응용통계연구 제 18권 2호, 2005년, pp. 329-341

군집수의 예측에 관한 방법의 제안 및 비교

채성산[1] 임남규[2]

요 약

군집방법의 비교시 사용되는 Rand(1971)의 $C_k, k = 2, 3, \ldots, N-1$ 통계량에 대한 점근 결과를 이용하여 자료에 존재하는 군집수를 예측하는 방법을 제안하였다. 제안된 방법과 C_k 통계량의 변화 형태에 따라 군집수를 예측하는 Chae와 Warde(1991)와 허명회와 이용구(2004)의 방법을 비교하기 위하여 모의실험을 하였다. 현실적인 문제를 고려하여 실제자료에 대해서는 계속적인 재표본의 형성을 위하여 붓스트랩방법을 사용하였다.

주요용어: 계통적 군집방법, Rand의 C_k 통계량

1. 서론

군집분석의 영역에서 대부분의 문제는 군집방법의 적용에 따른 결과를 비교하는 것이며, Rand(1971)의 C는 군집방법을 비교, 혹은 군집방법의 실행성 평가에 타당한 통계량으로 알려져 있다. 이러한 C 통계량은 $0.0 \leq C \leq 1.0$ 에 속하며, 동일한 자료에 두가지 다른 군집방법을 적용한 결과, C가 1.0이면 군집방법의 결과가 완전히 일치됨을 의미하고, 0.0이면 완전한 불일치를 나타낸다. 이를 이용한 군집수의 예측에 관한 타당성은 Chae와 Warde(1991), 이석훈 등(1995), 허명회와 이용구(2004)에 의하여 언급되었다. 최근, DuBien *et al.*(2004)는 Fowlkes와 Mallows(1983)에 의해 주어진 Rand C 통계량의 평균과 분산이 DuBien과 Warde(1981)에 의하여 주어진 Rand C 통계량의 평균과 분산의 특별한 경우임을 증명하였다.

본 연구에서는, C 통계량이 동일한 자료에 대한 두가지 다른 군집방법의 적용 결과로서 예측된 군집수 k에 따라 변하기 때문에 $C_k, k = 1, 2, .., K, ..., N$ 로 표현하였다. $C_k, k = 1, 2, .., K, ..., N$ 를 계산하기위하여 자료집합에 6개의 서로 다른 군집방법을 적용하였다. 만약, 주어진 자료에 타당한 구조가 존재하고 있다면 군집수의 예측 반복수는 크고, 타당한 구조가 존재하지 않는 다면 예측 반복수는 작을 것이다.

본 연구의 진행을 위하여, 우선적으로 6개의 계통적 군집방법을 선택하였고, DuBien과 Warde(1987), DuBien *et al.*(2004), Chae *et al.*(2004)에 의하여 연구된 비교통계량 Rand $C_k, k = 1, 2, .., K, ..., N$ 의 점근성을 이용하였다. 즉, N이 상당히 큰 경우, 비교통계량이 정규분포를 따른다고 가정하고, 이의 분산을 이용한 군집수의 예측이라는 새로운 방법을 제

1) (300-716) 대전시 동구 용운동 96-3, 대전대학교 정보통계학과, 부교수
E-mail: chae@dju.ac.kr
2) (300-716) 대전시 동구 용운동 96-3, 대전대학교 정보통계학과, 강의전담교수
E-mail: nklim@dju.ac.kr

안하였다. 모의실험을 실행하고, Chae와 Warde(1991)의 결과 및 허명회와 이용구(2004)의 엔트로피(entropy) 기준의 결과와 비교하여 본 연구에서 제안된 방법의 타당성을 살펴보았다.

2. 군집방법의 선택

주어진 자료에 군집방법을 적용하려면, 개체들 간의 거리를 측정하는 측정치를 정의하여야 한다. 본 연구의 목적상, 개체들 간의 거리측정에는 Euclidean 거리가 사용되었으며, 개체 혹은 군집 i 와 j가 결합되어 있고 새로운 개체 혹은 군집 k가 이들과 결합하였을 때의 거리, $d_{(ij)k}$, 는 Lance and Williams(1967)의 다음 공식을 이용하였다.

$$d_{(ij)k} = \alpha_i d_{ik} + \alpha_j d_{jk} + \beta d_{ij} + \pi |d_{ik} - d_{jk}| \tag{2.1}$$

위의 공식 (2.1)에서 d_{ij}는 n_i개의 개체를 포함하는 군집 Y_i 와 n_j개의 개체를 포함하는 군집 Y_j 간의 거리이며, α_i, α_j, β 와 π는 계통적군집방법을 정의하는 모수들이다.

