Trường Đại học Khoa học Tự nhiên
Khoa Công nghệ thông tin
Bộ môn Khai thác dữ liệu và ứng dụng

BÁO CÁO
BÀI TẬP THỰC HÀNH 4:
PHÂN CỤM DỮ LIỆU

GVHD: Lê Ngọc Thành, Nguyễn Ngọc Thảo

TP. Hồ Chí Minh, ngày 02 tháng 06 năm 2019

Mục lục

I. Thông tin thành viên:

1. Chu Phúc Nguyên 1512353
2. Võ Nhật Vinh 1612815

II. Báo cáo tổng quát:

Sinh viên đã hoàn thành đầy đủ yêu cầu các yêu cầu trong đề bài.

III. Báo cáo chi tiết:

A. Sử dụng công cụ WEKA để khảo sát thực nghiệm về hiệu quả của các giải thuật gom nhóm trên nhiều tập dữ liệu khác nhau.

Quy ước: số thứ tự cluster trong báo cáo sẽ = số thứ tự cluster trong weka/csv + 1. Ví dụ trong weka cluster bắt đầu từ 0 và trong báo cáo bắt đầu từ 1.

1. Thử nghiệm các giá trị k từ 3 đến 8,

k	SSE	Cluster centroids							
			Home	Products	Search	Prod_A	Prod_B	Prod_C	Cart	Purchase
3	128.85	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	121.77	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	113.58	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	109.36	1	0.931	0.7241	0.7241	0.5862	0.3448	0.6552	0.5172	0.1379

		2	0.9583	0.875	0.0833	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	93.79	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	0.7143
		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	88.93	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	0.3333	0	0	0
		8	1	0.2727	1	0.1818	0.2727	0	0.5455	0.2727

k=3:

```
Number of iterations: 7
Within cluster sum of squared errors: 128.85810810810813

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

Missing values globally replaced with mean/mode

Final cluster centroids:
                        Cluster#
Attribute    Full Data          0          1          2
                (100.0)     (32.0)     (37.0)     (31.0)
=============================================================
Home                0.6     0.0938     0.7838     0.9032
Products           0.72          1     0.6757     0.4839
Search             0.43      0.375          0          1
Prod_A             0.53     0.0313     0.8919     0.6129
Prod_B             0.55          1     0.3784     0.2903
Prod_C             0.45     0.6563     0.2162     0.5161
Cart               0.61     0.4375     0.6757     0.7097
Purchase           0.39     0.2813     0.4324     0.4516
```

k=4

```
Number of iterations: 8
Within cluster sum of squared errors: 121.77671880091235

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

Missing values globally replaced with mean/mode

Final cluster centroids:
                        Cluster#
Attribute    Full Data          0          1          2          3
               (100.0)     (31.0)     (44.0)      (9.0)     (16.0)
=============================================================================
Home                0.6     0.0645       0.75          1          1
Products           0.72          1     0.7273     0.2222     0.4375
Search             0.43     0.3548     0.2273          1     0.8125
Prod_A             0.53     0.0323     0.9773          1          0
Prod_B             0.55          1     0.3409     0.6667     0.1875
Prod_C             0.45     0.6774     0.2955     0.7778       0.25
Cart               0.61     0.4516     0.7727     0.4444     0.5625
Purchase           0.39     0.2903     0.5455     0.2222       0.25
```

k=5

```
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:
                      Cluster#
Attribute    Full Data          0          1          2          3          4
               (100.0)     (26.0)     (27.0)      (5.0)     (14.0)     (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
```

k=6

```
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:
                        Cluster#
Attribute    Full Data          0          1          2          3          4          5
               (100.0)     (29.0)     (24.0)      (4.0)      (6.0)     (28.0)      (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
```

k=7 :

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:

		Cluster#						
Attribute	Full Data	0	1	2	3	4	5	6
	(100.0)	(21.0)	(16.0)	(4.0)	(7.0)	(28.0)	(7.0)	(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

k=8

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:

		Cluster#							
Attribute	Full Data	0	1	2	3	4	5	6	7
	(100.0)	(18.0)	(16.0)	(6.0)	(2.0)	(28.0)	(7.0)	(12.0)	(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

Sinh viên chọn k = 7 để trả lời các câu hỏi bên dưới:

2. **Giả sử bạn quan sát thấy một người dùng mới đã truy cập các trang là Home => Search => Prod_B. Bạn sẽ giới thiệu sản phẩm nào đến người này? Giải thích cụ thể theo số liệu tính toán về tính gần của mẫu vừa quan sát với các cụm đã có.**

Do người dùng truy cập các trang **Home => Search => Prod_B** nên ta chọn giá trị ở các ô này là 1.

