



ĐẠI HỌC KHOA HỌC TỰ NHIÊN, ĐHQG-HCM
KHOA CÔNG NGHỆ THÔNG TIN

KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG
Bài tập Thực hành 4

CLUSTERING

Nhóm thực hiện

1. Hồng Thanh Hoài 1612855
 2. Huỳnh Minh Huân 1612858
-

Giáo viên lý thuyết
PGS.TS Lê Hoài Bắc

Giáo viên hướng dẫn
Nguyễn Ngọc Thảo

Tháng 06 năm 2019

Lời cảm ơn

Trong quá trình thực hiện bài tập này, nhóm chúng em đã nhận được rất nhiều sự giúp đỡ cũng như hỗ trợ từ các thầy cô Trường Đại học Khoa học Tự nhiên, ĐHQG-HCM và các bạn bè trong trường. Nhóm chúng em xin bày tỏ lòng cảm ơn chân thành đến mọi người vì đã hướng dẫn, chỉ bảo rất tận tình.

Đặc biệt, nhóm chúng em xin bày tỏ lòng biết ơn sâu sắc đến các thầy cô khoa Công nghệ thông tin, cụ thể hơn là thầy Lê Hoài Bắc đã giảng dạy rất kỹ lưỡng để chúng em nắm rõ kiến thức và cô Nguyễn Ngọc Thảo đã hướng dẫn chúng em thực hiện bài tập này rất nhiệt tình.

Một lần nữa, chúng em xin bày tỏ lòng biết ơn sâu sắc đến với các thầy cô và bạn bè.

Tháng 06 năm 2019,

Đại học Khoa học Tự nhiên, ĐHQG-HCM.

Mục lục

Lời cảm ơn	i
1 Giới thiệu nhóm và phân công công việc	1
1.1 Giới thiệu nhóm	1
1.2 Phân công công việc	1
2 Nội dung	2
2.1 Weka	2
2.1.1 Câu 1	2
2.1.2 Câu 2	5
2.1.3 Câu 3	6
2.1.4 Câu 4	6
2.1.5 Câu 5	6
2.1.6 Câu 6	7
2.2 Phần cài đặt	7
2.2.1 Kết quả chạy với <i>random-initial-starting-points</i>	8
2.2.2 Kết quả chạy với <i>Weka-initial-starting-points</i>	11
2.2.3 Nhận xét	14
3 Đánh giá	15
Tài liệu tham khảo	16

1 Giới thiệu nhóm và phân công công việc

1.1 Giới thiệu nhóm

Nhóm gồm 2 thành viên.

STT	Họ và tên	MSSV	Email	SĐT
1	Hồng Thanh Hoài	1612855	hthoai1006@gmail.com	0965596807
2	Huỳnh Minh Huân	1612858	minhhuanhuynh289@gmail.com	0824540646

1.2 Phân công công việc

STT	Họ và tên	Công việc
1	Hồng Thanh Hoài	Phần Weka.
2	Huỳnh Minh Huân	Phần cài đặt.

