응용통계연구 제 5권 2호
1992년 pp 211-227

집락분석법에 있서서 비유사도 와 계층적 응집법의 관계에 관한 연구1)

조 완 현2)

요 약

본논문은 지금까지 집락분석방법에서 많이 사용되는 유사도 또는 비유사도의 정의 및 이들의 수학적 성질을 알아보고, 또한 이들에 의해서 생성되는 비유사도 행렬을 이용하여 계층적 집락분석을 실시하였다. 이 경우에 가정된 초기의 집락구조를 정확하게 잘 재생시킬 수 있는 비유사도의 측정방법과 계층적 응집법의 상호 관계를 질적자료와 양적자료 각각에 대하여 고찰하고, 이들에 관련된 시뮬레이션 결과를 제시하였다.

1. 서론

집락분석(cluster analysis)은 측정된 변수들의 값을 사용하여 특정한 법칙에 따라 개체들을 비슷한 집락으로 분류시키는 방법을 말하며, 이때 각 집락안에서 개체들은 동질한 성격을 가지며 동시에 다른 집락간의 개체들은 가능한 한 서로 이질적 성격을 갖도록 분류하는 방법을 총칭하는 것으로 수치분석법(numerical classification), 패턴인식(pattern recognition) 등의 여러가지 이름으로 불리워지고 있다.

일반적으로 집락분석에 있어서 많은 경우에 주어진 개체간의 집락구조가 특정한 유사도와 집락알고리즘에 따라서 다르게 표현될 수 있다. 따라서 우리는 주어진 개체간의 유사성이나 비유사성을 측정할 수 있는 유사도 또는 비유사도의 다양한 측정방법과 그들의 성질을 고찰해 보고, 여러가지 계층적 집락알고리즘들의 성질을 요약하였으며, 또한 주어진 초기의 집락구조를 재생시키는데 이들 두가지 요인들이 어떻게 영향을 미치는가를 고찰하고자 한다. 즉 특별히 집락구조의 재생에 영향을 미치는 요인들을 시뮬레이션을 통하여 알아보고, 이들 상호간에 관련성을 생각하여 보았다.

본 논문은 4 개의 절로 구성되어 있으며 제2절은 유사도 또는 비유사도의 측정방법을 고찰하였고, 제3절에서는 여러가지 계층적 집락알고리즘을 제안하고, 동시에 두 집락구조간의 유사성을 측정하는데 사용될 수 있는 두가지 통계량을 정의하고, 제4절

1) 본 연구는 1990년도 학술진흥재단 자유공모과제 연구비에 의해서 수행되었음
2) 500-757 광주시 북구 용봉동 300 전남대학교 자연과학대학 통계학과.

에서는 몬테칼로 시뮬레이션 절차를 제시하였으며, 제5절에서는 분석결과를 정리하였다.

2. 유사도 또는 비유사도 측정

우리들은 대부분의 경우에 관습적으로 p개의 변수에 대하여 측정된 n개의 개체들을 벡터 $x_i=(x_{i1},\cdots,x_{ip})$, $i=1,\cdots,n$ 로 많이 표현한다. 이때 많은 집락분석 기법은 주어진 개체들로 부터 두 개체간의 떨어진 정도(distance)나 이들 서로간의 유사성(proximity)의 크기를 측정하는 과정을 포함하고 있다.

먼저 유사도(similarity)는 두 개체사이의 유사성 또는 관련성을 표현하는 측도로서, 임의의 두 개체 x_i 와 x_j 의 어떤함수 $s_{ij} = f(x_i,x_j)$로 정의되며 다음의 조건을 만족한다.

① $0 \le s_{ij} \le 1$, ② $s_{ij} = s_{ji}$, ③ $x_i = x_j \Rightarrow s_{ij} = 1$

또한 유사도 s_{ij}가 다음의 두 조건을 더 만족하면 이것을 거리(metric) 이라 한다

④ $s_{ij} = 1 \Rightarrow x_i = x_j$, ⑤ $(s_{ij} + s_{jk}) \cdot s_{ik} \ge s_{ij} \cdot s_{jk}$

그리고 많은 경우에 유사도의 역개념으로 두 개체간의 거리의 측도로 비유사도(dissimilarity) $d_{ij} = 1 - s_{ij}$ 를 정의할 수 있다. 그때 이것은 유사도의 성질로 부터 다음의 조건을 만족한다.

① $0 \le d_{ij} \le 1$, ② $d_{ij} = d_{ji}$, ③ $x_i = x_j \Rightarrow d_{ij} = 0$

또한 이렇게 정의되는 비유사도 d_{ij}가 다음의 두 조건을 더 만족하면 이것을 거리이라 부른다.

④ $d_{ij} = 0 \Rightarrow x_i = x_j$, ⑤ $d_{ij} + d_{jk} \ge d_{ik}$

2.1 질적자료에 대한 유사도의 측정

먼저 가장 단순한 유사도는 각 변수가 오직 두가지 특성값 만을 가지는 이진수(dichotomous)변수에 관한 것이다. 즉 p개의 성분을 가지는 임의의 두 개체 벡터 $x_i=(x_{i1},\cdots x_{ip})$ 와 $x_j=(x_{j1},\cdots,x_{jp})$ 에서 이들 각각의 성분의 값은 0 또는 1를 취할 때, 이들 두 개체간의 유사도를 정의하기 위해서는 먼저 아래와 같은 (2×2) 분할표(association table)가 필요하다.