Khoảng cách giữa centroids của mẫu đang xét với các thuộc tính Home, Search, Prod_b = (1,1,1). Ta dùng khoảng cách Euclid:

Cụm	Khoảng cách đến (1,1,1)
1	0.6884
2	1.0402
3	0
4	1
5	1.2085
6	1.4982
7	0.9737

Centroids của cluster 3 có khoảng cách đến mẫu đang xét bằng nhỏ nhất (=0), nên ta chọn cluster3.

Home	Products	Search	Prod_A	Prod_B	Prod_C	Cart	Purchase
1	0	1	1	1	0.75	0.75	0.5

Sinh viên nhận thấy Prod_A = 1 (cao hơn Prod_C = 0.75), nên sinh viên sẽ giới thiệu Sản phẩm A đến người dùng này.

3. **Tương tự Câu 2, lần này người dùng truy cập các trang Products => Prod_C.**

Do người dùng truy cập các trang **Products => Prod_C** nên ta chọn tương ứng các ô này với giá trị là 1.

Khoảng cách giữa centroids đến mẫu đang xét với 2 thuộc tính Products, Prod_C:

Cụm	Khoảng cách đến (1,1,1)
1	0.1346
2	0.8777
3	1.0307
4	1.0101
5	0.3214
6	1.3171
7	1.1602

Centroids của cluster 1 có khoảng cách đến mẫu đang xét bằng nhỏ nhất (=0.1346), nên ta chọn cluster1.

Home	Products	Search	Prod_A	Prod_B	Prod_C	Cart	Purchase
0.9048	**0.9048**	**0.7143**	**0.7619**	**0.381**	**0.9048**	**0.6667**	**0.1905**

Sinh viên nhận thấy Prod_A = 7619 (cao hơn Prod_B = 0.381), nên sinh viên sẽ giới thiệu Sản phẩm A đến người dùng này.

4. **Kết quả gom cụm mà bạn chọn có thể nhận diện được các hình mẫu người dùng dưới đây hay không? Nếu có, xu hướng thanh toán của những mẫu người này cao hay thấp? Dẫn chứng cụ thể bằng một số mẫu đại diện.**

Kết quả gom cụm có thể nhận diện những mẫu người dùng dưới đây. Cụ thể:

- **Người dùng thông thường** (window shopper, xem nhiều sản phẩm):
  - Ở cluster 3, tỉ lệ người xem từng loại sản phẩm (A, B, C) = (1, 1, 0.75). Như vậy mỗi người xem ít nhất là 2 sản phẩm A và B.
  - Xu hướng thanh toán của những người này ở mức trung bình (0.5)
  - Dẫn chứng:

  (STT, Home, Products, Prod_A, Prod_B, Prod_C, Cart, Purchase)

  4,1,0,1,1,1,0,1,1
  9,1,0,1,1,1,1,1,1
  10,1,0,1,1,1,1,1,0

- **Người dùng tập trung** (biết rõ cần mua sản phẩm gì):

  Cluster 3, cluster 5, cluster 6 thỏa.

  - **Cluster 3:**

    + Tỉ lệ người xem từng loại sản phẩm (A, B, C) = (1, 1, 0.75). Mỗi người xem ít nhất là 2 sản phẩm A và B.

+ Xu hướng thanh toán ở mức trung bình: 0.5

+ Dẫn chứng:

(STT, Home, Products, Prod_A, Prod_B, Prod_C, Cart, Purchase)

4,1,0,1,1,1,0,1,1

9,1,0,1,1,1,1,1,1

10,1,0,1,1,1,1,1,0

- **Cluster 5**:

+ Tỉ lệ người xem từng loại sản phẩm (A, B, C) = (0, 1, 0.6786). Mỗi người đều xem sản phẩm B.