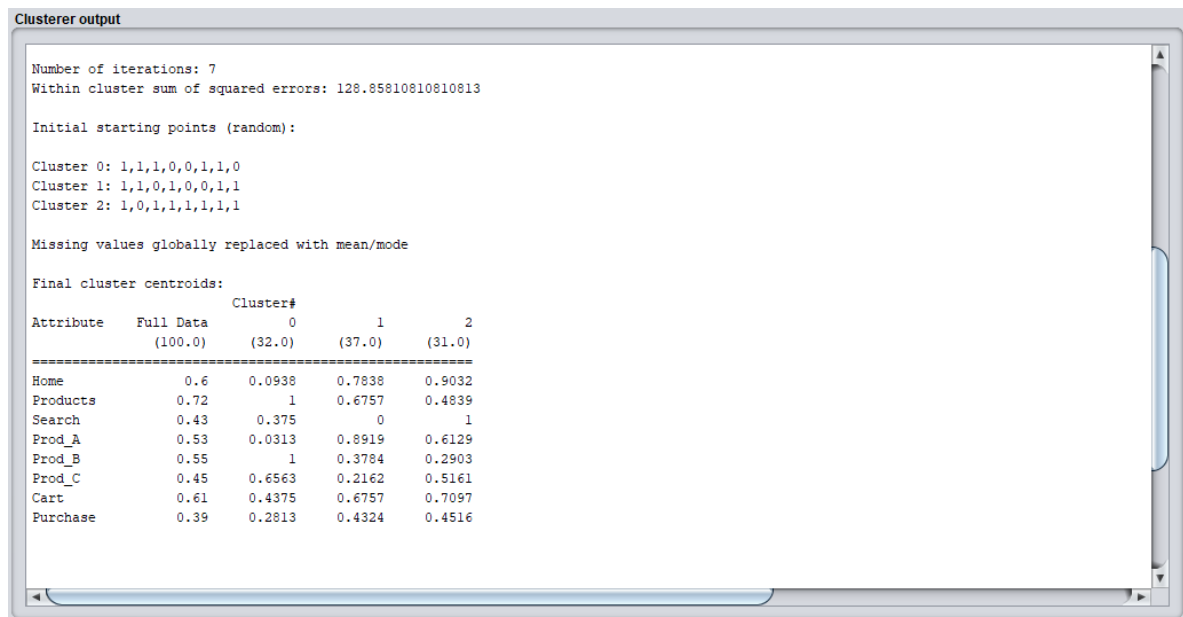
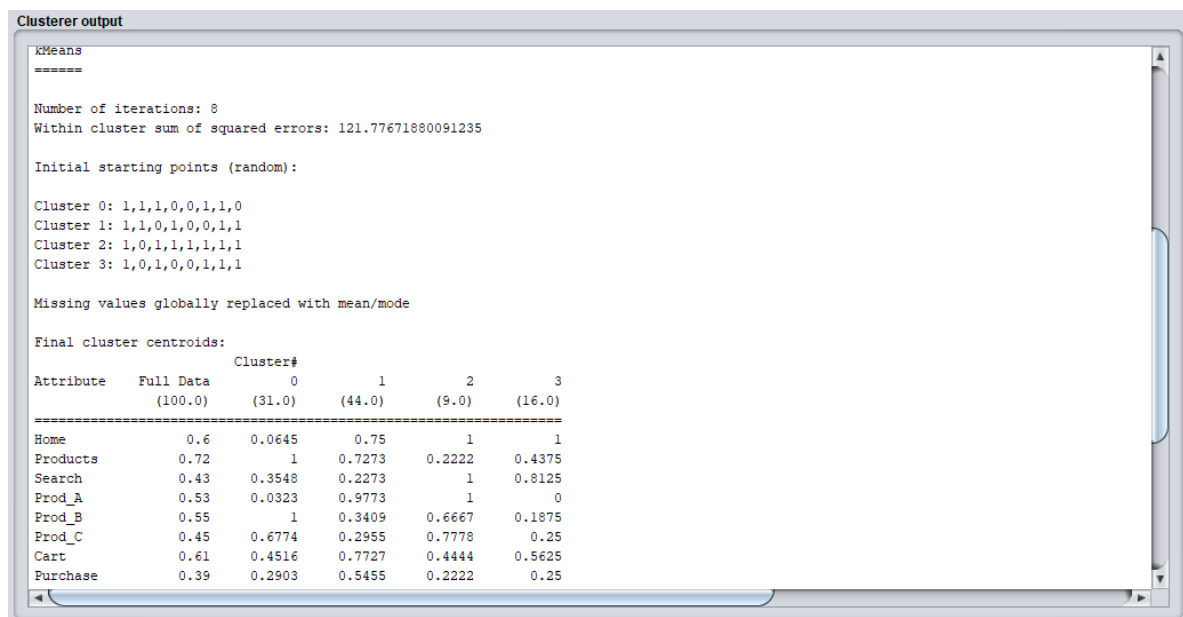
2 Nội dung

2.1 Weka

2.1.1 Câu 1

Bảng 1: Bảng với các giá trị k từ 3–8.

k	SSE	Cluster centroids								
			Home	Products	Search	Prod_A	Prod_B	Prod_C	Cart	Purchase
3	128.8581	1	0.0938	1	0.375	0.0313	1	0.6563	0.4375	0.2813
		2	0.7838	0.6757	0	0.8919	0.3784	0.2162	0.6757	0.4324
		3	0.9032	0.4839	1	0.6129	0.2903	0.5161	0.7097	0.4516
4	121.7767	1	0.0645	1	0.3548	0.0323	1	0.6774	0.4516	0.2903
		2	0.75	0.7273	0.2273	0.9773	0.3409	0.2955	0.7727	0.5455
		3	1	0.2222	1	1	0.6667	0.7778	0.4444	0.2222
		4	1	0.4375	0.8125	0	0.1875	0.25	0.5625	0.25
5	113.5826	1	0.9615	0.6923	0.6538	0.4615	0.3846	0.5385	0.4615	0
		2	0.6667	0.6667	0	0.963	0.4444	0	0.6296	0.5185
		3	1	0	1	1	0.8	0.8	0.8	0.4
		4	0.8571	0.5714	0.8571	0.7143	0.0714	0.5714	1	1
		5	0	1	0.3214	0	1	0.6786	0.5	0.3214
6	109.3612	1	0.931	0.7241	0.7241	0.5862	0.3448	0.6552	0.5172	0.1379
		2	0.9583	0.875	0.083	0.9167	0.3333	0.0833	0.7083	0.5833
		3	1	0	1	1	1	0.75	0.75	0.5
		4	1	0.1667	1	0.1667	0	0.3333	1	0.8333
		5	0	1	0.3214	0	1	0.6786	0.5	0.3214
		6	0	0.1111	0.1111	1	0.5556	0	0.6667	0.5556
7	93.7901	1	0.9048	0.9048	0.7143	0.7619	0.381	0.9048	0.6667	0.1905
		2	1	0.9375	0.125	0.875	0.4375	0.125	1	0.8125
		3	1	0	1	1	1	0.75	0.75	0.5
		4	1	0.2857	1	0.1429	0	0.2857	1	0.7142
		5	0	1	0.3214	0	1	0.6786	0.5	0.3214
		6	0	0.1429	0.1429	1	0.2857	0	1	0.8571
		7	0.8235	0.4118	0.2941	0.6471	0.3529	0	0	0
8	88.9319	1	0.8889	1	0.6667	0.7778	0.3889	0.9444	0.6667	0.2222
		2	1	0.9375	0.125	0.875	0.4375	0.125	1	0.8125
		3	1	0	1	1	0.6667	0.8333	0.6667	0.3333
		4	1	0.5	1	0	0	1	1	1
		5	0	1	0.3214	0	1	0.6786	0.5	0.3214
		6	0	0.1429	0.1429	1	0.2857	0	1	0.8571
		7	0.75	0.5	0	0.8333	0.8333	0	0	0
		8	1	0.2727	1	0.1818	0.2727	0	0.5455	0.2727