표 1. 두 개체 벡터에 대한 2X2 분할표

x_i \ x_j	1	0	합계
1	a	b	a+b
0	c	d	c+d
합계	a+c	b+d	p

그러면 우리는 이들 4가지 요소 a,b,c,d 들을 적절히 결합하여 두 개체벡터 사이의 여러가지 유사도를 다음과 같이 정의할 수 있다.

표 2. 질적자료에 대한 유사도

거리의 성질	0-0 대응의 분자, 분모에 포함여부	유 사 도		
만족함	포함하지 않음	1. Jaccard $\frac{a}{a+b+c}$	2. Anderberg $\frac{a}{a+2(b+c)}$	3. Russel and Rao $\frac{a}{a+b+c+d}$
	포함함	4. Simple matching $\frac{a+d}{a+b+c+d}$	5. Rogers-Tanimoto $\frac{a+d}{a+2(b+c)+d}$	6. Hamman $\frac{a+d-(b+c)}{a+b+c+d}$
만족하지 않음	포함하지 않음	7. Dice and Cze $\frac{a}{a+\frac{1}{2}(b+c)+d}$	8. Kulczynski $\frac{1}{2}\left(\frac{a}{a+b}+\frac{a}{a+c}\right)$	9. Ochiai $\frac{a}{\sqrt{(a+b)(a+c)}}$
	포함함	10. Sneath and Sokal $\frac{a+d}{a+\frac{1}{2}(b+c)+d}$	11. Anderberg $\frac{1}{4}\frac{a}{a+b}+\frac{a}{a+d}+\frac{d}{c+d}+\frac{d}{b+d}$	
		12. $\frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$	13. Yale $\frac{ad-bc}{ad+bc}$	
		14. Pearson's ϕ $\frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$		

두번째로, 변수들이 두개 이상의 수준을 가지는 명명변수(nominal variable)의 유사도는 k번째 성분에 대하여 적절한 스코어를 배정하는 방법으로 정의할 수 있다. 예를 들어서 스코어 함수 $s_k(x_{ik}, x_{jk})$를 다음과 같이 정의하자.

$$S_k(X_{ik},X_{jk})=\begin{cases}1, & X_{ik}=X_{jk}\\ & k=1,\cdots,p\\ 0, & X_{ik}\neq X_{jk}\end{cases}$$

이때 두 벡터의 유사도는 모든 변수에 관하여 스코어들의 평균으로 정의된다.

$$S_{ij}=\frac{1}{p}\sum_{k=1}^{p}s_k(x_{ik},x_{jk})$$

또한 보다 일반적으로 우리는 k 번째 성분에 대하여 사전의 중요성이나 신뢰성을 반영하는 가중 w_k를 사용할 수 있다. 즉, 가중함수 $w_k(x_{ik},x_{jk})$를 k번째 성분에 대응시키면, 이 경우에 두 개체간의 유사도는 다음과 같다.

$$S_{ij}=\frac{\sum_{k=1}^{p}w_k(x_{ik},x_{jk})s_k(x_{ik},x_{jk})}{\sum_{k=1}^{p}w_k(x_{ik},x_{jk})}$$

세번째로, 자료가 오직 순서변수만을 포함하는 경우에 대하여는 각 성분 x_{ik} 와 x_{jk} 가 $x_{ik} \geq x_{jk}$ 이거나 또는 $x_{ik} < x_{jk}$이고, 이들의 최소값 $\min(x_{ik},x_{jk})$이 의미를 갖는다. 따라서 두 개체 벡터 x_i 와 x_j 의 비유사도 $d_0(x_i,x_j)$를 다음과 같이 정의할 수 있으며,

$$d_0(x_i,x_j)=\frac{\sum_{k=1}^{p}x_{ik}+\sum_{k=1}^{p}x_{jk}-2\sum_{k=1}^{p}\min(x_{ik},x_{jk})}{\sum_{k=1}^{p}x_{ik}+\sum_{k=1}^{p}x_{jk}-\sum_{k=1}^{p}\min(x_{ik},x_{jk})}$$

이 정의는 이진자료에 대한 Jccard 유사도로 부터 유도되는 비유사도에 대응한다.

2.2 양적자료에 대한 비유사도

각 변수가 양적으로 측정될 수 있을때는 두 개체간의 거리를 비유사도로 측정하는 것이 유사도보다 자연스러운 일이다. 이때 사용가능한 비유사도 계수들이 다음표에서 주어졌고, 또한 비유사도 D_2, D_3, D_4 에서 사용되는 r_k 는 각 변수들의 서로 다른 척도의 효과를 제거하기 위해서 도입되는 표준화 상수이다. r_k 에 대한 논리적 제약조건은 모든 변수 X_k가 똑같은 측정단위를 갖도록 취하는 것이다. 보통 사용되는 방법은 r_k 를 X_k 의 표본표준편차나 표본범위로 취한다.

