	Lâm Văn Dậu	8	7.7	5.9		8.3		5.7	6.3	5.7	6.2	6.1	4.5	6	6.9	8	7.4	7.7	6.5	7	7.1		7.2		6.2	
2001140456	Quách Đình	5.4	7.8	8.3		8.7	6.5	7.2	7.1	5.8	8.6	6.1	7.5	8.2	7.6	6.9	7.9	6.7		9			7		6.2	
2001140556	Lê Trung Đô	9.1	9.7	9.2		8.2		8.7	8	8.1	8.8	7.4	9	7.7	7.6	8.6	8.2	10	8.5	8.9			8.7		9.2	
2001140467	Võ Kim Đồng	5.8	6.5	9.3		8.5		8.8	8.6	8.8	6.1	7.7	6.8	6.9	8.6	8.6	7.2	8					5		9	
2001140458	Trương Hùng Dũ	9.5	7.8	8.5		8.7	9.3	8.7	8.1	8.4	9.5	7	9	7.5	7.8	9.1	8.6	9.7		9.1			9.2		8.5	

2.3. Ứng dụng K-Means trong phân cụm chuyên ngành

Thuật toán K-Means được áp dụng phân cụm cho từng chuyên ngành, mỗi chuyên ngành có kết quả là 2 cụm, một cụm gồm các sinh viên có khả năng theo học chuyên ngành đó và cụm còn lại là sinh viên không có khả năng học, lặp lại thuật toán cho đến hết các chuyên ngành. Trọng tâm ban đầu của mỗi cụm trong từng chuyên ngành được chỉ định với 2 mức gọi là ngưỡng trên và ngưỡng dưới. Ngưỡng trên sẽ gom cụm gồm những sinh viên có

khả năng học chuyên ngành, ngưỡng dưới gom cụm gồm những sinh viên không có khả năng học chuyên ngành.

Thuật toán K-Means áp dụng vào bài toán được trình bày như sau:

Input:

- Bảng điểm sinh viên tổng hợp đã qua bước tiền xử lý
- Danh sách các môn học được chọn theo từng chuyên ngành.
- Trọng tâm của 2 cụm ứng với ngưỡng trên (có khả năng) và ngưỡng dưới (không có khả năng)

Output: Danh sách sinh viên được phân cụm theo từng chuyên ngành và danh sách sinh viên không thuộc chuyên ngành nào.

Begin

Bước 1: Khởi tạo

- Trọng tâm ban đầu theo từng chuyên ngành $C_{ki}(c_0, c_1, \dots, c_m)$; trong đó ki là trọng tâm của cụm i thuộc chuyên ngành k; m là điểm số thứ m thuộc trọng tâm C_{ki}
- Danh sách điểm số $D_{ki}(d_0, d_1, \dots, d_m)$; trong đó ki là nhóm điểm thứ i thuộc chuyên ngành k; m là điểm số thứ m thuộc nhóm điểm D_{ki}

Bước 2: Phân cụm cho mỗi chuyên ngành

For k = 1 to N do *//Lặp N chuyên ngành*

R_k = K-Means (Q_k); *//Xử lý phân cụm cho chuyên ngành Q_k*

// R_k gồm 2 tập: R_{k0} gồm những sinh viên được phân cụm vào chuyên ngành Q_k, R_{k1} gồm những sinh viên không thuộc chuyên ngành Q_k