Hình 1: $k = 3$.Hình 2: $k = 4$.

Clusterer output

Number of iterations: 8
Within cluster sum of squared errors: 113.58260073260074

Initial starting points (random):

Cluster 0: 1,1,1,0,0,1,1,0
Cluster 1: 1,1,0,1,0,0,1,1
Cluster 2: 1,0,1,1,1,1,1,1
Cluster 3: 1,0,1,0,0,1,1,1
Cluster 4: 0,1,1,0,1,1,1,1

Missing values globally replaced with mean/mode

Final cluster centroids:

Attribute	Full Data (100.0)	Cluster# 0 (26.0)	1 (27.0)	2 (5.0)	3 (14.0)	4 (28.0)
Home	0.6	0.9615	0.6667	1	0.8571	0
Products	0.72	0.6923	0.6667	0	0.5714	1
Search	0.43	0.6538	0	1	0.8571	0.3214
Prod_A	0.53	0.4615	0.963	1	0.7143	0
Prod_B	0.55	0.3846	0.4444	0.8	0.0714	1
Prod_C	0.45	0.5385	0	0.8	0.5714	0.6786
Cart	0.61	0.4615	0.6296	0.8	1	0.5
Purchase	0.39	0	0.5185	0.4	1	0.3214

Hình 3: $k = 5$.

Clusterer output

Number of iterations: 4
Within cluster sum of squared errors: 109.36117952928299

Initial starting points (random):

Cluster 0: 1,1,1,0,0,1,1,0
Cluster 1: 1,1,0,1,0,0,1,1
Cluster 2: 1,0,1,1,1,1,1,1
Cluster 3: 1,0,1,0,0,1,1,1
Cluster 4: 0,1,1,0,1,1,1,1
Cluster 5: 0,0,0,1,1,0,1,1

Missing values globally replaced with mean/mode

Final cluster centroids:

Attribute	Full Data (100.0)	Cluster# 0 (29.0)	1 (24.0)	2 (4.0)	3 (6.0)	4 (28.0)	5 (9.0)
Home	0.6	0.931	0.9583	1	1	0	0
Products	0.72	0.7241	0.875	0	0.1667	1	0.1111
Search	0.43	0.7241	0.0833	1	1	0.3214	0.1111
Prod_A	0.53	0.5862	0.9167	1	0.1667	0	1
Prod_B	0.55	0.3448	0.3333	1	0	1	0.5556
Prod_C	0.45	0.6552	0.0833	0.75	0.3333	0.6786	0
Cart	0.61	0.5172	0.7083	0.75	1	0.5	0.6667
Purchase	0.39	0.1379	0.5833	0.5	0.8333	0.3214	0.5556

Hình 4: $k = 6$.

Clusterer output

Number of iterations: 4
Within cluster sum of squared errors: 93.79009103641458

Initial starting points (random):

Cluster 0: 1,1,1,0,0,1,1,0
Cluster 1: 1,1,0,1,0,0,1,1
Cluster 2: 1,0,1,1,1,1,1,1
Cluster 3: 1,0,1,0,0,1,1,1
Cluster 4: 0,1,1,0,1,1,1,1
Cluster 5: 0,0,0,1,1,0,1,1
Cluster 6: 0,0,0,1,1,0,0,0

Missing values globally replaced with mean/mode

Final cluster centroids:

Attribute	Full Data (100.0)	Cluster# 0 (21.0)	1 (16.0)	2 (4.0)	3 (7.0)	4 (28.0)	5 (7.0)	6 (17.0)
Home	0.6	0.9048	1	1	1	0	0	0.8235
Products	0.72	0.9048	0.9375	0	0.2857	1	0.1429	0.4118
Search	0.43	0.7143	0.125	1	1	0.3214	0.1429	0.2941
Prod_A	0.53	0.7619	0.875	1	0.1429	0	1	0.6471
Prod_B	0.55	0.381	0.4375	1	0	1	0.2857	0.3529
Prod_C	0.45	0.9048	0.125	0.75	0.2857	0.6786	0	0
Cart	0.61	0.6667	1	0.75	1	0.5	1	0
Purchase	0.39	0.1905	0.8125	0.5	0.7143	0.3214	0.8571	0

Hình 5: $k = 7$.