DuBien(1976), DuBien과 Warde(1987)는 위의 공식 (2.1)에 타당한 제한조건을 설정하고 다음의 공식을 유도하였다.

$$d_{(ij)k} = \frac{1-\beta+2\pi}{2} d_{jk} + \frac{1-\beta-2\pi}{2} d_{ik} + \beta d_{ij} \tag{2.2}$$

위의 공식 (2.2) 에서 $d_{ij} < d_{ik} < d_{jk}$ 를 만족하며, 현재 사용하고 있는 대부분의 계통적 군집방법들이 위의 공식에 의하여 정의된다. 여기서, 공식 (2.2)를 만족하는 무수한 군집방법들을 (β, π)-군집방법족(family)이라고 부른다. 군집방법족에 대한 자세한 설명은 DuBien과 Warde(1987)을 참고하기로 하고, 본 연구에서는 (β, π)로 정의되는 방법중에서 잘 알려진, 단일연결법-(.0 ,-.5), 평균연결법-(.0,.0), 최장연결법-(.0,.5), 유동연결법-(-.25,.0)과 (-.5,.0), 그리고 새로운 방법-(-0.5, 0.25)을 사용하였다.

3. 비교통계량과 점근분포

본 연구의 목적은 주어진 자료에 대하여 붓스트랩기법을 사용하여 재표본을 추출하고, 추출된 재표본에 다른 성질을 갖는 다양한 군집방법을 적용하여 자료에 존재하는 군집수를 예측하는 것이다. 이를 위하여 군집화의 재현성(retrieval ability 혹은 reproducibility)과 동의성(agreement) 평가시 이용되는 Rand $C_k, k = 1, 2, ..., K, ..., N$ 을 사용하였고, 군집수 K를 예측하는 새로운 방법을 제시하였다. 먼저, 임의의 k에 대하여 Rand C_k를 살펴보면,

$$C_k = \frac{\binom{N}{2} - \frac{1}{2}(\sum_i^k n_{i\cdot}^2 + \sum_j^k n_{\cdot j}^2) + \sum_{i \neq j}^k n_{ij}^2}{\binom{N}{2}} \tag{3.1}$$

위에서, n_{ij}는 주어진 자료(X)에 대하여 하나의 군집방법의 적용에 의한 결과인 군집체(Y) 중에서 i 번째 군집에, 다른 군집방법의 적용에 의한 결과인 군집체(Y') 중에서 j 번째 군집에 속하는 개체의 수이다. 이러한 경우 $C_k(Y, Y')$는 k개 군집으로 이루어진 군집체의 동의

성을 나타낸다. 만약, Y가 군집수가 K인 자료의 군집체이고, Y'이 군집방법의 적용에 의해 $k = K$개 군집으로 형성된 군집체이면 $C_k(Y, Y')$는 재현성을 나타낸다. 이때, N개의 개체로 구성된 자료에 K개의 군집이 존재한다고 가정하면, 즉, $[N, K]$-모집단을 가정한다면, 공식 (3.1)은 이지변수에 의한 표현을 이용하여 다음과 같이 나타낼 수 있다.

$$C_k = \frac{\sum_{i=1}^{N-1}\sum_{i=2}^{N} M_{ij}}{\binom{N}{2}} \tag{3.2}$$

위에서 M_{ij}는 서로 다른 군집방법의 적용에 의하여 생성된 군집체에 대한 이지변수 표현을 이용하여 얻을 수 있다. 군집체 Y에 대한 이지변수 표현으로 나타낸 벡터를 U, 군집체 Y'에 대한 이지변수 표현으로 나타낸 벡터를 V라 했을 때, $u_{ij} = v_{ij}$이면 $M_{ij} = 1$ 이고, 그렇지 않으면 $M_{ij} = 0$ 이다. 이때, $[N, K]$-모집단에서 생성되는 군집체의 수는 *Stirling*의 2번째 종류의 수이며, 이는

$$S(N, K) = \frac{1}{K!}\sum_{j=0}^{K}\binom{K}{j}(-1)^j(K-j)^N,\ K = 1, 2, \ldots, N$$

이다. 물론, $[N]$-모집단에서 생성되는 군집체의 총수는

$$L_N = \sum_{K=1}^{N} S(N, K)$$

이며, Becker와 Riordan(1934)의 *Bell Number* 이다. K가 상당히 큰 경우의 *Stirling*의 2 번째 종류의 수에 관한 연구에 관심이 있다면 Lengyel(1984)을 참고하기 바랍니다.

본 연구에서는, $[N, K]$-모집단 만을 고려하려고 하기 때문에,

1. 군집체 Y 가 발생할 확률은, $P_r\{Y\} = \frac{1}{S(N,K)}$,
2. $[N, K]$-모집단에서, 두 군집체 Y, Y'는 임의복원추출되며,
3. $[N, K]$-모집단에서, 두 군집체 Y, Y'의 동시 발생 확률은, $P_r\{(Y, Y')\} = \frac{1}{S(N,K)^2}$.

을 이용하여 C_k 통계량의 확률분포를 살펴볼 수도 있다.

표3.1은 $N = 4$인 경우, 생성 가능한 군집체와 이지변수 표현이며, 이를 이용하여 $\sum_{i=1}^{N-1}\sum_{i=2}^{N} M_{ij}$를 계산한 것이 표3.2 이다. 표3.2에서 $[N]$-모집단에서 C 의 확률분포, $f(c; N, L_N)$와 $[N, K]$-모집단에서 $C_k, k = K = 2, 3$ 의 확률분포, $f(c_k; N, K, S(N, K))$ 를 구할 수 있다. 본 연구에서의 관심은 $[N, K]$-모집단에서 군집수 K에 대한 예측에 한정하였다. 따라서, 생성되는 모든 군집체에 대하여 위와 같은 과정을 통하여 $N \geq 5$인 경우에 대해서도 C_k 의 확률분포(채성산, 1997)를 구할 수 있다.