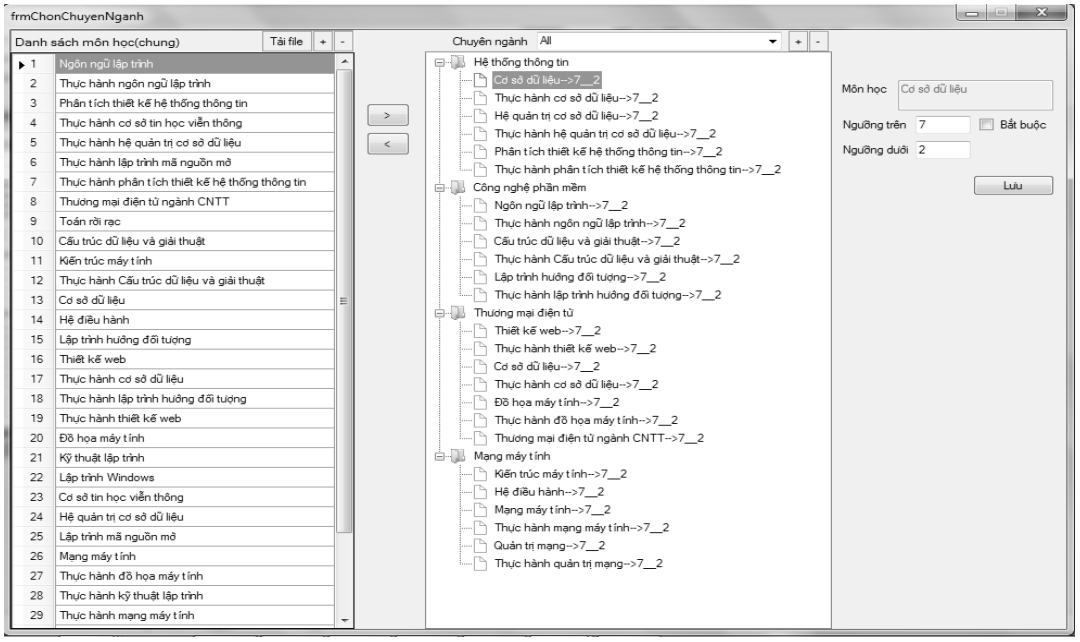
Bước 3: Xử lý kết quả

For i = 1 to M do*//Duyệt qua danh sách M sinh viên đầu vào*

Nếu $SV_i \notin$ {tập tất cả các chuyên ngành đã phân cụm}

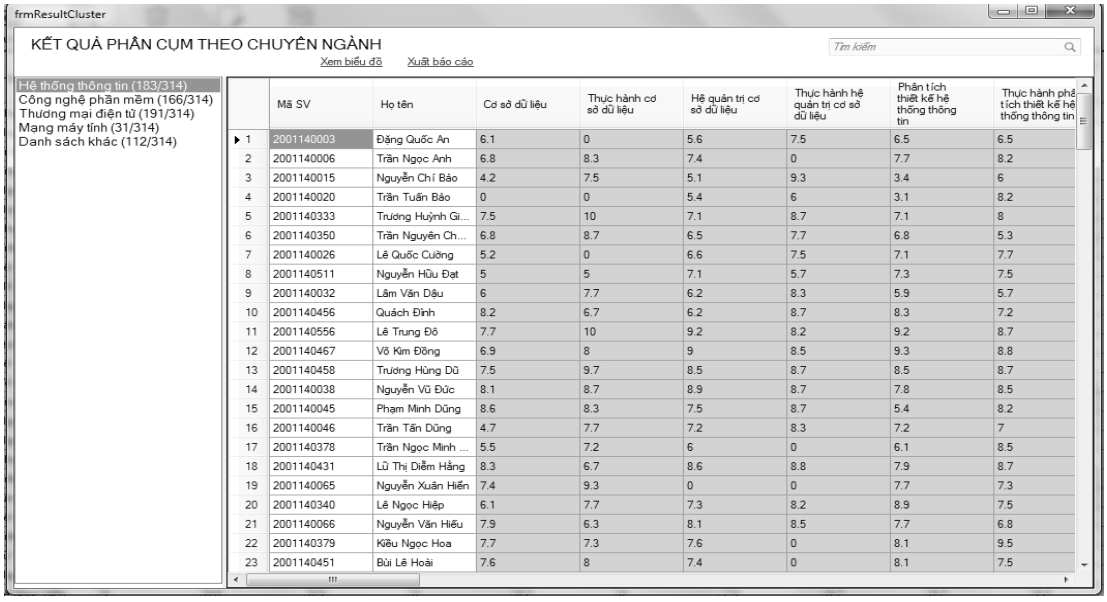
$SV_i \in$ {Danh sách khác}

End



Hình 1. Màn hình tạo các cụm cho các chuyên ngành

Kết quả phân cụm cho từng chuyên ngành:



	Mã SV	Họ tên	Cơ sở dữ liệu	Thực hành cơ sở dữ liệu	Hệ quản trị cơ sở dữ liệu	Thực hành hệ quản trị cơ sở dữ liệu	Phân tích thiết kế hệ thống thông tin	Thực hành phân tích thiết kế hệ thống thông tin
1	2001140003	Đặng Quốc An	6.1	0	5.6	7.5	6.5	6.5
2	2001140006	Trần Ngọc Anh	6.8	8.3	7.4	0	7.7	8.2
3	2001140015	Nguyễn Chí Bảo	4.2	7.5	5.1	9.3	3.4	6
4	2001140020	Trần Tuấn Bảo	0	0	5.4	6	3.1	8.2
5	2001140333	Trương Huỳnh Gi...	7.5	10	7.1	8.7	7.1	8
6	2001140350	Trần Nguyên Ch...	6.8	8.7	6.5	7.7	6.8	5.3
7	2001140026	Lê Quốc Cường	5.2	0	6.6	7.5	7.1	7.7
8	2001140511	Nguyễn Hữu Đạt	5	5	7.1	5.7	7.3	7.5
9	2001140032	Lâm Văn Dậu	6	7.7	6.2	8.3	5.9	5.7
10	2001140456	Quách Đình	8.2	6.7	6.2	8.7	8.3	7.2
11	2001140556	Lê Trung Đô	7.7	10	9.2	8.2	9.2	8.7
12	2001140467	Võ Kim Đồng	6.9	8	9	8.5	9.3	8.8
13	2001140458	Trương Hùng Dũ	7.5	9.7	8.5	8.7	8.5	8.7
14	2001140038	Nguyễn Vũ Đức	8.1	8.7	8.9	8.7	7.8	8.5
15	2001140045	Phạm Minh Dũng	8.6	8.3	7.5	8.7	5.4	8.2
16	2001140046	Trần Tấn Dũng	4.7	7.7	7.2	8.3	7.2	7
17	2001140378	Trần Ngọc Minh ...	5.5	7.2	6	0	6.1	8.5
18	2001140431	Lữ Thị Diễm Hằng	8.3	6.7	8.6	8.8	7.9	8.7
19	2001140065	Nguyễn Xuân Hiển	7.4	9.3	0	0	7.7	7.3
20	2001140340	Lê Ngọc Hiệp	6.1	7.7	7.3	8.2	8.9	7.5
21	2001140066	Nguyễn Văn Hiếu	7.9	6.3	8.1	8.5	7.7	6.8
22	2001140379	Kiều Ngọc Hoa	7.7	7.3	7.6	0	8.1	9.5
23	2001140451	Bùi Lê Hoài	7.6	8	7.4	0	8.1	7.5