표 3. 양적 자료에 대한 비유사도

거리의 성질	종	류
만족함	1. Euclidean $\frac{1}{p}\sum_{k=1}^{p}(x_{ik}-x_{jk})^2$	2. Standardized Euclidean $\frac{1}{p}\sum_{k=1}^{p}(x_{ik}-x_{jk})^2/r_k^2$
	3. Minkowski $\frac{1}{p}\sum_{k=1}^{p}\lvert x_{ik}-x_{jk}\rvert^t/r_k^t$	4. Canberra $\sum_{k=1}^{p}\frac{\lvert x_{ik}-x_{jk}\rvert}{\lvert x_{ik}\rvert+\lvert x_{jk}\rvert}$
만족하지않음	5. Divergence $\frac{1}{p}\sum_{k=1}^{p}\frac{(x_{ik}-x_{jk})^2}{(x_{ik}+x_{jk})^2}$	6. $\frac{1}{p}\sum_{k=1}^{p}\frac{\lvert x_{ik}-x_{jk}\rvert}{\lvert x_{ik}+x_{jk}\rvert}$
	7. Ware and Hedges $\frac{1}{p}\sum_{k=1}^{p}\{1-\frac{\min(x_{ik},x_{jk})}{\max(x_{ik},x_{jk})}\}$	
	8. Soergel $\frac{\sum_{k=1}^{p}\lvert x_{ik}-x_{jk}\rvert}{\sum_{k=1}^{p}\max(x_{ik},x_{jk})}$	9. Bray and Curtis $\frac{\sum_{k=1}^{p}\lvert x_{ik}-x_{jk}\rvert}{\sum_{k=1}^{p}(x_{ik}+x_{jk})}$

3. 계층적 응집법의 종류과 두 집락구조의 유사성을 측도할 수 있는 통계량

계층적 응집법에 의한 집락분석은 n 개의 개체에 p 개의 변수의 값이 주어진 $(n\times p)$ 의 자료행렬로 부터 개체상의 유사성을 표시하는 비유사도, d_{ij} 를 계산하고, 이것을 이용하여 수형도 (dendrogram)이라고 부르는 나무가지 모양의 계층적 구조를 구성하는 것을 그 목적으로 하는 분석방법이다.

이러한 응집형 집락구조를 형성할때 사용되는 알고리즘은 먼저 n 개의 개체로부터 비유사도 행렬 $D = (d_{ij})$ 를 계산하고, 이들로 부터 각 단계 k (k=1,..,n-1)에서 두 집락 C_p 와 C_q 간의 거리가 최소로 되는 집락을 병합하여 새로운 집락 $C_t = C_p \cup C_q$ 을 만든다. 다음 단계 k+1 에서 새로운 집락 C_t 와 다른 임의의 집락 C_r 간의 비유사도를 적절한 갱신공식을 사용하여 다시 계산한다. 이와같은 방법을 한개의 집락이 얻어질때 까지 계속한다. 이경우 병합된 새로운 집락 C_t 와 다른집락 C_r 간의 비유사도를 갱신하는 방법에 따라 다음과 같은 여러가지 분석방법을 고려할 수 있다.

Lance 와 Williams(1967)는 지금까지 제안된 알고리즘들을 다음과 같은 공통의 한 식으로 표현하였다.

$$d_{tr} = \alpha_p d_{pr} + \alpha_q d_{qr} + \beta d_{pq} + \gamma d_{pr} - d_{qr}$$

이때 이들 4개의 모수 $\alpha_p,\alpha_q,\beta,\gamma$ 들은 $\alpha_p + \alpha_q + \beta = 1$, $\alpha_p = \alpha_q$, $\beta < 1$, $\gamma = 0$ 의 조건을 만족하는 범위에서 임의의 값을 취하도록 제약조건을 적용하되, 이 중 그들에 의하여 추천되는 적절한 값은 $\alpha_p = \alpha_q = 0.5$, $\beta = -0.25$ 이다.

지금까지 사용되어 온 여러가지 방법들에 대하여 위의 모델에 의해서 모수의 값을 계산하여 열거한 표는 다음과 같다. 또한 이들 방법들을 공간의 수축(space-contracting), 보존(conserving) 또는 확장(dilating)의 개념에서 분류한 결과도 함께 표시하였다.

표 4. 계층적 집락분석법에 대응하는 4 개의 모수의 값

방법	조합적 방법의 모수				공간의 수축및 확장	기 타
	α_p	α_q	β	γ		
최단거리	1/2	1/2	0	- 1/2	수축	
최장거리	1/2	1/2	0	1/2	확장	
군 평균	n_p/n_t	n_q/n_t	0	0	보존	
중 심	n_p/n_t	n_q/n_t	$-n_p n_q/n_t^2$	0	보존	
메디안	1/2	1/2	- 1/4	0	보존	
워 드	$\frac{n_q+n_r}{n_t+n_r}$	$\frac{n_q+n_r}{n_t+n_r}$	$-\frac{n_r}{n_t+n_r}$	0	확장	
가 변	$(1-\beta)/2$	$(1-\beta)/2$	$\beta < 1$	0	$\beta<0$, 확장 $\beta=0$, 보존 $\beta>0$, 수축	

두가지 이상의 집락 분석방법들의 재생(retrieval)능력을 평가하기 위해서 사용될 수 있는 두개의 측도방법들을 정의하여보자.