이론적으로 N개의 관측치에 대한 모든 계산은 가능하지만, C 혹은 C_k 통계량에 대하여 모든 경우를 살펴보려면 컴퓨터를 사용한다 하더라도 많은 시간이 소요되는 것이 사실이며, 군집체 Y 와 군집체 Y' 간의 $\sum_{i=1}^{N-1}\sum_{i=2}^{N} M_{ij}$를 구하는 것도 정말 힘든 작업이다.

표 3.1: $N = 4$ 인 경우, 가능한 모든 군집체와 이지변수표현

군집수	번호	가능군집	$U = (u_{12}, u_{13}, u_{14}, u_{23}, u_{24}, u_{34})$
K=1	1	(X_1, X_2, X_3, X_4)	(1 1 1 1 1 1)
K=2	2	$(X_1, X_2, X_3)(X_4)$	(1 1 0 1 0 0)
	3	$(X_1, X_2, X_4)(X_3)$	(1 0 1 0 1 0)
	4	$(X_1, X_3, X_4)(X_2)$	(0 1 1 0 0 1)
	5	$(X_2, X_3, X_4)(X_1)$	(0 0 0 1 1 1)
	6	$(X_1, X_2)(X_3, X_4)$	(1 0 0 0 0 1)
	7	$(X_1, X_3)(X_2, X_4)$	(0 1 0 0 1 0)
	8	$(X_1, X_4)(X_2, X_3)$	(0 0 1 1 0 0)
K=3	9	$(X_1, X_2)(X_3)(X_4)$	(1 0 0 0 0 0)
	10	$(X_1, X_3)(X_2)(X_4)$	(0 1 0 0 0 0)
	11	$(X_1, X_4)(X_2)(X_3)$	(0 0 1 0 0 0)
	12	$(X_2, X_3)(X_1)(X_4)$	(0 0 0 1 0 0)
	13	$(X_2, X_4)(X_1)(X_3)$	(0 0 0 0 1 0)
	14	$(X_3, X_4)(X_1)(X_2)$	(0 0 0 0 0 1)
K=4	15	$(X_1)(X_2)(X_3)(X_4)$	(0 0 0 0 0 0)

표 3.2: N=4 인 경우, 군집체 Y 와 군집체 Y' 의 $\sum_{i=1}^{N-1}\sum_{i=2}^{N} M_{ij}$

Y/Y'	K	1	2							3						4
K	번호	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1	6	3	3	3	3	2	2	2	1	1	1	1	1	1	0
2	2	3	6	2	2	2	3	3	3	4	4	2	4	2	2	3
	3	3	2	6	2	2	3	3	3	4	2	4	2	4	2	3
	4	3	2	2	6	2	3	3	3	2	4	4	2	2	4	3
	5	3	2	2	2	6	3	3	3	2	2	2	4	4	4	3
	6	2	3	3	3	3	6	2	2	5	3	3	3	3	5	4
	7	2	3	3	3	3	2	6	2	3	5	3	3	5	3	4
	8	2	3	3	3	3	2	2	6	3	3	5	5	3	3	4
3	9	1	4	4	2	2	5	3	3	6	4	4	4	4	4	5
	10	1	4	2	4	2	3	5	3	4	6	4	4	4	4	5
	11	1	2	4	4	2	3	3	5	4	4	6	4	4	4	5
	12	1	4	2	2	4	3	3	5	4	4	4	6	4	4	5
	13	1	2	4	2	4	3	5	3	4	4	4	4	6	4	5
	14	1	2	2	4	4	5	3	3	4	4	4	4	4	6	5
4	15	0	3	3	3	3	4	4	4	5	5	5	5	5	5	6

실질적으로, $[N,K]$-모집단에서도 생성되는 군집체의 수는 $S(N,K)$, 즉, *Stirling*의 2 번째 수로서 증가하기 때문에 그 실체를 파악하기는 거의 불가능하다고 할 수도 있다. 여기서 $S(N,K)$에 대한 몇 가지 사실을 살펴보면,

$$\begin{aligned} S(N,0) &= 0; \\ S(N,N+b) &= 0 \quad \text{if} \quad b>0; \\ S(N+1,K) &= K\,S(N,K)+S(N,K-1); \\ S(N,K) &\approx \frac{K^N}{K}=K^{N-1} \quad \text{if} \quad N\to\infty, \end{aligned}$$

이다.