Hình 2. Danh sách sinh viên thuộc các chuyên ngành sau khi phân cụm

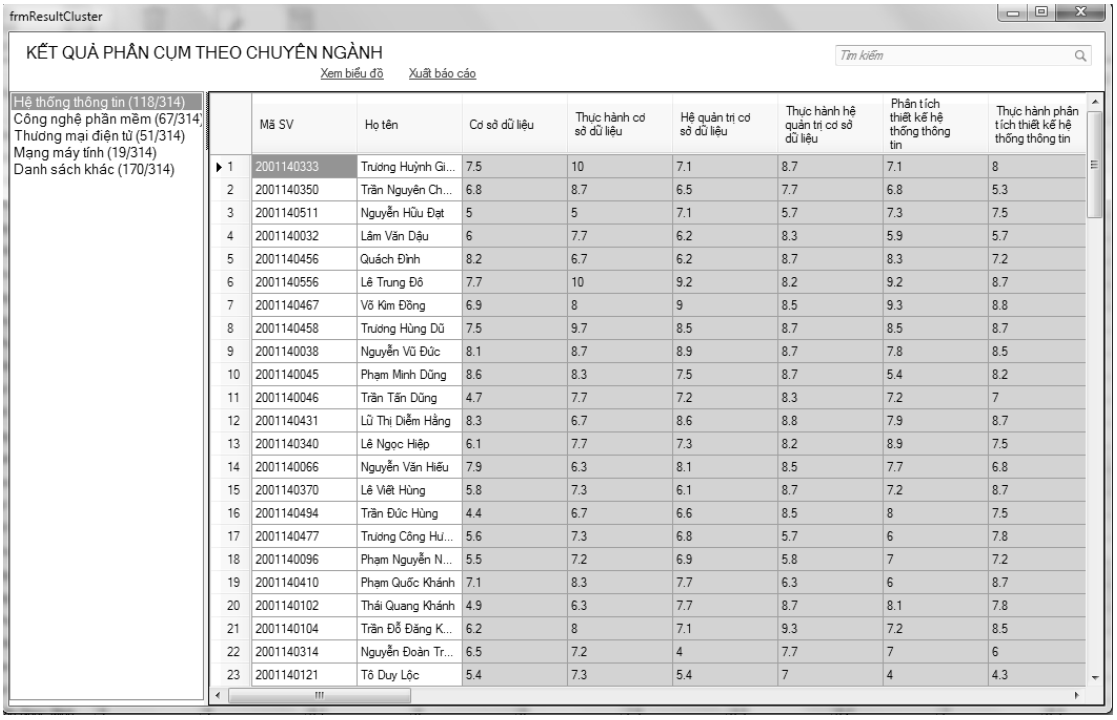
2.4. Cải tiến thuật toán K-Means

Kết quả phân cụm cho thấy phần lớn sinh viên được phân hoạch vào cụm thuộc chuyên ngành Hệ thống thông tin đều có điểm các môn học trong khoảng từ 6.0 đến 10.0 điểm (Hình 2). Tuy nhiên, vài trường hợp có điểm một số môn rất thấp (dưới 2.0) như sinh viên có mã 2001140003 có điểm môn Thực hành hệ quản trị cơ sở dữ liệu đạt 7.5 điểm, trong khi môn Thực hành cơ sở dữ liệu chỉ đạt 0.0 điểm nhưng sinh viên này vẫn được phân hoạch vào cụm Hệ thống thông tin. Lý do là việc phân cụm này sử dụng khoảng cách Euclid.

Theo công thức (2), giả sử có 2 trọng tâm: trọng tâm (6, 6, 6) thuộc phân cụm Hệ thống thông tin và trọng tâm (2, 2, 2) không thuộc phân cụm Hệ thống thông tin. Nếu một sinh viên thi 3 môn có điểm lần lượt là 2, 7, 10 thì khoảng cách Euclid từ sinh viên này đến hai trọng tâm là $\sqrt{33}$, $\sqrt{89}$. Kết quả sinh viên được phân hoạch vào cụm Hệ thống thông tin. Tuy nhiên, nếu sinh viên này học chuyên ngành Hệ thống thông tin thì sẽ không thể đáp ứng được yêu cầu của chuyên ngành vì có một môn cơ sở ngành chỉ đạt 2.0 điểm.

Đây là điểm hạn chế khi áp dụng thuật toán K-Means trong bài toán này. Để khắc phục, nhóm tác giả đề xuất phương pháp loại bỏ những trường hợp nhiễu như ví dụ trên. Theo quy chế đào tạo Tín chỉ, một sinh viên đạt một môn học nếu có điểm học phần đạt từ 4.0 trở lên [6]. Nếu sinh viên có điểm thi dưới 4.0 thì sẽ phải học lại môn học đó. Do đó, nếu nhóm điểm theo chuyên ngành của sinh viên có điểm không đạt thì sinh viên khó có thể học tiếp vào chuyên ngành. Vì thế, thực hiện loại bỏ những sinh viên có điểm trong nhóm điểm theo chuyên ngành nhỏ hơn 4.0 và không tính khoảng cách Euclid từ những sinh viên này đến các cụm trong quá trình thực hiện thuật toán K-Means nhằm giảm thời gian tính toán đối với chuyên ngành đang xét. Tuy nhiên, những sinh viên này vẫn được xét lại khi phân cụm cho chuyên ngành khác. Để thực hiện điều này, cần bổ sung vào thuật toán K-Means tại bước 2 như sau:

Đối với mỗi nhóm điểm $D(d_0, d_1, \dots, d_m)$ thuộc chuyên ngành đang xét, nếu $\exists d_i < 4$ thì loại D ra khỏi tập phân cụm. Kết quả thực hiện như sau:

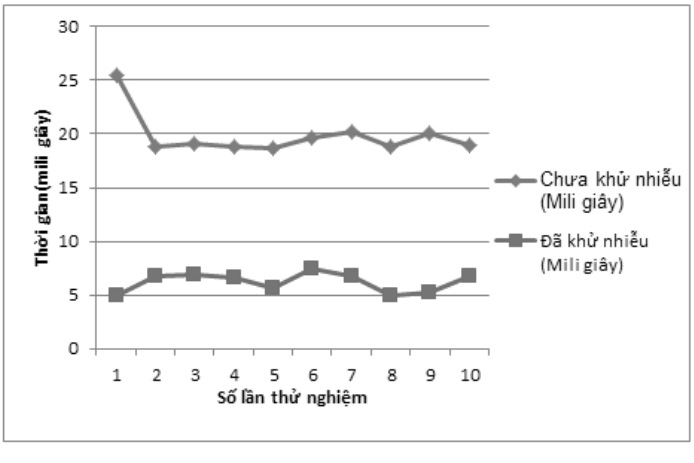


	Mã SV	Họ tên	Cơ sở dữ liệu	Thực hành cơ sở dữ liệu	Hệ quản trị cơ sở dữ liệu	Thực hành hệ quản trị cơ sở dữ liệu	Phân tích thiết kế hệ thống thông tin	Thực hành phân tích thiết kế hệ thống thông tin
1	2001140333	Trương Huỳnh Gi...	7.5	10	7.1	8.7	7.1	8
2	2001140350	Trần Nguyên Ch...	6.8	8.7	6.5	7.7	6.8	5.3
3	2001140511	Nguyễn Hữu Đạt	5	5	7.1	5.7	7.3	7.5
4	2001140032	Lâm Văn Dậu	6	7.7	6.2	8.3	5.9	5.7
5	2001140456	Quách Đình	8.2	6.7	6.2	8.7	8.3	7.2
6	2001140556	Lê Trung Đô	7.7	10	9.2	8.2	9.2	8.7
7	2001140467	Võ Kim Đồng	6.9	8	9	8.5	9.3	8.8
8	2001140458	Trương Hùng Dũ	7.5	9.7	8.5	8.7	8.5	8.7
9	2001140038	Nguyễn Vũ Đức	8.1	8.7	8.9	8.7	7.8	8.5
10	2001140045	Phạm Minh Dũng	8.6	8.3	7.5	8.7	5.4	8.2
11	2001140046	Trần Tấn Dũng	4.7	7.7	7.2	8.3	7.2	7
12	2001140431	Lữ Thị Diễm Hằng	8.3	6.7	8.6	8.8	7.9	8.7
13	2001140340	Lê Ngọc Hiệp	6.1	7.7	7.3	8.2	8.9	7.5
14	2001140066	Nguyễn Văn Hiếu	7.9	6.3	8.1	8.5	7.7	6.8
15	2001140370	Lê Viết Hùng	5.8	7.3	6.1	8.7	7.2	8.7
16	2001140494	Trần Đức Hùng	4.4	6.7	6.6	8.5	8	7.5
17	2001140477	Trương Công Hư...	5.6	7.3	6.8	5.7	6	7.8
18	2001140096	Phạm Nguyễn N...	5.5	7.2	6.9	5.8	7	7.2
19	2001140410	Phạm Quốc Khánh	7.1	8.3	7.7	6.3	6	8.7
20	2001140102	Thái Quang Khánh	4.9	6.3	7.7	8.7	8.1	7.8
21	2001140104	Trần Đỗ Đăng K...	6.2	8	7.1	9.3	7.2	8.5
22	2001140314	Nguyễn Đoàn Tr...	6.5	7.2	4	7.7	7	6
23	2001140121	Tô Duy Lộc	5.4	7.3	5.4	7	4	4.3