먼저 n개의 자료점들로 구성하는 초기의 집락구조를 Y로 나타내고 , 또한 주어진 자료점들에 대하여 특정한 유사도 와 분석방법을 적용하여 얻어지는 새로운 집락구조를 Y′으로 나타내면, 그때 이들 두 집락구조 Y 와 Y′의 유사성을 계산하는 측도로 Rand(1971)는 다음과 같은 C-통계량을 제안하였다.

(정의1) 똑같은 n개의 점들로 구성되는 두 계층적 집락구조 Y 와 Y′에 대하여, 각 계층 k (k=2, .. ,n-1)에서 n_{ij} 를 Y 의 i번째 집락 C_i 와 Y′의 j번째 집락 C_j에 속하는 개체의 수라 하면, 그때 우리는 다음과 같은 (k × k) 행렬

$$N = [\, n_{ij} \,] \quad (i, j = 1, \ldots, k)$$

을 얻게 된다. 그러면 Rand 의 통계량 C_k는

$$C_k(Y,Y') = [T_k - P_k/2 - Q_k/2 + {}_nC_2] / {}_nC_2$$

으로 주어진다. 여기서

$$T_k = \sum_i^k \sum_j^k n_{ij} - n, \quad P_k = \sum_i^k n^2_{i\cdot} - n, \quad Q_k = \sum_j^k n^2_{\cdot j} - n$$

$$n_{i\cdot} = \sum_j^k n_{ij}, \quad n_{\cdot j} = \sum_i^k n_{ij}$$

이다.

또 다른 두 집락구조의 유사성의 측정방법은 Fowles 와 Mallow(1983)에 의해서 제안되었는데, 이것은 위에서 정의된 통계량 T_k 를 다음과 같이 다르게 척도화 한 것이다.

(정의 2) 각 계층 k (k = 2,...,n-1) 에서, 두 집락구조 Y 와 Y′ 의 유사성은

$$B_k = \frac{T_k}{\sqrt{P_k \cdot Q_k}}$$

이다. 여기서 T_k, P_k 와 Q_k 은 위에서 정의된 값들이다.

이때 통계량 B_k 와 C_k 의 기대값 과 분산 및 통계적 성질들은 위의 두 논문에 잘 요약되어 있다.

4. 시뮬레이션 연구 : 자료의 계획 및 생성

유사도 또는 비유사도의 측정방법과 계층적 집락알고리즘의 선택이 자료의 초기 집락구조를 재생하는데 어떠한 영향을 미치는가를 고찰하기 위해서, 먼저 질적자료와 양적자료 각각에 대하여 시뮬레이션 자료를 생성해 보자.

4.1 질적자료

먼저 k (k ≥ 2) 개의 집락구조를 가지는 n개의 p 차원 이진수 개체벡터 x_i, $i=1,\cdots,n$, 를 다음과 같이 생성한다.

① 우리는 이 경우 차원의 수를 p = 8로 택하였으며, 각 변수 x_{ij} (j=1,..8) 의 각 특성값이 0 또는 1 을 취하므로, 출현가능한 서로 다른 이진수 개체벡터의 총수는 2^8=256 개의 측정값이 가능하다. 이들을 x_1 = (0,0,...,0) 에서 x_{256} =(1,1,..,1) 까지 (256×8) 의 2차원 배열 로 구성하였다.

② 두개 이상인 k 개의 집락구조를 가지는 초기의 시뮬레이션 표본을 생성 하기 위해서, 256 개의 배열을 k 개의 군으로 나누고, 각 군에서 크기 10 인 표본을 난수를 추출하여 이들을 잘 섞어서 초기의 집락을 취한다. 예를 들어서 집락의 수가 k = 3 또는 4 인 경우 각 집락에 속하는 개체벡터들의 형태는 다음과 같이 취한다.

집락의수	3	4
개체벡터의 형태	(0,0,0,0,0,0,0,1) (0,0,0,0,1,1,1,1) (0,1,1,1,1,1,1,1)	(0,0,0,0,0,0,0,1) (0,0,0,0,0,1,1,1) (0,0,0,1,1,1,1,1) (0,1,1,1,1,1,1,1)

4.2 양적자료

양적자료의 성질을 고찰하는 경우에는 대부분 다변량 정규분포를 따르는 n 개의 개체벡터를 많이 사용하며, 이들은 다음과 같은 단계를 통하여 생성하였다.

① 서로 독립인 표준 정규난수 Z_1, Z_2를 IMSL의 부 프로그램으로 부터 생성한다.