+ Xu hướng thanh toán ở mức thấp: 0.3214

+ Dẫn chứng:

(STT, Home, Products, Search, Prod_A, Prod_B, Prod_C, Cart, Purchase)

85,0,1,0,0,1,1,0,0

86,0,1,0,0,1,1,0,0

87,0,1,0,0,1,1,1,1

- **Cluster 6**:

+ Tỉ lệ người xem từng loại sản phẩm (A, B, C) = (1, 0.2857, 0). Mỗi người đều xem sản phẩm A.

+ Xu hướng thanh toán ở mức cao: 0.8571

+ Dẫn chứng:

(STT, Home, Products, Prod_A, Prod_B, Prod_C, Cart, Purchase)

72,0,0,0,1,0,0,1,1

75,0,1,0,1,0,0,1,1

76,0,0,0,1,0,0,1,0

- **Người dùng tìm kiếm** (sử dụng chức năng search để tìm sản phẩm cần mua)

  Cluster 3, cluster 4 thỏa chức năng search = 1.

  - **Cluster 3:**

    + Xu hướng thanh toán ở mức trung bình: 0.5

    + Dẫn chứng:

    (STT, Home, Products, Prod_A, Prod_B, Prod_C, Cart, Purchase)

    4,1,0,1,1,1,0,1,1

    9,1,0,1,1,1,1,1,1

    10,1,0,1,1,1,1,1,0

  - **Cluster 4:**

    + Xu hướng thanh toán ở mức cao: 0.7143

    + Dẫn chứng:

    (STT, Home, Products, Prod_A, Prod_B, Prod_C, Cart, Purchase)

    17,1,0,1,0,0,0,1,0

    20,1,1,1,0,0,1,1,1

    21,1,0,1,0,0,1,1,1

5. **Từ kết quả gom cụm mà bạn chọn, có cụm nào thể hiện sở thích mua hàng cụ thể của người dùng đối với sản phẩm đơn lẻ hay nhóm các sản phẩm hay không? Nếu có, nhận diện đặc điểm hành vi duyệt trang và xu hướng thanh toán của những người dùng trong nhóm này. Dẫn chứng cụ thể bằng một số mẫu đại diện.**

Có 3 cụm: 4,6,7 thể hiện sở thích mua hàng cụ thể của người dùng với sản phẩm đơn lẻ và nhóm sản phẩm.

Cụm thể hiện sở thích người dùng với sản phẩm đơn lẻ: là cụm 5, 6

- **Cụm 5:**

+ Tỉ lệ người xem từng loại sản phẩm (A, B, C) = (0, 1, 0.6786). Tất cả mọi người ở cụm này đều xem sản phẩm B và không ai xem sản phẩm A, tỉ lệ xem sản phẩm C hơn mức trung bình.

+ Hành vi duyệt trang: (Home, Product, Search) = (0, 1, 0.3214). Tất cả người ở cụm này sẽ vào Product xem mà không qua trang chủ Home, tỉ lệ tìm kiếm Seach cho sản phẩm thấp.

+ Xu hướng thanh toán ở mức thấp: 0.3215

+ Dẫn chứng:

(STT, Home, Products, Prod_A, Prod_B, Prod_C, Cart, Purchase)

39,0,1,0,0,1,0,0,0
40,0,1,0,0,1,0,0,0
41,0,1,0,0,1,1,0,0

Cụm 6:

+ Tỉ lệ người xem từng loại sản phẩm (A, B, C) = (1, 0.2857, 0). Tất cả mọi người đều xem sản phẩm A và không ai xem sản phẩm C.

+ Hành vi duyệt trang: (Home, Product, Search) = (0.1429, 0.1429, 1). Tất cả mọi người ở cụm này sẽ vào trang Search để tìm kiếm, có một số ít vào thông qua trang Home, và trang Product.