Clusterer output

Within cluster sum of squared errors: 88.93190836940838

Initial starting points (random):

Cluster 0: 1,1,1,0,0,1,1,0
Cluster 1: 1,1,0,1,0,0,1,1
Cluster 2: 1,0,1,1,1,1,1,1
Cluster 3: 1,0,1,0,0,1,1,1
Cluster 4: 0,1,1,0,1,1,1,1
Cluster 5: 0,0,0,1,1,0,1,1
Cluster 6: 0,0,0,1,1,0,0,0
Cluster 7: 1,0,1,0,0,0,1,1

Missing values globally replaced with mean/mode

Final cluster centroids:

Attribute	Full Data (100.0)	Cluster# 0 (18.0)	1 (16.0)	2 (6.0)	3 (2.0)	4 (28.0)	5 (7.0)	6 (12.0)	7 (11.0)
Home	0.6	0.8889	1	1	1	0	0	0.75	1
Products	0.72	1	0.9375	0	0.5	1	0.1429	0.5	0.2727
Search	0.43	0.6667	0.125	1	1	0.3214	0.1429	0	1
Prod_A	0.53	0.7778	0.875	1	0	0	1	0.8333	0.1818
Prod_B	0.55	0.3889	0.4375	0.6667	0	1	0.2857	0.3333	0.2727
Prod_C	0.45	0.9444	0.125	0.8333	1	0.6786	0	0	0
Cart	0.61	0.6667	1	0.6667	1	0.5	1	0	0.5455
Purchase	0.39	0.2222	0.8125	0.3333	1	0.3214	0.8571	0	0.2727

Hình 6: $k = 8$.

Từ các câu tiếp theo, nhóm chọn $k = 5$ để trả lời.

2.1.2 Câu 2

- New-user: [1, 0, 1, 0, 1, 0, 0, 0]
- Khoảng cách giữa điểm mới với các cụm: {1.3, 1.859, 1.574, 2.0, 1.809}

→ New-user như mô tả của câu này thuộc cụm 1 (cluster0). Ta có $Prod_A = 0.4615 < Prod_C = 0.5385$ nên sẽ giới thiệu $Prod_C$ đến người dùng này.

2.1.3 Câu 3

- New-user: [0, 1, 0, 0, 0, 1, 0, 0]
- Khoảng cách giữa điểm mới với các cụm: {1.49, 1.829, 2.3409, 2.08, 1.24}

→ New-user như mô tả của câu này thuộc cụm 5 (cluster4). Ta có $Prod_B = 1 < Prod_A = 0$ nên sẽ giới thiệu $Prod_B$ đến người dùng này.

2.1.4 Câu 4

- Người dùng thông thường (window shopper, xem nhiều sản phẩm): Cụm 3 (cluster2) vì tỉ lệ xem $Prod_A$, $Prod_B$ và $Prod_C$ cao hơn hẳn các cụm còn lại.

Dẫn chứng:

- + Mẫu 09: [1,0,1,1,1,1,1,1]
- + Mẫu 10: [1,0,1,1,1,1,1,0]
- + Mẫu 14: [1,0,1,1,0,1,1,0]

- Người dùng tập trung (biết rõ cần mua sản phẩm gì): Cụm 2 (cluster1) vì có tỉ lệ $Cart$ và $Purchase$ cao trong khi $Search$ thấp.

Dẫn chứng:

- + Mẫu 69: [1,1,0,1,0,0,1,1]
- + Mẫu 71: [1,1,0,1,1,0,1,1]
- + Mẫu 72: [0,0,0,1,0,0,1,1]

- Người dùng tìm kiếm (sử dụng chức năng search để tìm sản phẩm cần mua): Cụm 3 (cluster2) vì có tỉ lệ người dùng sử dụng chức năng search là 100%.

Dẫn chứng:

- + Mẫu 09: [1,0,1,1,1,1,1,1]
- + Mẫu 10: [1,0,1,1,1,1,1,0]
- + Mẫu 14: [1,0,1,1,1,1,0,0]

2.1.5 Câu 5

- Sản phẩm đơn lẻ - cụm 2 (cluster1): $Prod_A$.
 - + Đặc điểm hành vi duyệt trang: vào thẳng trang $Prod_A$.

- + Xu hướng thanh toán: trung bình (khoảng 51%).
- + Dẫn chứng:
 - * Mẫu 53: [1,1,0,1,0,0,0,0]
 - * Mẫu 54: [0,1,0,1,1,0,0,0]
 - * Mẫu 55: [0,0,0,1,1,0,1,1]
- Sản phẩm đơn lẻ - cụm 5 (cluster4): *Prod_B*.
 - + Đặc điểm hành vi duyệt trang: *Product* \rightarrow *Prod_B*.
 - + Xu hướng thanh toán: trung bình (khoảng 32%).
 - + Dẫn chứng:
 - * Mẫu 82: [0,1,1,0,1,0,0,0]
 - * Mẫu 83: [0,1,0,0,1,1,1,1]
 - * Mẫu 84: [0,1,0,0,1,1,0,0]
- Nhóm các sản phẩm - cụm 3 (cluster2): *Prod_A*, *Prod_B*, *Prod_C*.
 - + Đặc điểm hành vi duyệt trang: *Home* \rightarrow *Search* \rightarrow các trang *Prod_A*, *Prod_B*, *Prod_C*.
 - + Xu hướng thanh toán: trung bình (khoảng 40%).
 - + Dẫn chứng:
 - * Mẫu 09: [1,0,1,1,1,1,1,1]
 - * Mẫu 10: [1,0,1,1,1,1,1,0]
 - * Mẫu 14: [1,0,1,1,1,1,0,0]