위의 결과를 이용하여 N개의 개체에 K개의 군집이 존재하는 경우의 $[N,K]$-모집단, $K=2,3,...,N-1$ 에서, N이 상당히 큰 경우 C_k 통계량의 점근적 평균과 분산은 다음과 같다.

$$E_{Asy}(C_k) = \frac{1+(K-1)^2}{K^2}=p, \tag{3.3}$$

$$V_{Asy}(C_k) = \frac{1}{\binom{N}{2}}\left\{\frac{2(K-1)[1+(K-1)^2]}{K^4}\right\}=\frac{p(1-p)}{\binom{N}{2}}=\frac{p(1-p)}{n}. \tag{3.4}$$

참고적으로, Chae *et al.*(2004)에서

$$V(C_k)=\frac{p(1-p)}{n}+o(n^{-2}) \quad \text{as} \quad N\to\infty \quad \text{for} \quad K\geq 2, \quad N\geq 4$$

이며, N이 증가하면 두 군집체 Y와 Y'의 공분산이 0에 근접한다. 또한, Idrissi(2000)는 Kolmogorov의 적합도 검정을 이용하여 수정된 C_k 통계량이 점근적으로 정규분포를 따른다고 하였다. 즉,

$$\frac{C_k-E_{Asy}(C_k)}{\sqrt{V_{Asy}(C_k)}}\to N(0,1), \quad \text{if} \quad n\to\infty$$

을 가정하고,

$$C_k+z\sqrt{V_{Asy}(C_k)}\geq 1.0, \; k=2,3,\ldots,N-1 \tag{3.5}$$

을 만족하면 자료에 존재하는 군집수를 k라 예측한다.

표3.3과 표3.4는 $N=15,25,35$이고 $k=2,3,\ldots,10$인 경우의 $E(C_k)$, $E_{Asy}(C_k)$ 와 $\sqrt{V(C_k)}$, $\sqrt{V_{Asy}(C_k)}$ 의 값을 나타낸 것이다. 여기서 N이 증가하면 $E_{Asy}(C_k)$ 와 $\sqrt{V_{Asy}(C_k)}$는 $E(C_k)$ 와 $\sqrt{V(C_k)}$에 접근함을 알 수 있다.

Chae *et al.*(2004)에서 언급되었듯이 C_k가 0.0 과 1.0 사이의 어떤 값, 예를 들면 0.9, 혹은 0.8 이라는 것이 명백한 의미를 갖지는 않는다. 단지, 동일 자료에 대하여 두 개의 군집방법들을 적용하였을 때, 군집방법들 간의 실행에 대한 동의성을 측정할 따름이다. 즉, $C_k=1.0$ 이면 완전한 일치성, 혹은 $C_k=0.0$ 이면 완전한 불일치성을 나타낼 경우에 그 값이 갖는 의미가 명백하다. 한편으로, $K=k$일 경우 완전한 불일치성을 나타내는 경우는 발생하지 않기 때문에, 완전한 일치성을 의미하는 $C_k=1.0$ 인 경우, 즉, 공식 (3.5)를 만족하는 경우에 대하여 살펴보는 것이 타당하다.

표 3.3: $E(C_k)$, $E_{Asy}(C_k)$ for $N = 15, 25, 35$

	$E(C_k)$			$E_{Asy}(C_k)$
k/N	15	25	35	
2	.5000	.5000	.5000	.5000
3	.5563	.5556	.5556	.5556
4	.6297	.6263	.6250	.6250
5	.6920	.6812	.6801	.6800
6	.7439	.7251	.7227	.7222
7	.7878	.7606	.7562	.7551
8	.8256	.7899	.7833	.7813
9	.8587	.8148	.8058	.8025
10	.8882	.8362	.8249	.8200

표 3.4: $\sqrt{V(C_k)}$, $\sqrt{V_{Asy}(C_k)}$ for $N = 15, 25, 35$

N	15		25		35	
k	$\sqrt{V}(C_k)$	$\sqrt{V_{Asy}(C_k)}$	$\sqrt{V}(C_k)$	$\sqrt{V_{Asy}(C_k)}$	$\sqrt{V}(C_k)$	$\sqrt{V_{Asy}(C_k)}$
2	.0488	.0448	.0289	.0289	.0205	.0205
3	.0480	.0485	.0287	.0287	.0204	.0204
4	.0449	.0472	.0278	.0280	.0198	.0198
5	.0404	.0455	.0262	.0269	.0190	.0191
6	.0354	.0437	.0244	.0259	.0181	.0184
7	.0305	.0420	.0225	.0248	.0171	.0176
8	.0258	.0403	.0207	.0239	.0160	.0169
9	.0212	.0389	.0188	.0230	.0150	.0163
10	.0170	.0375	.0171	.0222	.0140	.0158

4. 모의실험의 계획 및 결과

본 연구의 목적상, 다음과 같은 구조의 자료를 생성하였다. 즉,

$$X_{ki} \sim MVN(\mu_k, \mathbf{\Sigma}_k)$$

를 따르고, $i = 1, 2, \ldots, 30$ 이며, 부모집단의 수는 $K = 3$이다. 3개의 부모집단에 속하는 자료의 수는 $(n_1; n_2; n_3) = (10; 10; 10)$ 이고, $k = 3$인 각 부모집단의 평균은 다음과 같다.

$$\mu_1 = \begin{pmatrix} 0.0 \\ 0.0 \end{pmatrix}, \quad \mu_2 = \begin{pmatrix} \delta \\ 0.0 \end{pmatrix}, \quad \mu_3 = \begin{pmatrix} \frac{\delta}{2} \\ \frac{\delta}{2}\sqrt{3.0} \end{pmatrix}$$

이때, 평균간의 거리는 $\delta = \delta_k = 5.0, 6.0$이며, 분산-공분산 행렬, $\mathbf{\Sigma}_k$, $k = 1, .., 3$ 는

$$\mathbf{\Sigma_k} = \mathbf{\Sigma} = \begin{pmatrix} 1.0 & \rho \\ \rho & 1.0 \end{pmatrix}, \quad \rho = 0.0, 0.4, 0.8.$$

이다.