② 이들 생성된 두개의 난수를 다음의 변환식 $X_1 = \sigma_1 \cdot Z_1$, $X_2 = \rho\,\sigma_1 \cdot Z_1 + \sigma_2(1-\rho^2)^{1/2} \cdot Z_2$ 을 사용하여 평균벡터가 (0,0) 이고, 공분산행렬

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

을 가지는 이변량 정규분포의 표본으로 바꾼다. 여기서 표준편차는 $\sigma_1 = \sigma_2$ = 1.0 이고, 상관계수는 ρ = 0.0 과 0.5으로 취한다.

③ 집락의 수가 k (k ≥ 2) 개인 초기의 집락구조를 가지는 표본을 생성하기 위하여, 다음의 평균벡터를 가지는 분포에서, 각각 크기 10인 표본을 추출하여 이들을 잘 섞어서 초기의 집락구조를 만든다.

집락의수	3	4
$\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}$	$\begin{pmatrix}0.0\\0.0\end{pmatrix}\begin{pmatrix}4.0\\0.0\end{pmatrix}\begin{pmatrix}0.0\\4.0\end{pmatrix}$	$\begin{pmatrix}0.0\\0.0\end{pmatrix}\begin{pmatrix}4.0\\0.0\end{pmatrix}\begin{pmatrix}0.0\\4.0\end{pmatrix}\begin{pmatrix}4.0\\4.0\end{pmatrix}$

다음으로 우리는 위에서 생성된 자료를 초기의 집락구조 Y 로 사용하여, 먼저 비유사도를 계산하고, 이들 각각의 비유사도에 대하여 7 가지의 계층적 집락분석법을 적용하여 최종적인 집락구조 Y'를 얻는다. 따라서 이들 두개의 집락구조 Y 와 Y' 대하여 통계량 B_k, C_k를 계산한다. 이러한 절차를 10 번 반복 측정하여 주어진 통계량의 값들을 집락의 수 k 에 대하여 두 요인 분산분석을 실시하였다. 본 논문에서는 통계량 C_k 의 값을 이용하여 분석한 결과만을 제시하였다. 또한 지금까지의 모든 계산은 Trigem 386V 를 사용하였으며, 일량난수와 표준정규난수는 개인용 컴퓨터 IMSL 을 이용하였다.

5. 분석결과

5.1 질적자료

유사도의 측정방법과 집락알고리즘의 종류가 초기의 집락구조의 재생에 대하여 어떤 영향을 주고, 또한 각 요인의 어떤 수준조합에서 통계량의 값을 최대로 하는가를 알아 보기 위해서, 측정된 자료에 대한 분산분석을 다음과 같이 수행하였다.

먼저 두 요인의 수준은 다음과 같이 정의하였다

A : 유사도의 측정방법

A_1 : Russell and Rao , A_2 : Jaccard , A_3 : simple matching

A_4 : Anderberg , A_5 : Rogers-Tanimoto, A_6 : Sorensen,Dice and Czekanowski , A_7 : Sneath and Sokal, A_8 : Hamman , A_9 : Ochiai

B : 집락알고리즘의 종류

B_1 : Single Linkage , B_2 : Complete Linkage , B_3 : Average Linkage , B_4 : Centroid , B_5 : Median , B_6 : Ward , B_7 : Flexible

실험은 초기의 집락의 수가 3 인 경우와 4 인 경우에 매 실험을 10회 반복 실시하여 이들 각각에 대하여 통계량 C_k를 계산하고, 측정된 자료에 대하여 분산분석표를 작성한 결과가 다음표에 주어져 있다.

표 5. 집락수가 3과 4인 질적자료의 분산분석표

요 인	SS		DF	ME		F		Pr > F
	3	4		3	4	3	4	
A	16.7796	0.0799	8	2.0975	0.0100	264.51	5.30	0.001
B	4.8323	10.2762	6	0.8054	1.7127	101.57	909.14	0.001
A × B	2.9894	0.1532	48	0.0623	0.0032	7.85	1.69	0.001
E	4.4962	1.0682	567	0.0079	0.0019			
T	29.0976	11.5774	629					

먼저 초기의 집락수가 3 인 경우는 각 집락간의 특성이 비교적 잘 구분되는 경우로 볼 수 있고, 이때는 집락 재생능력이 유사도의 종류 및 집락방법의 수준에 따라 매우 유의한 차가 있으나 유사도 정의 방법에 더 많은 영향을 받고 있음을 알 수 있고, 또한 A × B 교호작용도 대단히 유의하다.

두번째로 초기 집락의 수가 4 인 경우는 각 집락간의 구별이 비교적 뚜렷하지 않은 경우로 생각할 수 있으며, 이 경우는 집락재생능력이 집락알고리즘의 방법에 대해서는 많은 영향을 받지만 유사도의 정의방법에 대해서는 적은 영향을 받음을 알 수 있다. 또한 두 요인의 교호작용의 효과도 비교적 낮음을 알 수 있다.

따라서 유사도 측정방법과 집락알고리즘의 종류에 대한 보다 구체적인 특성을 파악하기 위해서 다중비교를 실시한 결과는 다음과 같다.