2.1.6 Câu 6

Có thể, đó là cụm 5 (cluster4) vì có tỉ lệ vào 3 trang *Home*, *Product*, *Search* thấp. Chiến dịch quảng cáo cho *Prod_B* thành công hơn vì có tỉ lệ truy cập ở cụm này là 100%.

2.2 Phần cài đặt

Chạy chương trình cài đặt với tập dữ liệu `sessions.csv`. Đối chiếu kết quả phát sinh được với kết quả của Weka trên cùng giá trị k (từ 3 đến 8).

2.2.1 Kết quả chạy với *random-initial-starting-points*

```

1 | Within cluster sum of squared errors: 127.712307
2 | Cluster centroids:
3 |   Cluster#
4 | Attribute  0          1          2
5 |           (24)       (29)       (47)
6 | =====
7 | Home       0.9167      0.0       0.8085
8 | Products   0.5        1.0       0.6596
9 | Search     0.5        0.3103    0.4681
10 | Prod_A     0.625      0.0345    0.7872
11 | Prod_B     0.4167      1.0       0.3404
12 | Prod_C     0.3333     0.6552    0.383
13 | Cart       0.0        0.4828    1.0
14 | Purchase   0.0        0.3103    0.6383
15 |

```

Iteration 4:
SSE = 85.43425324675326

SSE value = 85.43425324675326

E:\Visual Studio\Workspace\HCMUS\HK6\KTDL\HCMUS_HK6_KTDL\Labs\Lab04\src>
E:\Visual Studio\Workspace\HCMUS\HK6\KTDL\HCMUS_HK6_KTDL\Labs\Lab04\src>

Hình 7: Random-initial-starting-points: $k = 3$.

```

1 | Within cluster sum of squared errors: 116.145227
2 | Cluster centroids:
3 |   Cluster#
4 | Attribute  0          1          2          3
5 |           (34)       (27)       (11)       (28)
6 | =====
7 | Home       0.8235     0.2222     0.0909     0.8929
8 | Products   0.7059     0.8889     1.0        0.4643
9 | Search     0.0        0.5185     0.0909     1.0
10 | Prod_A     0.9118     0.1852     0.0        0.6071
11 | Prod_B     0.3235      1.0        1.0        0.2143
12 | Prod_C     0.2353     0.5185     0.8182     0.5
13 | Cart       0.7059     0.1481     1.0        0.7857
14 | Purchase   0.4412     0.0        0.9091     0.5
15 |

```

Iteration 4:
SSE = 85.43425324675326

SSE value = 85.43425324675326

E:\Visual Studio\Workspace\HCMUS\HK6\KTDL\HCMUS_HK6_KTDL\Labs\Lab04\src>
E:\Visual Studio\Workspace\HCMUS\HK6\KTDL\HCMUS_HK6_KTDL\Labs\Lab04\src>

Hình 8: Random-initial-starting-points: $k = 4$.

```

1 | Within cluster sum of squared errors: 105.759343
2 | Cluster centroids:
3 |
4 |   Attribute   | Cluster#
5 |   |   |   |   |   |   |
6 |   |   |   |   |   |   |
7 |   |   |   |   |   |   |
8 |   |   |   |   |   |   |
9 |   |   |   |   |   |   |
10 |   |   |   |   |   |   |
11 |   |   |   |   |   |   |
12 |   |   |   |   |   |   |
13 |   |   |   |   |   |   |
14 |   |   |   |   |   |   |
15 |

```

Attribute	0	1	2	3	4
	(23)	(23)	(17)	(22)	(15)
Home	0.9565	0.913	0.7647	0.1364	0.0667
Products	0.3043	0.9565	0.5882	1.0	0.7333
Search	1.0	0.3478	0.0	0.5	0.0667
Prod_A	0.5217	0.9565	0.8235	0.0	0.3333
Prod_B	0.3478	0.3478	0.2941	1.0	0.8
Prod_C	0.4348	0.5652	0.0588	0.5909	0.5333
Cart	0.6087	1.0	0.2353	0.2273	1.0
Purchase	0.3043	0.7391	0.0	0.0	1.0

```

Iteration 4:
SSE = 85.43425324675326
-----
SSE value = 85.43425324675326

E:\Visual Studio\Workspace\HCMUS\HK6\KTDL\HCMUS_HK6_KTDL\Labs\Lab04\src>
E:\Visual Studio\Workspace\HCMUS\HK6\KTDL\HCMUS_HK6_KTDL\Labs\Lab04\src>