표본의 반복 추출은 1000번 실행하였으며, 각 반복 추출된 표본에 대하여 위에서 선택된 6개의 군집방법을 동시적으로 적용하였으며, 군집방법들의 각 쌍들의 실행 결과 중, 모든 $K = k$에 대하여 동의성을 이용한 Chae와 Warde(1991)에 의한 다음의 조건,

$$C_{k-1} \leq C_k \quad 와 \quad C_k > C_{k+1}, k = 2, 3, \ldots, N-1 \tag{4.1}$$

허명회와 이용구(2004)에 의하여 제안된 엔트로피(entropy) 기준을 활용한 다음의 조건,

$$AvgEnt_{k-1} \geq AvgEnt_k \quad 와 \quad AvgEnt_k < AvgEnt_{k+1}, k = 2, 3, \ldots, N-1 \tag{4.2}$$

을 만족하는 반복수를 계산하였다. 이때,

$$AvgEnt = (RowEnt + ColEnt)/2$$

이며, $RowEnt$와 $ColEnt$는 각각 행 다항분포와 열 다항분포의 엔트로피를 결합한 것으로

$$\begin{aligned} RowEnt &= \sum_{i=1}^{K} \frac{n_{i.}}{N} \left[\sum_{j=1}^{K} -\frac{n_{ij}}{n_{i.}} log\left(\frac{n_{ij}}{n_{i.}}\right) \right], \\ ColEnt &= \sum_{j=1}^{K} \frac{n_{.j}}{N} \left[\sum_{i=1}^{K} -\frac{n_{ij}}{n_{.j}} log\left(\frac{n_{ij}}{n_{.j}}\right) \right] \end{aligned}$$

이다.

또한, 본 연구에서 제안된,

$$C_k + z_\alpha \sqrt{V_{Asy}(C_k)} \geq 1.0, \ k = 2, 3, \ldots, N-1$$

조건을 만족하는 반복수를 계산하였으며, 정규분포표에서의 값 z_α, $\alpha = 0.05$ 를 이용하였다. 각각의 경우에 대한 비교 검토를 위하여 $K = 3$으로 고정하였다.

지금까지의 실행이 생성된 자료에 대한 방법들 간의 비교라고 한다면, 실제자료에 대하여 그 결과를 비교 검토할 필요성이 있을 것이다. 본 연구에서 제안된 방법의 타당성을 비교 검토하기 위하여, 표4.1의 실제자료(SPSS, 1995)를 이용하였다. 자료의 반복적인 추출을 위하여 붓스트랩기법을 적용하였으며, $k = 2, 3, 4, 5$ 의 각 경우에 대하여 이를 반복시행 하였으나, $k = 2, 3, 4$ 에 대한 결과만을 제시하였다. 각 반복마다 C_k에 주어진 각 방법에 대한 조건을 만족하는 예측반복비율을 계산하였으며, 그 결과를 비교 검토하였다.

표 4.1: Consumer Report of Beer in U.S.A. in 1983(SPSS, 1986)

관측수	맥주회사	칼로리	소다	알콜	가격
1	budweiser	144	15	4.7	0.43
2	schlitz	151	19	4.9	0.43
3	lowenbrau	157	15	4.9	0.48
4	kronenbourg	170	7	5.2	0.73
5	heineken	152	11	5.0	0.77
6	old milwaukee	145	23	4.6	0.28
7	augsberger	175	24	5.5	0.40
8	strohs bohemtan style	149	23	4.7	0.42
9	miler lite	99	10	4.3	0.43
10	budweiser light	113	8	3.7	0.44
11	coors	140	18	4.6	0.44
12	coors light	102	15	4.1	0.46
13	michelob light	135	11	4.2	0.50
14	becks	150	19	4.7	0.76
15	kirin	149	6	5.0	0.79
16	pabst extra light	68	15	2.3	0.38
17	hamms	136	19	4.4	0.43
18	heilemans old style	144	24	4.9	0.43
19	olympia gold light	72	6	2.9	0.46
20	schlitz light	97	7	4.2	0.47

5. 분석결과 및 결론

본 연구에서는 모의실험을 통하여, Rand C_k 통계량을 이용하여 군집수를 예측하는 방법을 제안하고 이를 기존의 방법과 비교 하였다. 물론, 제안된 방법의 이론적인 측면은 C_k 통계량에 대한 점근성에 따라 그 모습을 달리 할 수도 있을 것이다.