Hình 3. Kết quả phân cụm chuyên ngành Hệ thống thông tin sau khi loại điểm nhiễu

Việc loại bỏ những điểm nhiễu trong quá trình phân cụm làm giảm đáng kể thời gian thực hiện thuật toán. Theo thử nghiệm, khi thực hiện thuật toán 10 lần trên cùng một tập dữ liệu đầu vào, cùng một máy tính và đo thời gian (tính bằng mili giây) với 2 phương pháp: không khử nhiễu và có khử nhiễu. Kết quả được thể hiện trong Hình 4:

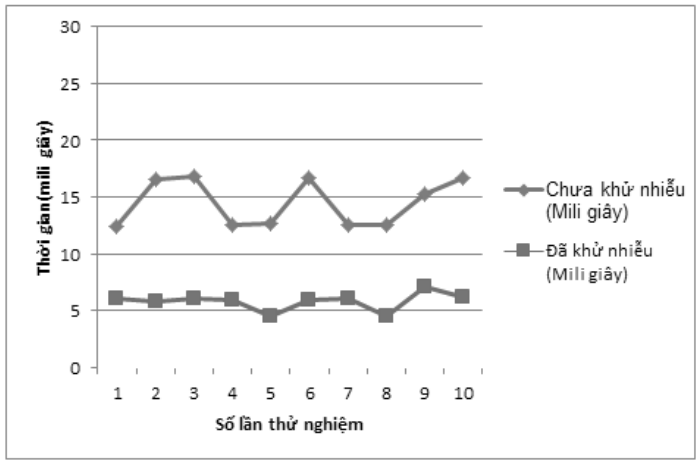
Số lần	Chưa khử nhiễu (Mili giây)	Đã khử nhiễu (Mili giây)
1	25,5025	4,9846
2	18,8674	6,7111
3	19,1128	6,9094
4	18,8481	6,6384
5	18,6355	5,6284
6	19,7128	7,5108
7	20,2436	6,7159
8	18,8703	4,9726
9	20,0742	5,2767
10	18,9044	6,7323
Trung bình	19,87716	6,20802



Hình 4. Biểu đồ so sánh thời gian thực hiện thuật toán với tập dữ liệu ban đầu

Để kiểm tra tính hiệu quả của việc khử nhiễu, 94 sinh viên trong danh sách sinh viên đầu vào bị loại bỏ, những sinh viên này đều có điểm các môn dưới 4.0, nghĩa là giảm số lượng phần tử nhiễu, còn lại 220 sinh viên và tiếp tục phân cụm với 2 phương pháp như trên. Kết quả được thể hiện trong Hình 5:

Số lần	Chưa khử nhiễu (Mili giây)	Đã khử nhiễu (Mili giây)
1	12,4334	6,0254
2	16,5957	5,7531
3	16,7959	6,0538
4	12,611	5,9176
5	12,6437	4,4543
6	16,7362	5,9893
7	12,6037	6,0268
8	12,5595	4,539
9	15,297	7,0431
10	16,6404	6,2506
Trung bình	14,49165	5,8053



Hình 5. Biểu đồ so sánh thời gian thực hiện thuật toán sau khi loại một số phần tử nhiễu

Kết quả trên cho thấy, thời gian thực thi thuật toán đã khử nhiễu tương đối ổn định không phụ thuộc vào dữ liệu nhiễu. Trong khi thuật toán chưa khử nhiễu có thời gian thực thi tăng tuyến tính với lượng dữ liệu đầu vào.