```

Hình 9: Random-initial-starting-points: $k = 5$.

```

1 | Within cluster sum of squared errors: 102.164234
2 | Cluster centroids:
3 |
4 |   Attribute   | Cluster#
5 |   |   |   |   |   |   |
6 |   |   |   |   |   |   |
7 |   |   |   |   |   |   |
8 |   |   |   |   |   |   |
9 |   |   |   |   |   |   |
10 |   |   |   |   |   |   |
11 |   |   |   |   |   |   |
12 |   |   |   |   |   |   |
13 |   |   |   |   |   |   |
14 |   |   |   |   |   |   |
15 |

```

Attribute	0	1	2	3	4	5
	(30)	(17)	(6)	(23)	(13)	(11)
Home	0.7667	0.5882	1.0	0.0	0.8462	0.9091
Products	0.6	0.5882	0.1667	1.0	0.6923	1.0
Search	0.4667	0.0	1.0	0.3043	1.0	0.2727
Prod_A	0.8333	0.6471	1.0	0.0	0.0	1.0
Prod_B	0.3333	0.4118	0.3333	1.0	0.5385	0.5455
Prod_C	0.3	0.0	0.6667	0.8261	0.3077	0.8182
Cart	1.0	0.1176	0.5	0.6087	0.3077	0.7273
Purchase	1.0	0.0	0.0	0.3913	0.0	0.0

```

Iteration 4:
SSE = 85.43425324675326
-----
SSE value = 85.43425324675326

E:\Visual Studio\Workspace\HCMUS\HK6\KTDL\HCMUS_HK6_KTDL\Labs\Lab04\src>
E:\Visual Studio\Workspace\HCMUS\HK6\KTDL\HCMUS_HK6_KTDL\Labs\Lab04\src>

```

Hình 10: Random-initial-starting-points: $k = 6$.

```

model_test_5.txt  model_test_7.txt  model_test_8.txt  model_3.txt  model_8.txt  model_7.txt x
1 | Within cluster sum of squared errors: 88.006536
2 | Cluster centroids:
3 |
4 | Attribute      0      1      2      3      4      5      6
5 |                (20)   (10)   (20)   (9)   (15)   (9)   (17)
6 | =====
7 | Home           0.7     0.0     0.9     0.7778  0.8     1.0     0.0
8 | Products       0.65    1.0     0.7     0.8889  0.4667  0.3333  1.0
9 | Search         0.25    0.5     0.65    1.0     0.0667  0.8889  0.1176
10 | Prod_A         0.95    0.1     0.85    0.3333  0.8667  0.0     0.0
11 | Prod_B         0.5     1.0     0.25    0.8889  0.2     0.2222  1.0
12 | Prod_C         0.0     0.0     0.95    0.8889  0.0667  0.0     1.0
13 | Cart           1.0     0.4     0.95    0.0     0.1333  0.6667  0.5882
14 | Purchase       0.95    0.1     0.45    0.0     0.0     0.2222  0.4706
15 |
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  1: cmd
Iteration 4:
SSE = 85.43425324675326
-----
SSE value = 85.43425324675326
E:\Visual Studio\Workspace\HCMUS\[HK6]\[KTDL]\HCMUS_HK6_KTDL\Labs\Lab04\src>
E:\Visual Studio\Workspace\HCMUS\[HK6]\[KTDL]\HCMUS_HK6_KTDL\Labs\Lab04\src>

```

Hình 11: Random-initial-starting-points: $k = 7$.

```

del_test_6.txt  model_test_5.txt  model_test_7.txt  model_test_8.txt  model_3.txt  model_8.txt x
1 | Within cluster sum of squared errors: 85.434253
2 | Cluster centroids:
3 |
4 | Attribute      0      1      2      3      4      5      6      7
5 |                (16)   (7)   (6)   (22)   (11)   (11)   (16)   (11)
6 | =====
7 | Home           1.0     0.8571  0.6667  0.0     0.7273  0.0     1.0     0.9091
8 | Products       0.5625  1.0     0.6667  0.9091  0.0     0.8182  0.75    1.0
9 | Search         1.0     0.8571  0.5     0.3636  0.7273  0.0909  0.0     0.0909
10 | Prod_A         0.375    0.8571  1.0     0.1364  0.6364  0.1818  0.8125  0.9091
11 | Prod_B         0.5625  0.0     0.0     1.0     0.1818  1.0     0.3125  0.5455
12 | Prod_C         0.5625  0.8571  0.6667  0.5     0.2727  0.7273  0.25    0.0
13 | Cart           0.25    1.0     1.0     0.2273  1.0     1.0     0.375    1.0
14 | Purchase       0.0     1.0     0.0     0.0     0.9091  1.0     0.0     1.0
15 |
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  1: cmd
Iteration 4:
SSE = 85.43425324675326
-----
SSE value = 85.43425324675326
E:\Visual Studio\Workspace\HCMUS\[HK6]\[KTDL]\HCMUS_HK6_KTDL\Labs\Lab04\src>
E:\Visual Studio\Workspace\HCMUS\[HK6]\[KTDL]\HCMUS_HK6_KTDL\Labs\Lab04\src>

```

Hình 12: Random-initial-starting-points: $k = 8$.