그러나, Chae *et al.*(2004)의 연구에서 N 이 증가하면 C_k 통계량이 정규성을 갖는다는 것이 사실이라는 가정하에, 여러 가지 군집방법을 동시적용하였다. 설정된 모수하에서 자료의 생성을 위하여, 1000번의 반복추출이 실행되었으며, 6개의 군집방법을 동시적용하였다. 군집방법의 각 쌍에 대하여 조건을 만족하는 예측반복비율이 계산되었으며, 생성된 자료에 대한 세가지 방법, Chae와 Warde(1991), 허명회와 이용구(2004), 그리고 본 연구에서 제안된 방법에 대한 결과는 표5.1에 요약 정리되었다.

표5.1에 제시된 결과를 비교해보면, 생성된 자료에 대하여 두 개의 군집방법을 동시 적용하여 군집수를 예측한 경우, 본 연구에서 제안된 방법이 군집수의 예측에 더욱 효과적임을 알 수 있다. 특이한 점으로 (.0, -.5)-단일 연결법과 다른 군집방법을 적용하였을 경우, 허

명희와 이용구(2004)의 엔트로피 기준을 활용한 군집수의 예측도 좋아 보인다는 점이다.

표 5.1: $K = 3$일 때, 세가지 방법에 의한 예측반복비율

δ	ρ		0.0			0.4			0.8		
	(β, π)	(β, π)	CW	Ent	New	CW	Ent	New	CW	Ent	New
		(.0, .0)	.343	.332	.348	.357	.370	.422	.382	.443	.461
		(.0, .5)	.463	.480	.380	.479	.539	.396	.458	.638	.405
	(.0,-.5)	(-.25,.0)	.420	.330	.380	.444	.355	.430	.443	.459	.470
		(-.5,.0)	.446	.331	.365	.468	.351	.414	.440	.446	.449
		(-.5,.25)	.486	.385	.360	.497	.412	.401	.455	.515	.438
		(.0,.5)	.402	.396	.666	.389	.413	.680	.326	.392	.635
		(-.25,.0)	.437	.365	.706	.365	.328	.710	.368	.348	.688
5.0	(.0,.0)	(-.5,.0)	.553	.423	.648	.480	.412	.660	.475	.431	.636
		(-.5,.25)	.592	.485	.614	.529	.476	.625	.489	.483	.598
		(-.25,.0)	.407	.434	.714	.371	.466	.721	.314	.435	.688
	(.0,.5)	(-.5,.0)	.493	.465	.672	.457	.517	.664	.427	.492	.633
		(-.5,.25)	.522	.487	.641	.494	.519	.643	.437	.476	.608
	(-.25,.0)	(-.5,.0)	.308	.215	.811	.270	.246	.812	.264	.273	.813
		(-.5,.25)	.409	.388	.739	.376	.383	.730	.324	.418	.740
	(-.5,.0)	(-.5,.25)	.232	.313	.821	.219	.303	.816	.202	.326	.825
		(.0, .0)	.509	.398	.769	.499	.429	.737	.441	.460	.696
		(.0, .5)	.642	.571	.737	.642	.687	.716	.569	.783	.664
	(.0,-.5)	(-.25,.0)	.641	.373	.764	.613	.378	.741	.552	.493	.712
		(-.5,.0)	.712	.373	.746	.659	.372	.720	.575	.482	.701
		(-.5,.25)	.736	.473	.731	.684	.466	.700	.601	.551	.682
		(.0,.5)	.477	.462	.871	.424	.484	.844	.362	.505	.766
		(-.25,.0)	.500	.359	.880	.401	.351	.864	.368	.342	.815
6.0	(.0,.0)	(-.5,.0)	.667	.467	.848	.544	.412	.829	.513	.412	.781
		(-.5,.25)	.717	.522	.824	.600	.502	.808	.543	.508	.746
		(-.25,.0)	.460	.481	.879	.375	.534	.878	.348	.544	.821
	(.0,.5)	(-.5,.0)	.579	.518	.847	.499	.561	.834	.476	.586	.782
		(-.5,.25)	.599	.520	.839	.544	.538	.815	.513	.564	.770
	(-.25,.0)	(-.5,.0)	.361	.243	.911	.308	.232	.901	.287	.245	.884
		(-.5,.25)	.476	.451	.876	.397	.414	.861	.375	.467	.837
	(-.5,.0)	(-.5,.25)	.269	.355	.910	.221	.352	.904	.213	.372	.895

이러한 결과는 자료의 생성시 군집수를 3 개로 고정하였기 때문에 쉽게 파악될 수 있다. 그

러나, 현실적인 측면을 고려하였을 경우에 취득된 자료가 포함하고 있는 정보를 파악하기는 쉽지 않으며, 군집수에 대한 정보도 모르는 경우가 대부분이다. 따라서, 제안된 방법의 실제적인 응용 및 그 타당성을 살펴보기 위하여, 표4.1의 실제자료를 이용하였으며, 그 결과를 표5.2과 표5.3에 제시하였다. 표5.2은 원래의 자료를 이용하였고, 표5.3은 통계적 표준화를 실시한 후 군집수를 예측한 것이다.

