군집분석

p개의 변수로 구성된 N개의 개체(항목)를 유사성(similarity)이 높은 개체들로 여러개의 군집으로 나누는 통계적 방법. (개체간 유사성이 높다 ≡ 개체간 거리가 가깝다)

①계층적(hierachical) 군집:
분할적(divisisive) 방법과 병합적(agglomerative) 방법(Start with n clusters, SPSS).
②비계층적 군집방법: eg: k-means방법

개체간(u,v) 거리:

유클리디안 거리(Euclidean): $d_{uv} = \sqrt{\sum_{k=1}^{p}(u_k - v_k)^2}$

제곱 유클리디안 거리: $d_{uv} = \sum_{k=1}^{p}(u_k - v_k)^2$

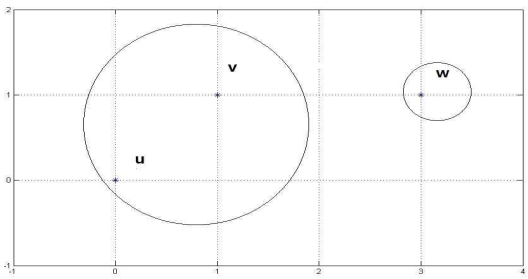
블록거리: $d_{uv} = \sum_{k=1}^{p}|u_k - v_k|$

군집간((u,v),w) 거리:

평균연결법(average linkage): $d_{(u,v)w} = \frac{1}{2}(d_{uw} + d_{vw})$(집단간),

$$d_{(u,v)w} = \frac{1}{3}(d_{uv} + d_{uw} + d_{vw})$$(집단내)

최단연결법(single linkage): $d_{(u,v)w} = \min(d_{uw}, d_{vw})$(가장가까운 항목)
최장연결법(complete linkage): $d_{(u,v)w} = \max(d_{uw}, d_{vw})$(가장먼 항목)
중심연결법(centroid linkage): $d_{(u,v)w} = d_{\overline{uv},w}$ (유클리드 거리)
eg) N=3, p=2



개체간 거리=제곱유클리디안 거리(Euclidean):

$d_{uv} = \sum_{k=1}^{2}(u_k - v_k)^2 = (0-1)^2 + (0-1)^2 = 2$, $d_{uw} = \sum_{k=1}^{2}(u_k - w_k)^2 = (0-3)^2 + (0-1)^2 = 10$

$d_{vw} = \sum_{k=1}^{2}(v_k - w_k)^2 = (1-3)^2 + (1-1)^2 = 4$

군집간 거리:

평균(집단간)연결법: $d_{(u,v)w} = \frac{1}{2}(d_{uw} + d_{vw}) = \frac{1}{2}(10+4) = 7$,

평균(집단내)연결법: $d_{(u,v)w} = \frac{1}{3}(d_{uv} + d_{uw} + d_{vw}) = \frac{1}{3}(2+10+4) = 5.333$

최단(최근접이웃))연결법: $d_{(u,v)w} = \min(d_{uw}, d_{vw}) = \min(10,4) = 4$
최장(가장먼이웃)연결법: $d_{(u,v)w} = \max n(d_{uw}, d_{vw}) = \max(10,4) = 10$
중심연결법: $\overline{uv} = (u+v)/2 = ((0,0)+(1,1))/2 = (0.5,0.5)$

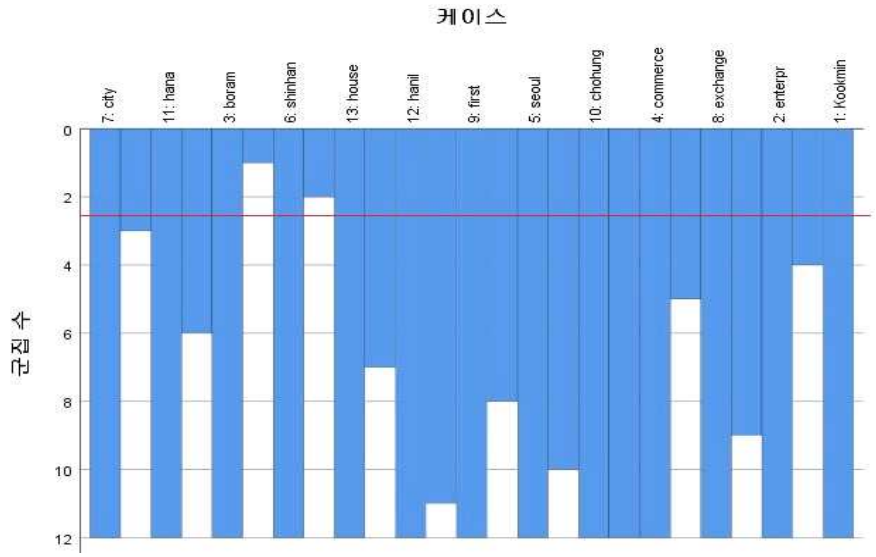
$d_{(u,v)w} = d_{\overline{uv},w}$ = $(0.5-3)^2+(0.5-1)^2$=6.5

