군집수 결정 문제

How to determine the number of clusters

윤복식
홍익대학교 기초과학과
bsyoon@hongik.ac.kr

초 록

주어진 데이터를 일정한 기준에 따라 여러 개 군집으로 분할할 때 대부분 경우는 군집수에 대한 사전정보가 없이 군집화를 실시하게 된다. 적절한 군집수의 결정은 군집화 결과의 타당성에 전제가 되는 매우 중요한 문제이나 내재된 복잡성 때문에 실제 적용에 간편한 방법을 찾기 힘들고 더구나 다양한 형태의 데이터에 보편적으로 적합한 방법을 찾기는 더욱 어렵다. 본 연구에서는 기존에 제시된 군집수 결정방법 들의 아이디어 들을 소개하고 주어진 데이터의 종류에 관계없이 일반적으로 적용할 수 있는 새로운 군집수 결정기법을 제시한다. 대부분의 경우 군집수 결정은 군집화와 동시에 이루어지게 되므로 이것을 한꺼번에 처리하는 범용의 방법도 소개한다. 적용 예제들을 통한 타당성 검증도 이루어진다.

1. 서론

주어진 데이터를 여러 개 군집으로 분할할 때 대부분 경우는 군집수에 대한 사전정보가 없이 군집화를 실시하게 된다. 그러나 적절한 군집수의 결정은 군집화 결과의 타당성 여부에 매우 중요하고 군집화 기법 중에는 사전에 군집수를 정해 주도록 설계된 것들도 많다. 적절한 군집수의 결정에 대한 연구는 지금까지 매우 활발히 진행되어 왔고, 다양한 방법 및 판단기준들이 제시된 바 있으나 (Milligan & Cooper(1985), Milligan(1996), Peck(1989), Bock(1996), Hardy(1996) 등) 아직까지 해결이 안 된 난제로 남아 있다. 사실 군집수 결정 문제는 최적의 군집화와 군집화 결과의 타당성 문제와 연관되므로 독자적으로 접근하기 곤란한 문제이다.

2. 군집화 문제

각각 m개의 속성을 가진 n개의 개체의 데이터 $\boldsymbol{x}_i \in R^m, i=1,,,n$을 $S_1, S_2, \ldots, S_c$의 c개의 군집(cluster)으로 겹치지 않도록 나누되 가장 적합하게 분할(c-분할)하는 것이 군집화(clustering) 문제라고 할 수 있다. 그런데 어떻게 나누어야 가장 적절한 군집화인지를 판단하기가 쉬운 일은 아니다. 대략적으로 같은 군집에 속한 개체들은 동질성(similarity)과 서로 다른 군집에 있는 개체들 사이의 이질성(dissimilarity)을 극대화하면 최적의 군집화라고 볼 수 있을 것이다. 따라서 우선 이 개념을 잘 반영할 수 있는 기준함수를 설정해야 하는데

$$S=\{S_1, S_2, \ldots S_c\}, X=\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$$

로 표기하면 기준함수를 $f(c, S|X)$로 나타낼 수 있다. 이제 군집화 문제는 군집수와 군집의 원소들을 변동시키면서 기준함수 $f(c, S|X)$를 최대화 또는 최소화하는 조합최적화 문제로 파악할 수 있다. 여기서 사전에 군집수가 알려져 있지 않을 경우 에는 적어도 다음 2가지 접근이 가능할 것이다.

(1)한 가지는 군집수와 군집을 한꺼번에 고려하여 전역최적해(global optimum)을 구한다.

(2)군집수 c를 먼저 1,2,3..., $c_{\max}$ 등으로 고정시키고 최적의 군집화를 수행하여 이들의 기준함수 값을 비교하여 최적의 값을 주는 c를 구한다.

(2)는 계층적 주로 군집화에서 적용하는 접근법이고 (2)는 비교적 최근의 연구인 Nakamura (1998), Kothari(1999)와 본 논문에서 실험결과를 제시하게 되는 ASA 군집화 방법(윤복식(2004)) 등에 따르는 방법이다. (1)의 접근을 위해서는 기준함수를

$$f(c, S) = w^{(1)}(c) + w_c^{(2)}(S)$$

또는

$$f(c, S) = w^{(1)}(c) \cdot w_c^{(2)}(S)$$

형태로 설정할 필요가 있다. 여기서 $w^{(1)}(c)$는 c의 증가에 따른 비용이고, $w_c^{(2)}(S)$는 c개의 군집으로 군집화된 결과의 적합성 정도를 나타내는데 척도이다(Bock (1996), Milligan(1996), Geva(2000) 등 참조). .