표5.2에서 Chae와 Warde(1991)와 허명회와 이용구(2004)의 엔트로피 기준을 활용한 군집수의 예측 결과가 아주 유사하게 나타났다. 그러나 제안된 방법의 예측률이 $K = 2$ 에서 상당히 높게 나타났음도 살펴볼 수 있다. 이러한 결과는 일차적으로 주어진 자료의 구조에 영향을 받으며, 이차적으로 적용된 군집방법들의 성향에 따라 다르게 나타날 수 있다. 그러나, 본 연구에서 적용된 군집방법들은 군집분석의 영역에서 잘 알려져 있고, 그 이용 빈도도 상당히 높은 방법들이다. 이러한 관점에서, 현재 사용된 여러 가지 군집방법의 적용 결과가 비슷하다면, 그 결과들에 근거한 판단은 상당한 합리성을 지닌다고 할 수 있다.

표 5.2: Beer data에서 $k = 2, 3, 4$ 에 대한 예측반복비율

		$k=2$			$k=3$			$k=4$		
(β, π)	(β, π)	CW	Ent	New	CW	Ent	New	CW	Ent	New
	(.0 , .0)	.157	.157	.583	.302	.302	.801	.358	.354	.667
	(.0 , .5)	.332	.332	.558	.305	.311	.576	.517	.473	.509
(.0, -.5)	(-.25, .0)	.300	.300	.553	.285	.300	.512	.484	.439	.503
	(-.5 , .0)	.367	.379	.551	.276	.304	.359	.415	.385	.328
	(-.5 , .25)	.397	.367	.542	.245	.286	.288	.390	.348	.231
	(.0, .5)	.238	.238	.973	.188	.188	.752	.509	.508	.771
(.0 , .0)	(-.25, .0)	.348	.348	.968	.206	.206	.640	.481	.478	.730
	(-.5 , .0)	.538	.538	.960	.238	.239	.446	.406	.426	.476
	(-.5 , .25)	.603	.603	.943	.217	.218	.362	.370	.398	.353
	(-.25, .0)	.298	.298	.993	.184	.184	.698	.424	.427	.740
(.0 , .5)	(-.5 , .0)	.488	.488	.985	.229	.230	.506	.382	.403	.494
	(-.5 , .25)	.550	.550	.968	.225	.226	.429	.337	.375	.386
(-.25, .0)	(-.5 , .0)	.277	.277	.990	.216	.217	.721	.219	.227	.638
	(-.5 , .25)	.388	.388	.973	.253	.252	.593	.245	.266	.490
(-.5 , .0)	(-.5 , .25)	.191	.191	.976	.155	.155	.769	.138	.146	.743

표준화를 실시한 후 군집방법을 적용한 결과인 표5.3을 살펴보면, 실제자료에 존재하는 군집수를 예측하기는 쉽지 않다. 어떠한 군집 방법을 적용했는가에 따라 $K = 3$ 혹은 $K = 4$ 으로 결정될 것이다. 이러한 판단은 이석훈, 박래현과 김응환(1995)의 연구에서 제시된 $K = 4$ 와 다르며, 본 연구에서 사용된 군집분석방법과는 다른 분석방법이었다. 즉, 주관적으로 어떠한 군집분석방법을 선택하느냐에 따라서 그 결과도 서로 다를 수 있는 것이

다. 단지, 군집수의 예측 관점에서 (-.25, .0)과 (-.5,.0)-유동연결법과 (-.5, .25)-새로운 군집방법을 동시 적용하였다면, 주어진 자료에서의 군집수는 $K = 3$ 이라 판단되며, 자료에 타당한 군집의 형태를 결정하는 것은 3가지 군집방법의 적용 결과를 검토하는 연구자의 몫이다. 또한, 주어진 자료에 직접적으로 혹은 표준화를 실시한 후에 군집방법을 적용할 것인지에 대한 판단도 분석대상으로 주어지는 자료의 특성을 이해한 후에 결정하여야 한다. 참고적으로, Chae와 Warde(2005)에 의하면 Mahalanobis 표준화는 군집분석의 측면에서 가능하면 사용하지 않을 것을 권하고 있다.

표 5.3: Beer data에서 $k = 2, 3, 4$ 에 대한 예측반복비율: 표준화된 경우

		$k=2$			$k=3$			$k=4$		
(β, π)	(β, π)	CW	Ent	New	CW	Ent	New	CW	Ent	New
	(.0 , .0)	.158	.158	.347	.302	.320	.414	.168	.181	.307
	(.0 , .5)	.136	.136	.215	.191	.222	.223	.288	.319	.259
(.0, -.5)	(-.25, .0)	.140	.140	.216	.189	.235	.214	.290	.320	.229
	(-.5 , .0)	.118	.118	.154	.222	.255	.137	.332	.378	.154
	(-.5 , .25)	.110	.110	.131	.254	.263	.116	.329	.384	.126
	(.0, .5)	.107	.107	.660	.090	.101	.640	.322	.337	.792
(.0 , .0)	(-.25, .0)	.077	.077	.522	.174	.190	.615	.412	.417	.743
	(-.5 , .0)	.083	.084	.309	.271	.297	.490	.432	.436	.539
	(-.5 , .25)	.079	.080	.238	.292	.318	.439	.403	.407	.442
	(-.25, .0)	.070	.070	.518	.174	.182	.640	.381	.394	.696
(.0 , .5)	(-.5 , .0)	.077	.078	.348	.299	.311	.561	.393	.402	.513
	(-.5 , .25)	.088	.089	.278	.338	.345	.509	.346	.374	.419
(-.25, .0)	(-.5 , .0)	.087	.088	.630	.210	.211	.771	.247	.248	.698
	(-.5 , .25)	.103	.104	.490	.283	.280	.693	.271	.278	.567
(-.5 , .0)	(-.5 , .25)	.074	.074	.752	.148	.143	.829	.148	.151	.770