2.2.2 Kết quả chạy với *Weka-initial-starting-points*

```

37     - idCluster: [] chứa giá trị id của cluster mà dataPoint thuộc về.
38     - nCluster: [] chứa số lượng phân tử thuộc mỗi cluster.
39     - SSE: giá trị Sum of Squared Error.
40
41     """
42     # Initialize
43     # random k giá trị làm centroid_id và lấy ra các centroid initial
44     #print(len(data))
45     initial_id = np.random.choice(len(data), size = k, replace = False)
46     centroids = [[1,1,1,0,0,1,1,0],
47                  [1,1,0,1,0,0,1,1],
48                  [1,0,1,1,1,1,1,1]],
49                  # [1,0,1,0,0,1,1,1],
50                  # [0,1,1,0,1,1,1,1],
51                  # [0,0,0,1,1,0,1,1],
52                  # [0,0,0,1,1,0,0,0],
53                  # [1,0,1,0,0,0,1,1]]
54     #centroids = []
55     #for id in initial_id:
56     #    centroids.append(data[id])
57
58     # Khởi tạo các cluster ban đầu là rỗng
59     clusters = [[] for x in range(k)]

```

```

Iteration 9:
SSE = 120.8135761213031
SSE value = 120.8135761213031

```

E:\Visual Studio\Workspace\HCMUS\ [HK6] [KTDL] \HCMUS_HK6_KTDL\Labs\Lab04\src>

Hình 13: Ở đoạn code này, nhóm thiết lập các initial-starting-points giống với bên Weka để kiểm tra phần cài đặt giải thuật k-mean của nhóm (comment phần màu đỏ, copy từ file test_case_k_init.txt ứng với số lượng k cụm muốn phân).

```

1 Within cluster sum of squared errors: 128.858108
2 Cluster centroids:
3   Cluster#
4 Attribute  0      1      2
5           (32)   (37)   (31)
6
7 Home       0.0938  0.7838  0.9032
8 Products   1.0     0.6757  0.4839
9 Search     0.375   0.0     1.0
10 Prod_A    0.0312  0.8919  0.6129
11 Prod_B    1.0     0.3784  0.2903
12 Prod_C    0.6562  0.2162  0.5161
13 Cart      0.4375  0.6757  0.7097
14 Purchase  0.2812  0.4324  0.4516
15

```

```

Iteration 7:
SSE = 128.8581081081081
SSE value = 128.8581081081081

```

E:\Visual Studio\Workspace\HCMUS\ [HK6] [KTDL] \HCMUS_HK6_KTDL\Labs\Lab04\src>

Hình 14: Weka-initial-starting-points: $k = 3$.

```

1 | Within cluster sum of squared errors: 121.776719
2 | Cluster centroids:
3 |   Cluster#
4 | Attribute  0      1      2      3
5 |           (31)  (44)  (9)   (16)
6 | =====
7 | Home      0.0645  0.75   1.0   1.0
8 | Products  1.0    0.7273 0.2222 0.4375
9 | Search    0.3548  0.2273 1.0    0.8125
10 | Prod_A    0.0323  0.9773 1.0    0.0
11 | Prod_B    1.0    0.3409 0.6667 0.1875
12 | Prod_C    0.6774  0.2955 0.7778 0.25
13 | Cart      0.4516  0.7727 0.4444 0.5625
14 | Purchase  0.2903  0.5455 0.2222 0.25
15 |

```

```

-----
Iteration 8:
SSE = 121.77671880091228
-----
SSE value = 121.77671880091228
E:\Visual Studio\Workspace\HCMUS\[HK6][KTDL]\HCMUS_HK6_KTDL\Labs\Lab04\src>

```

Hình 15: Weka-initial-starting-points: $k = 4$.

```

1 | Within cluster sum of squared errors: 113.582601
2 | Cluster centroids:
3 |   Cluster#
4 | Attribute  0      1      2      3      4
5 |           (26)  (27)  (5)   (14)  (28)
6 | =====
7 | Home      0.9615  0.6667  1.0   0.8571  0.0
8 | Products  0.6923  0.6667  0.0   0.5714  1.0
9 | Search    0.6538  0.0     1.0   0.8571  0.3214
10 | Prod_A    0.4615  0.963   1.0   0.7143  0.0
11 | Prod_B    0.3846  0.4444  0.8   0.0714  1.0
12 | Prod_C    0.5385  0.0     0.8   0.5714  0.6786
13 | Cart      0.4615  0.6296  0.8   1.0     0.5
14 | Purchase  0.0     0.5185  0.4   1.0     0.3214
15 |

```

```

-----
Iteration 8:
SSE = 113.5826007326008
-----
SSE value = 113.5826007326008
E:\Visual Studio\Workspace\HCMUS\[HK6][KTDL]\HCMUS_HK6_KTDL\Labs\Lab04\src>