표 6. 유사도의 측정방법에 대한 다중비교 : 집락수가 3 인 경우

유사도의 종류	7	8	3	5	4	1	2	6	9
거리의 성질	X	O	O	O	O	O	O	X	X
0-0대응의포함여부	O	O	O	O	X	X	X	X	X
통계량의 평균	0.8145	0.8045	0.8044	0.7760	0.4882	0.4814	0.4763	0.4643	0.4523
Tukey 의 그룹	A	A	A	A	B	B	B	B	B

표 7. 유사도의 측정방법에 대한 다중비교 : 집락수가 4 인 경우

유사도의 종류	7	5	8	3	9	1	4	2	6
거리의 성질	X	O	O	O	O	O	O	X	X
0-0대응의포함여부	O	O	O	O	X	X	X	X	X
통계량의 평균	0.5077	0.4953	0.4939	0.4939	0.4908	0.4868	0.4739	0.4738	0.4730
Tukey 의 그룹	A	(A,B)	(A,B)	(A,B)	(A,B)	(A,B)	(B,B)	B	B

위의 결과로 부터 집락수가 3 인 경우는 초기집락의 재생능력이 각 유사도의 정의 방법에 따라 매우 차이가 있음을 알 수 있다. 즉 유사도의 정의에서 분자, 분모에 0 - 0 대응의 성분의 수를 포함하고 있는 유사도들이 대체적으로 집락 재생능력이 탁월하며, 거리 성질의 만족성은 관련이 그렇게 높지 않음을 알 수 있다. 반면에 집락의 수가 4 인 경우는 집락 재생능력에 영향을 많이 미치는 유사도의 종류에 따른 순위간에 특별한 차이를 느낄 수 없다.

표 8. 응집형 집락알고리즘에 대한 다중비교

3	집락방법의 종류	6	2	1	3	7	4	5
	공간의 수축 팽창	확장	확장	수축	보존	확장	보존	보존
	통계량의 평균	0.7937	0.6783	0.6070	0.6022	0.5858	0.5618	0.4971
	Tukey 의 그룹	A	B	C	C	(C,D)	D	D
4	집락방법의 종류	6	3	7	2	5	1	4
	공간의 수축 팽창	확장	보존	확장	확장	보존	수축	보존
	통계량의 평균	0.6227	0.6093	0.6091	0.5442	0.3534	0.3446	0.3307
	Tukey 의 그룹	A	A	A	B	C	(C,D)	D

또한 계층적 집락방법은 전반적으로 여러방법들 사이에 많은 차이를 보이지 않고 있지만, 그러나 집락간의 구별이 뚜렷할 때에는 유사도 측정방법에 관계없이 Ward 또는 Comple Linkage 등의 공간확장에 속하는 방법들이 일반적으로 좋은 응집법임을 알 수 있다. 또한 집락들간의 성격이 서로 혼합되는 경우에는 Ward, Average Linkage, Flexible 집락방법들이 비교적 집락재생능력이 탁월함을 알 수 있다.

끝으로 실험의 분석결과에 의해서 교호작용이 유의하므로, 집락재생능력을 가장 크게 하는 A × B의 수준조합은 A_7*B_6 이고, 이것은 Sneath-Sokal 의 유사도에 대한 Ward 방법을 적용한 것이다.

5.2 양적자료

양적자료에 대해서도 초기의 집락구조 재생에 대하여 유사도의 측정방법과 집락알고리즘의 종류가 어떤 영향을 미치는 가를 알아보기 위해서, 두 요인의 수준들을 다음과 같이 정의하였다

A : 비유사도의 측정방법

A_1 : Squard Euclidean , A_2 : City-block , A_3 : Minknowsky

A_4 : Canberra , A_5 : Divergence, A_6 : No name , A_7 : Soergel,

A_8 : Ware and Hedges , A_9 : Bary and Curtis

B : 집락알고리즘의 종류

B_1 : Single Linkage , B_2 : Complete Linkage , B_3 : Average Linkage ,

B_4 : Centroid , B_5 : Median , B_6 : Ward , B_7 : Flexible

먼저 집락의 수가 3 과 4 이고, 두 변량간의 상관계수 ρ = 0.0 인 경우를 생각하였다. 두 변량간의 상관계수가 ρ = 0.5 인 경우는 각 집락간의 개체들이 너무 혼합되어서 초기의 집락들간의 구별이 비교적 확실하지 않는 경우이고, 또한 분석결과가 ρ = 0.0 인 경우와 비슷하므로 고려 대상에서 제외하였다.

두번째로 집락의 수가 3 인 경우는 초기의 집락들간의 구별이 비교적 확실한 집락구조로 생각할 수 있고, 집락의 수가 4 인 경우는 각 집락간의 개체들이 서로 혼합되는 경우로 생각할 수 있다. 매 실험은 10회 반복 실시하여 이들 각각에 대하여 통계량 C_k를 계산하였다.