3. KẾT LUẬN VÀ ĐÁNH GIÁ

Phân cụm dữ liệu hiện nay đang được ứng dụng rộng rãi trong nhiều lĩnh vực. Thuật toán K-means là một trong những thuật toán đơn giản của phân cụm nhưng có hiệu quả cao và được ứng dụng rộng rãi. Nghiên cứu này thực hiện phân cụm dữ liệu cho bài toán hỗ trợ sinh viên Khoa Công nghệ Thông tin, Trường Đại học Công nghiệp Thực phẩm Thành phố Hồ Chí Minh lựa chọn chuyên ngành. Ngoài ra, nhóm tác giả đề xuất thêm phương pháp khử bớt nhiễu dựa trên bài toán thực tế quản lý điểm sinh viên. Kết quả cho thấy thời gian thực hiện thuật toán sau khi khử nhiễu giảm đáng kể và gợi ý lựa chọn chuyên ngành của sinh viên khá chính xác. Ứng dụng trên sẽ giúp sinh viên có quyết định chắc chắn và hợp lý hơn khi lựa chọn chuyên ngành phù hợp với khả năng của mình.

TÀI LIỆU THAM KHẢO

1. Thuật toán K-Means, 2016.
(http://ungdung.khoa-hnvd.com/Hoc_thuat/KMeans.html).
2. Đinh Mạnh Tường - Học máy: các kỹ thuật cơ bản và hiện đại, Nhà xuất bản Đại học Quốc gia Hà Nội, 2015, tr. 480-481.
3. Jame McCaffrey - K-Means data clustering using C#, Visual Studio Magazine, 2013 (https://visualstudiomagazine.com/Articles/2013/12/01/K-Means-Data-Clustering-Using-C.aspx?Page=1).
4. Nguyễn Văn Chức - Thuật toán K-Means với bài toán phân cụm dữ liệu, BIS 2010 (http://bis.net.vn/forums/t/374.aspx).
5. Trường Đại học Công nghiệp Thực phẩm TP. Hồ Chí Minh.- Chương trình đào tạo ngành Công nghệ thông tin, Quyết định số 2352/QĐ-DCT ngày 30 tháng 9 năm 2014.
6. Trường Đại học Công nghiệp Thực phẩm TP. Hồ Chí Minh.- Quy chế Đào tạo Đại học theo hệ thống tín chỉ, Quyết định số 1603/QĐ-DCT ngày 23 tháng 08 năm 2017.

7. Dữ liệu kết quả học tập của sinh viên khóa 05DHTH, Khoa Công nghệ Thông tin, Trường Đại học Công nghiệp Thực phẩm TP.HCM, do Phòng Đào tạo cung cấp.

ABSTRACT

IMPROVEMENT OF K-MEANS ALGORITHM AND APPLICATIONS FOR STUDENTS TO CHOOSE MAJORS BASED ON CREDITS SYSTEM

Nguyen Van Le*, Manh Thien Ly
Nguyen Thi Dinh, Nguyen Thi Thanh Thuy
Ho Chi Minh City University of Food Industry
*Email: *lecntp@gmail.com*

K-means algorithm is very effectively used in many applications about database clustering. The authors applied this algorithm to professional clustering on the score data set, but the authorithms were inefficient in some cases, so the accuracy was not high. Therefore, in this paper, a clustering method on the group data set specific to each discipline was proposed. In addition, the improvement of K-Means algorithm for removing noise items was done in order to reduce the computation time of the algorithm. The result of this clustering will support students of Faculty of Information Technology, Ho Chi Minh City University of Food Industry in choosing their majors.

Keywords: K-Means, clustering, choose specialized, Euclidean distance, centroids.

Giấy phép xuất bản số 435/GP-BTTTT, ngày 23 tháng 10 năm 2013.
In tại Công ty TNHH Sản xuất Thương mại Dịch vụ Khang Hưng Phú.
Số lượng 250 cuốn, khổ 19 x 27 cm. In xong và nộp lưu chiểu tháng 6 năm 2018.