결론적으로, 본 연구에서 제안된 방법을 이용한 군집수의 예측은 자료가 지닌 특성 및 구조에 따라, 혹은 연구자가 주관적으로 선택한 군집방법에 따라 그 결과도 다르게 나타날 수도 있다. 이러한 문제점은 군집분석방법이 지닌 자체의 문제이며, 주어진 자료를 취급하는 연구자의 경험과 능력에 따라 극복되어야 할 것이다.

본 연구에서는 현실적으로 자주 이용되는 6개의 군집방법을 생성된 자료 및 실제자료에 동시적용하고, 각각의 군집방법의 적용에 의하여 생성되는 군집체들의 비교 결과로서 비교통계량의 값들을 계산하였다. 비교통계량이 점근적으로 정규분포를 따른다는 가정하에, 점근결과를 이용하여 자료에 존재하는 군집수의 예측방법을 제시하였고, 제안된 방법이 기존의 방법보다 자료에 존재하는 군집수의 예측에 더욱 효과적임을 알 수 있다.

6. 감사의 글

본 논문을 심사해주신 편집위원과 심사위원, 그리고 프랑스에서 학위논문을 보내 주신 Dr. Idrissi에게 깊은 감사를 드립니다.

참고문헌

이석훈, 박래현, 김응환 (1995). 쿨롱네트워크를 이용한 집락분석, <응용통계연구>, **8**, 39-50.

채성산 (1997). 재표본추출 및 검정을 통한 집락수의 예측, 대전대, <자연과학>, **8**, 73-88.

허명회, 이용구 (2004). K-평균 군집화의 재현성 평가 및 응용, <응용통계연구>, **17**, 135-144.

Becker, H. W., and Riodan, J. (1934). The arithmetic of bell and Stirling numbers, *American Journal of Mathematics*, **70**, 385-394.

Chae, S. S., DuBien, J. L. and Warde W. D. (2004). A method of predicting the number of clusters using asymptotic results on C_k, *Computational Statistics & Data Analysis*, (in revise)

Chae, S. S. and Warde W. D. (1991). A method to predict the number of clusters, *Journal of the Korean Statistical Society*, **20**, 162-176.

Chae, S. S. and Warde W. D. (2005). Effect of using principal coordinates and principal components on retrieval of clusters, *Computational Statistics & Data Analysis*, (in press)

DuBien J. L. and Warde, W. D. (1987). A comparison of agglomerative clustering method with respect to noise, *Communication in Statistics, Theory and Method*, **16**, 1433-1460.

DuBien, J. L., Warde, W. D. and Chae, S. S. (2004). Moments of Rand's C statistic in cluster analysis, *Statistics & Probability Letters*, **69**, 243-252.

Fowlkes, E. B. and Mallows, C. L. (1983). A method for comparing two hierarchical clusterings, *Journal of American Statistical Association*, **78**, 553-569.

Idrissi, A. (2000). *Contribution à l'Unification de Critères d'Association pour variables qualitatives*, Ph.D., Paris: Universitè Pierre et Marie Curie.

Lance, G. N. and Williams, W. T. (1967). A general theory of classificatory sorting strategies. 1. Hierarchical systems, *The Computer Journal*, **9**, 373-380.

Lengyel, T. (1984). On a recurrence involving Stirling numbers, *European Journal of Combinatorics*, **5**, 313-321.

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

[2004년 6월 접수, 2004년 12월 채택]

A Comparative Study of Determining the Number of Clusters with a Method Proposed

Seong San Chae[1] Nam Kyoo Lim[2]

ABSTRACT

A method of determining the number of clusters is proposed based on some asymptotic results on the Rand's(1971) C_k, $k = 2, 3, ..., N-1$, statistic. Simulation is conducted to compare the proposed method with Chae and Warde(1991), and Huh and Lee(2004).

Keywords: Agglomerative clustering algorithms, Rand's C_k statistics

1) Associate professor, Department of Information and Statistics, Daejeon University, 96-3, YoungUn-Dong, Dong-Gu, Daejeon, 300-716, Korea.
E-mail: chae@dju.ac.kr

2) Teaching professor, Department of Information and Statistics, Daejeon University, 96-3, YoungUn-Dong, Dong-Gu, Daejeon, 300-716, Korea.
E-mail: nklim@dju.ac.kr