표 9. 집락수가 3 인 양적자료에 대한 통계량 C_k 의 값

종류 \ 방법	B_1	B_2	B_3	B_4	B_5	B_6	B_7
A_1	0.3632	0.9205	0.9317	0.9366	0.9274	0.9266	0.9172
A_2	0.3632	0.9333	0.9372	0.8754	0.7536	0.9455	0.9356
A_3	0.3632	0.9283	0.9306	0.9306	0.9280	0.9319	0.9131
A_4	0.3632	0.7032	0.6494	0.4018	0.4009	0.6931	0.7317
A_5	0.3632	0.6205	0.6124	0.6124	0.6186	0.6124	0.6172
A_6	0.3632	0.6175	0.6124	0.6140	0.5916	0.6161	0.6172
A_7	0.3632	0.7462	0.4703	0.4545	0.4853	0.6099	0.6202
A_8	0.3632	0.6232	0.6128	0.5876	0.5906	0.5910	0.5956
A_9	0.3632	0.8163	0.6514	0.4310	0.4331	0.4438	0.4472

표 10. 집락수가 4 인 양적자료에 대한 통계량 C_k 의 값

종류 \ 방법	B_1	B_2	B_3	B_4	B_5	B_6	B_7
A_1	0.3154	0.9076	0.9138	0.9174	0.8896	0.9203	0.9191
A_2	0.3154	0.9101	0.8991	0.8941	0.8429	0.9236	0.9223
A_3	0.3154	0.9127	0.9032	0.9055	0.9001	0.9240	0.9165
A_4	0.3154	0.8018	0.6022	0.3737	0.4538	0.7572	0.7528
A_5	0.3154	0.7082	0.6038	0.6038	0.6038	0.6858	0.6362
A_6	0.3154	0.7521	0.6054	0.6040	0.6038	0.7019	0.6863
A_7	0.3154	0.7881	0.4196	0.3973	0.4327	0.7382	0.6600
A_8	0.3154	0.7455	0.6679	0.5864	0.5937	0.6581	0.6437
A_9	0.3154	0.7972	0.6586	0.4512	0.4776	0.5738	0.5494

먼저 통계량 C_k 의 값을 고찰해보면, 이때 Single Linkage 방법에 대한 통계량의 값은 집락의 수가 3 인 경우나 4 인 경우에 똑 같이 9 가지 유사도의 종류에 관계없이 변하지 않고 있다. 따라서 우리는 Single Linkage 방법에 대한 통계량 C_k 의 값을 제외한 나머지 측정된 자료에 대하여 분산분석을 실시한 결과가 다음 표에 주어져 있다.

표 11. 집락수가 3 과 4인 양적자료의 분산분석표

요 인	SS		DF	ME		F		Pr>F
	3	4		3	4	3	4	
A	13.3266	10.1453	8	1.6658	1.2682	290.79	222.12	0.0001
B	1.0297	2.1944	5	0.2059	0.4389	35.95	76.87	0.0001
A × B	2.4026	2.2820	40	0.0601	0.0670	10.48	9.00	0.0001
E	2.7841	2.7748	486	0.0057	0.0057			
T	19.5430	17.3965	539					

위의 분석결과로 부터 집락재생능력이 비유사도의 측정방법및 집락알고리즘의 종류에 따라 매우 대단히 유의한 차가 있음을 알수 있고, 또한 A × B 교호작용도 대단히 유의하다. 따라서 우리는 유사도의 측정방법과 집락알고리즘의 종류에 대하여 다중비교를 실시하였다.

표 12. 비유사도의 측정방법에 대한 다중비교 : 집락수가 3 인 경우

비유사도의 종류	3	1	2	5	6	8	4	7	9
metric의 성질	O	O	O	O	X	X	X	O	X
통계량의 평균값	0.9271	0.9267	0.8968	0.6156	0.6115	0.6002	0.5967	0.5644	0.5371
Tukey의 group	A	A	A	B	B	(B,C)	(B,C)	(C,D)	D

표 13. 비유사도의 측정방법에 대한 다중비교 : 집락수가 4 인 경우

Distance의 종류	1	3	2	6	8	5	4	9	7
metric의 성질	O	O	O	O	X	O	X	X	X
통계량의 평균값	0.9113	0.9103	0.8987	0.6589	0.6492	0.6403	0.6236	0.5846	0.5726
Tukey의 group	A	A	A	B	B	B	(B,C)	(C,D)	D

표 14. 집락알고리즘의 종류에 대한 다중비교

3	집락방법의 종류	2	3	7	6	4	5	1
	공간의 수축 팽창	확장	보존	확장	확장	보존	보존	수축
	통계량의 평균	0.7677	0.7121	0.7106	0.7078	0.6493	0.6366	0.3632
	Tukey 의 그룹	A	B	B	B	C	C	D
4	집락방법의 종류	2	6	7	3	5	4	1
	공간의 수축 팽창	확장	확장	확장	보존	보존	보존	수축
	통계량의 평균	0.8137	0.7648	0.7429	0.6971	0.6442	0.6371	0.3154
	Tukey 의 그룹	A	A	B	C	D	D	F

위의 두 분석결과로 부터 양적자료에 있어서는 비유사도의 측정방법중 거리의 성질을 만족하는 측도들이 다른 측도들 보다 재생능력이 비교적 탁월한 사실을 알 수 있고, 또한 그중에서도 Minkowsky, Squard Euclidean, City-block 거리등이 특별히 좋은 성격을 갖는다.