```

Hình 16: Weka-initial-starting-points: $k = 5$.

```

37 - idCluster: [] chứa giá trị id của cluster mà dataPoint thuộc về.
38 - nCluster: [] chứa số lượng phân tử thuộc mỗi cluster.
39 - SSE: giá trị Sum of Squared Error.
40
41 # Initialize
42 # random k giá trị làm centroid_id và lấy ra các centroid initial
43 #print(len(data))
44 initial_id = np.random.choice(len(data), size = k, replace = False)
45 centroids = [[1,1,1,0,0,1,1,0],
46              [1,1,0,1,0,0,1,1],
47              [1,0,1,1,1,1,1,1],
48              [1,0,1,0,0,1,1,1],
49              [0,1,1,0,1,1,1,1],
50              [0,0,0,1,1,0,1,1]]
51 # [0,0,0,1,1,0,0,0],
52 # [1,0,1,0,0,0,1,1]]
53 #centroids = []
54 #for id in initial_id:
55 #    centroids.append(data[id])
56
57 # Khởi tạo các cluster ban đầu là rỗng
58 clusters = [[] for x in range(k)]

```

Iteration 4:
SSE = 109.36117952928299
SSE value = 109.36117952928299

E:\Visual Studio\Workspace\HCMUS\[HK6][KTDL]\HCMUS_HK6_KTDL\Labs\Lab04\src>

Hình 17: Weka-initial-starting-points: $k = 6$.

```

1 Within cluster sum of squared errors: 93.790091
2 Cluster centroids:
3
4 Attribute      0      1      2      3      4      5      6
5 (21) (16) (4) (7) (28) (7) (17)
6
7 Home      0.9048      1.0      1.0      1.0      0.0      0.0      0.8235
8 Products  0.9048  0.9375      0.0  0.2857      1.0  0.1429  0.4118
9 Search    0.7143  0.125      1.0      1.0      0.3214  0.1429  0.2941
10 Prod_A    0.7619  0.875      1.0  0.1429      0.0      1.0  0.6471
11 Prod_B    0.381  0.4375      1.0      0.0      1.0  0.2857  0.3529
12 Prod_C    0.9048  0.125      0.75  0.2857  0.6786      0.0      0.0
13 Cart      0.6667      1.0      0.75      1.0      0.5      1.0      0.0
14 Purchase  0.1905  0.8125      0.5  0.7143  0.3214  0.8571  0.0
15

```

Iteration 4:
SSE = 93.79009103641458
SSE value = 93.79009103641458

E:\Visual Studio\Workspace\HCMUS\[HK6][KTDL]\HCMUS_HK6_KTDL\Labs\Lab04\src>

Hình 18: Weka-initial-starting-points: $k = 7$.


```

1 Within cluster sum of squared errors: 88.931908
2 Cluster centroids:
3
4 Attribute      0      1      2      3      4      5      6      7
5 (18) (16) (6) (2) (28) (7) (12) (11)
6 =====
7 Home      0.8889      1.0      1.0      1.0      0.0      0.0      0.75      1.0
8 Products      1.0      0.9375      0.0      0.5      1.0      0.1429      0.5      0.2727
9 Search      0.6667      0.125      1.0      1.0      0.3214      0.1429      0.0      1.0
10 Prod_A      0.7778      0.875      1.0      0.0      0.0      1.0      0.8333      0.1818
11 Prod_B      0.3889      0.4375      0.6667      0.0      1.0      0.2857      0.3333      0.2727
12 Prod_C      0.9444      0.125      0.8333      1.0      0.6786      0.0      0.0      0.0
13 Cart      0.6667      1.0      0.6667      1.0      0.5      1.0      0.0      0.5455
14 Purchase      0.2222      0.8125      0.3333      1.0      0.3214      0.8571      0.0      0.2727
15

```

```

-----
Iteration 6:
SSE = 88.93190836940839
-----
SSE value = 88.93190836940839

E:\Visual Studio\Workspace\HCMUS\HK6\KTDL\HCMUS_HK6_KTDL\Labs\Lab04\src>

```

Hình 19: Weka-initial-starting-points: $k = 8$.

2.2.3 Nhận xét

- Kết quả gom cụm và SSE giữa phần cài đặt của nhóm và Weka khác nhau vì các điểm khởi tạo ban đầu (init centroid) của nhóm được lấy một cách ngẫu nhiên từ tập dữ liệu, việc lấy random khác so với Weka dẫn đến lấy các điểm khởi tạo ban đầu khác. Trong thực tế, tùy vào các center ban đầu mà thuật toán có thể có tốc độ hội tụ rất chậm hoặc thậm chí cho chúng ta nghiệm không chính xác (chỉ là local minimum - điểm cực tiểu - mà không phải giá trị nhỏ nhất). Do đó người ta thường chạy k-mean nhiều lần với các center ban đầu khác nhau rồi chọn cách có hàm mất mát cuối cùng đạt giá trị nhỏ nhất.
- Khi nhóm thiết lập các initial starting points giống với bên Weka thì kết quả gom cụm và SSE giống với kết quả Weka thực hiện.

3 Đánh giá

STT	Nội dung	Hoàn thành
1	Phần Weka	100%
2	Phần cài đặt	100%
Mức độ hoàn thành tổng thể của bài tập:		100%

Tài liệu

- [1] Slide lý thuyết.
- [2] Trang chủ của Weka: <http://www.cs.waikato.ac.nz/ml/weka/>
- [3] J. Han and M. Kamber, *Data Mining, Concepts and Techniques, Second Edition*