eg bank.sav)
분석-분류분석-계층적 군집(H)
방법: 집단간 연결 & 제곱유틀리디안 거리

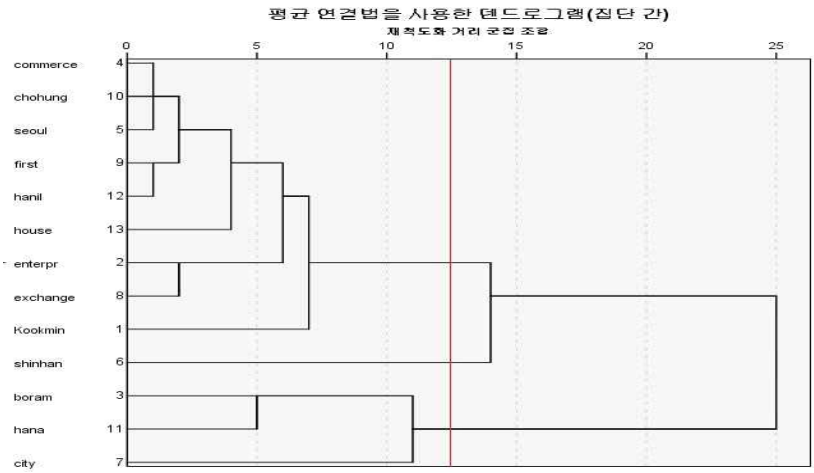
군집화 일정표

단계	결합 군집		계수	처음 나타나는 군집의 단계		다음 단계
	군집 1	군집 2		군집 1	군집 2	
1	4	10	17.470	0	0	3
2	9	12	30.300	0	0	5
3	4	5	32.275	1	0	5
4	2	8	51.480	0	0	8
5	4	9	51.490	3	2	6
6	4	13	73.710	5	0	8
7	3	11	96.200	0	0	10
8	2	4	105.013	4	6	9
9	1	2	137.226	0	8	11
10	3	7	192.520	7	0	12
11	1	6	250.771	9	0	12
12	1	3	455.023	11	10	0

12개에서 단계1에서 (4,10), (1),(2),(3), … (12)의 11개 군집으로. (min distance)
→ 단께2에서 (4,10), (9,12), (1),(2),(3), … (12)의 10개 군집으로 …



수직고드름도표(Vertical Icicle plot): 빈공간을 기준으로 군집경계 만들기
For k(군집수)=3 ~ (7, 11, 3), (6), 나머지:(13,12, …, 1) (빈공간 2개 포함하는 수평 cut)



For k=3 ~ (3,11,7), (6), (나머지)

저장(A)에서 원하는 군집수 지정.

	상호	편리성	신속성	친절	능률	패적	자동화	CLU3_1
1	Kookmin	71	59.4	63.7	54.3	66.9	62.6	1
2	enterpr	65	70.3	68.6	55.2	68.0	64.1	1
3	boram	67	79.6	78.5	62.4	79.8	62.4	2
4	commerce	61	65.0	65.6	54.4	64.5	63.9	1
5	seoul	63	66.5	67.9	56.0	59.7	62.0	1
6	shinhan	72	69.1	74.2	60.0	70.1	68.2	3
7	city	64	72.0	71.4	56.9	72.8	57.8	2
8	exchange	68	67.5	67.3	51.3	71.3	65.8	1
9	first	66	66.5	67.3	50.7	63.4	63.3	1
10	chohung	64	65.7	64.3	53.9	61.7	62.7	1
11	hana	69	74.3	80.5	63.6	75.7	55.9	2
12	hanil	63	65.5	68.3	49.8	64.6	59.1	1
13	house	64	64.8	67.8	59.7	65.7	61.8	1

다차원척도법(multidimensional scaling)

개체 간의 근접도 척도 집합 구조를 탐색. 이를 위해 개념적 저차원 공간에 있는 점 사이의 거리가 주어진 유사성 혹은 상이성 측도와 가능한 한 가깝게 일치되도록 이 공간의 특정 위치에 관측값이 할당된다. 그 결과 해당 저차원 공간에 개체의 최소제곱 표현이 얻어지며, 이 표현을 통해 데이터를 더 심도 있게 이해할 수 있다.

softdrink_mds.sav: 원자료 (58명, 1~7점)

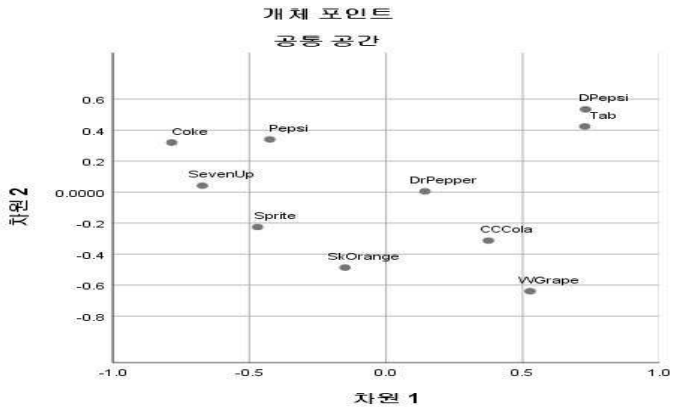
A: Flat~Fizzy, B:Expensive~inExpensive, C:hard to find~ easy

D: Bland~Flavorable, E: Lack of aftertaste~presence of aftertaste

F: Not sweet ~ sweet, G: Not satisfy thirst~satisfy thirst

분석-척도분석-다차원척도법(PROXSCAL) ~데이터형식-데이터로부터 근접행렬 작성

모형(M)에서



cities.sav: 9대 도시의 정치적성향 거리(상이성), 거리행렬

~데이터형식-데이터가 근접행렬 (default)