+ Xu hướng thanh toán ở mức cao: 0.8571

+ Dẫn chứng:

(STT, Home, Products, Prod_A, Prod_B, Prod_C, Cart, Purchase)

72,0,0,0,1,0,0,1,1
75,0,1,0,1,0,0,1,1
76,0,0,0,1,0,0,1,0

Cụm thể hiện sở thích người dùng với nhóm sản phẩm: Cụm 3

- **Cluster 3:**

+ Tỉ lệ người xem từng loại sản phẩm (A, B, C) = (1, 1, 0.75). Mỗi người xem ít nhất là 2 sản phẩm A và B, và tỉ lệ xem thêm sản phẩm C khá cao.

+ Hành vi duyệt trang: (Home, Product, Search) = (1, 0, 1). Tất cả mọi người ở cụm này sẽ vào trang Home, và trang Seach để tìm kiếm, không ai vào trang Product.

+ Xu hướng thanh toán ở mức trung bình: 0.5

+ Dẫn chứng:

(STT, Home, Products, Prod_A, Prod_B, Prod_C, Cart, Purchase)

4,1,0,1,1,1,0,1,1

9,1,0,1,1,1,1,1,1

10,1,0,1,1,1,1,1,0

6. **Giả sử rằng ABC đã đặt các banner quảng cáo lên một số trang nổi tiếng khác và những banner này trỏ trực tiếp đến trang của các sản phẩm A và B. Bạn có thể nhận diện cụm nào tương ứng với người dùng bị quảng cáo thu hút (tức là những người đến trực tiếp trang sản phẩm thay vì duyệt từ trang Home) hay không? Nếu có thể, cho biết chiến dịch quảng cáo cho sản phẩm nào thành công hơn**

Những cụm người dùng bị quảng cáo thu hút ứng với Home = 0 là cụm 5, 6. Ta xét trên 2 sản phẩm A và B.

- Cụm 5: (A, B) = (0, 1). Đây là cụm có chiến dịch quảng cáo cho sản phẩm B, nhưng với tỉ lệ thanh toán khá thấp 0.3214.
- Cụm 6: (A, B) = (1, 0.2857). Đây là cụm có chiến dịch quảng cáo cho sản phẩm A, và có 1 số ít sản phẩm B, với tỉ lệ thanh toán khá cao 0.8571.

Dựa vào tỉ lệ thanh toán ở cụm 6 cao hơn cụm 5, ta thấy rắng chiến dịch quảng cáo cho sản phẩm A thành công hơn sản phẩm B.

B. Thực hành: Cài đặt thuật toán K-mean:

Sinh viên chạy chương trinh cài đặt với tập dữ liệu sessions.csv. Đối chiếu kết quả phát sinh với kết quả của Weka (với k = 3 tới 8).

Cách chạy: Người dùng vào cmd và nhập câu lệnh:

```
python 1512353_1612815.py sessions.csv model.txt assignments.csv k
```

Với k là số cụm người dùng nhập vào.

K = 3:

```
Within cluster sum of squared errors: 137.363119
Cluster centroids:                                Cluster#
Attribute       0              1              2
                (39)           (32)           (29)
=============================================================
Home            0.666700       0.781200       0.310300
Products        0.666700       0.593800       0.931000
Search          0.564100       0.593800       0.069000
Prod_A          0.410300       0.843800       0.344800
Prod_B          0.564100       0.125000       1.000000
Prod_C          0.256400       0.468800       0.689700
Cart            0.179500       1.000000       0.758600
Purchase        0.000000       0.718800       0.551700
```

K = 4:

Within cluster sum of squared errors: 124.723300

Cluster centroids: Cluster#

Attribute	0 (30)	1 (13)	2 (34)	3 (23)
Home	0.933300	0.769200	0.088200	0.826100
Products	0.600000	0.153800	0.941200	0.869600
Search	0.633300	0.846200	0.352900	0.043500
Prod_A	0.733300	0.538500	0.088200	0.913000
Prod_B	0.233300	0.538500	1.000000	0.304300
Prod_C	0.700000	0.153800	0.617600	0.043500
Cart	0.566700	0.923100	0.411800	0.782600
Purchase	0.200000	0.769200	0.264700	0.608700

K = 5:

Within cluster sum of squared errors: 112.994491

Cluster centroids: Cluster#

Attribute	0 (10)	1 (27)	2 (14)	3 (34)	4 (15)
Home	1.000000	0.185200	0.857100	0.941200	0.066700
Products	0.500000	0.888900	0.857100	0.588200	0.733300
Search	0.700000	0.481500	0.642900	0.382400	0.066700
Prod_A	1.000000	0.111100	0.714300	0.735300	0.333300
Prod_B	0.800000	1.000000	0.000000	0.235300	0.800000
Prod_C	1.000000	0.481500	1.000000	0.000000	0.533300
Cart	0.600000	0.222200	0.857100	0.647100	1.000000
Purchase	0.100000	0.000000	0.571400	0.441200	1.000000

K = 6:

Within cluster sum of squared errors: 100.063505

Cluster centroids: Cluster#

Attribute	0 (15)	1 (30)	2 (23)	3 (10)	4 (11)	5 (11)
Home	0.066700	0.966700	0.739100	0.200000	0.818200	0.181800
Products	1.000000	0.633300	0.565200	0.700000	0.636400	1.000000
Search	0.400000	0.366700	0.304300	0.400000	1.000000	0.363600
Prod_A	0.000000	0.733300	0.869600	0.300000	0.727300	0.000000
Prod_B	1.000000	0.266700	0.391300	1.000000	0.181800	1.000000
Prod_C	0.666700	0.366700	0.087000	0.000000	1.000000	1.000000
Cart	1.000000	0.400000	1.000000	0.000000	1.000000	0.000000
Purchase	0.600000	0.000000	1.000000	0.000000	0.636400	0.000000

K = 7:

Within cluster sum of squared errors: 99.283848

Cluster centroids:

Attribute	0 (12)	1 (32)	2 (13)	3 (18)	4 (11)	5 (8)	6 (6)
Home	0.583300	0.468800	0.769200	0.833300	0.000000	0.875000	1.000000
Products	0.000000	0.937500	1.000000	1.000000	1.000000	0.000000	0.000000
Search	0.583300	0.312500	0.461500	0.500000	0.181800	0.625000	0.666700
Prod_A	0.750000	0.437500	0.461500	0.888900	0.000000	1.000000	0.000000
Prod_B	0.333300	0.812500	0.846200	0.000000	1.000000	0.250000	0.166700
Prod_C	0.250000	0.562500	0.000000	0.555600	0.909100	0.500000	0.000000
Cart	1.000000	0.093800	1.000000	1.000000	1.000000	0.375000	0.166700
Purchase	1.000000	0.000000	0.461500	0.666700	0.818200	0.000000	0.000000

K = 8:

Within cluster sum of squared errors: 87.949234

Cluster centroids:

Attribute	0 (15)	1 (21)	2 (17)	3 (16)	4 (11)	5 (11)	6 (9)
Home	0.866700	0.000000	0.529400	1.000000	1.000000	1.000000	0.000000
Products	0.666700	1.000000	0.882400	1.000000	0.545500	0.272700	0.111100
Search	0.733300	0.047600	1.000000	0.062500	0.090900	1.000000	0.111100
Prod_A	1.000000	0.047600	0.176500	0.875000	0.818200	0.181800	1.000000
Prod_B	0.133300	1.000000	0.941200	0.625000	0.090900	0.090900	0.444400
Prod_C	0.866700	0.761900	0.529400	0.187500	0.090900	0.272700	0.000000
Cart	0.933300	0.476200	0.294100	1.000000	0.000000	0.818200	0.777800
Purchase	0.533300	0.428600	0.000000	0.625000	0.000000	0.545500	0.666700

Nhìn chung có trường hợp cho độ lỗi lớn hơn (k = 3, 4), có trường hợp độ lỗi nhỏ hơn weka (k = 5, 6, 7, 8), nguyên nhân là do khởi tạo các cụm.

Tham Khảo:

[1] Slide bài giảng lý thuyết lý thuyết

[2] Trang chủ của WEKA:
http://www.cs.waikato.ac.nz/ml/weka/

[3] Đọc file csv: https://realpython.com/python-csv/

[4] Thuật toán Kmean: https://en.wikipedia.org/wiki/K-means_clustering

[5] Thuật toán Kmean:
https://machinelearningcoban.com/2017/01/01/kmeans/