집락 분석방법에서는 Comple Linkage , Ward, Average Linkage, Flexible 등의 공간확장 과 보존에 속하는 방법들이 좋은 성질을 소유함을 알 수 있다. 여기서 특히 집락들이 서로 교락되어 있는 경우에는 Ward 집락법이 비교적 재생능력이 탁월함을 알 수 있다.

마지막으로 본 실험의 분석결과에 의해서 교호작용이 유의하므로, 두 인자 A 와 B 의 수준조합의 통계표에 의해서 가장 높은 통계량의 평균은 A_2*B_7이고, 이것은 City-Block 거리에 대한 Ward 방법을 적용한 것이다.

7. 결 론

집락분석에 있어서 유사도 또는 비유사도의 측정방법과 계층적 집락알고리즘의 종류간의 관계를 고찰한 사실들을 요약하면 다음과 같다.

첫째, 질적자료에 있어서는 유사도의 측정방법중 거리의 성질을 만족하는 성격보다도, 두 개체간의 특성을 확실히 구별할 수 있는 정의방법, 즉 0 - 0 대응의 수가 포함되어 있는 유사도들이 집락재생능력이 탁월함을 알 수 있다.

둘째, 질적자료에서 각 집락간의 구별이 뚜렷한 경우에는 유사도의 정의방법이 집락알고리즘의 종류보다 많은 영향을 미치지만, 반대로 집락들이 서로 교락되어 있는 경우에는 계층적 집락알고리즘들이 많은 영향을 미친다. 이 중 공간팽창이나 보존등에 속하는 방법들이 특별히 좋은 성질을 갖는다.

셋째, 질적자료에서 특별히 발견되는 사실은 각 집락간의 특성이 잘 구별되는 경우에 0 - 0 대응의 수를 포함하는 비유사도의 정의방법에 대한 Single Linkage 집락분석법이 상당히 양호함을 알 수 있고, 이 사실은 지금까지 예견되지 않은 것이다.

네째 , 양적자료의 경우에는 유사도의 측정방법중 거리의 성질을 만족하는 측도들이 다른 측도들보다 재생능력이 탁월함을 알 수 있다.

다섯째 , 양적자료에서 각 집락간의 특성이 서로 구별되는 경우나 교락되어 있는 경우에 관계없이 공간팽창 및 보존에 속하는 집락방법들이 좋은 성격을 가짐을 알 수 있다.

끝으로 본 연구는 유사도의 종류와 계층적 집락분석방법에 관하여 연구되었지만, 다른 집락분석방법에 대한 연구도 진행중이다.

< 참고문헌 >

[1] Aldenderfer, M. S. and Blashfield, R. K. (1984), ***Cluster Analysis***, SAGE University Paper No44.

[2] Chae, S. S. (1988), A comparative study to predict the number of clusters in cluster analysis,Ph.D thesis,Oklahoma State University.

[3] DuBien, J. L. and Warde, W. D. (1987), "A comparison of agglomerative clustering methods with respect to noise", ***Communications in Statistics Theory and Methods***, vol 16,1433-1460

[4] Fowlkes, E. B. and Mallows, C. L. (1983), "A method for comparing two hierarchical clusterings ", *Journal of the American Statistical Association*, 78, 553-584.

[5] Gower, J. C. (1985). " Measures of similarity, dissimilarity and distance", *Encyclopedia of statistical sciences*, 5, 397-405.

[6] Gower, J. C. (1986), "Meric and Euclidean properties of dissimilarity coefficients", *Journal of Classification*, 3, 5-48.

[7] Lance, G. N. and Williams, W. T. (1967), "A general theory of sorting strategies 1.hierarchical systems," *Computer Journal* 9,373 - 380.

[8] Rand, W. M. (1971), "Objective criteria for the evaluation of clustering methods", *Journal of the American Statistical Association*, 66, 846-850.

[9] Spath, H.(1980). *Cluster Analysis Algorithms*, John Wiley and Sons

[10] 田中 豊 외 脇 本 和 昌(1984),ハソコソ 統計解析 ハソトフツワ 多變量 解析編, 共立出版株式會社.

A Study on the Relation between Dissimilarity and Hierarchical Agglomerative Methods in Clust analysis

Wan Hyun Cho[1)]

ASTRACT

In the this paper we consider the definition and mathematical properties of similarity or dissimilarity which have often used in clust analysis, and we apply a hierarchical agglomerative cluster algorithm to a dissimilarity metrx generated by these distance. Here we investigate the effect of relation between distance function and cluster algorithm on the retrieval ability of natural clusters. We present an empirical results for qualitative data as well as quantitative data.

1) Chonnam National University, Dept.of Statistics,300 YoungBong-Dong, Buk-gu, Kwangju, 500-757 KOREA.