기준함수의 $w_c^{(2)}(S)$의 전형적인 예로는

군집내부 편차

$$w_c(P) := \frac{1}{n} \sum_{i=1}^{c} \sum_{j \in S_i} \| x_j - \bar{x}_{S_i} \|^2$$

군집간 편차

$$b_c(P) = \frac{1}{n} \sum_{i=1}^{c} n_i \| \bar{x}_{S_i} - \bar{x} \|^2 \text{ , } (n_i = |S_i|)($$

비율기준

$$r_c(P) = \frac{b_c(P)}{w_c(P)}$$

등을 들 수 있다.

3. 방법의 실험

3.1 계층적 방법의 경우

3.1.1 기준함수를 이용한 그래프방법

군집화 기준에 의한 그래프방법은 결합형의 계층 군집화를 수행하며 기준함수의 값이 군집수 c에 따

라 변화는 과정을 그래프로 나타내고 변화폭이 급격히 줄어들기 시작하는 값을 적절히 선택하여 군집수를 정하는 선택하는 기법이다. 본 논문에서는

$$W_S^1 = \sum_{i=1}^{c} \frac{n(i)}{n} [\sum_{k=1}^{n(i)} \sum_{l>k}^{n(i)} \frac{2 \cdot \|x_k - x_l\|}{n(i) \cdot (n(i)-1)}]$$

를 기준함수로 설정하여 실험한다.

이 방법은 군집수를 1차원 그래프를 통해 결정하기 때문에 구조가 복잡한 데이터에 대한 분석에서 때로는 주관적인 판단에 의한 오류를 범할 수 있지만, 기준함수의 계산이 용이해서 큰 규모의 데이터에도 사용이 편리한 장점을 갖고 있다.

3.1.2 개선된 Mojena 방법

결합형의 계층적 군집화에서는 거리가 가장 짧은 두 군집을 제일 유사한 두 군집으로 간주하여 결합하게 된다. 각 결합 단계에서 결합된 두 군집간의 거리를 흔히 결합수준(fusion level)라고 부른다. 계층적 군집화에서, 적절한 군집수를 결정하는 방법은 결합(또는 분리)과정에 계층적 구조를 순차적으로 관찰하여 결합수준에 갑자기 큰 변화가 있을 때 결합과정을 종료하는 방법이다(Everitt(1991)참고). 이런 방법은 구조가 비교적 뚜렷한 데이터의 분석에서는 매우 간편하고 효과적일 수 있다

이것을 보다 개관화한 방법이 전체 결합기준치의 크기들을 통계적으로 비교하는 방법인데, 그 예로 Mojena(1977)의 방법을 들 수 있다. 이 방법은 $\alpha_0, \alpha_1, \ldots\ldots, \alpha_{n-1}$을 각 결합 단계에서 군집수 $n, n-1, \ldots, 1$에 대응되는 결합수준이라 하고, $\bar{\alpha}$와 s_α를 각기 결합기준치들의 평균 및 표준편차라고 할 때

$$\alpha_{j+1} > \bar{\alpha} + k\ s_\alpha$$

을 만족하는 최초의 j를 선택하여 $n-j$를 적절한 군집수로 제안하는 방법이다. 여기서 상수 k는 [2.75, 3.50]에서 취하도록 제안되었다.

본 논문에서는 결합기준치 대신에 결합기준치 증가폭의 상대적인 크기를 통계적으로 비교하여 적절한 군집수 판별기준을 얻는다. 이 방법은 Mojena의 방법처럼 전체 결합기준치를 일일이 비교할 필요가 없이 증가폭의 상대적인 크기만을 고려하기 때문에 군집수의 상한에서부터 시작하여 계산과정을 단순화시킬 수 있고 정확성에서도 더 우월할 것으로 기대된다. 이 방법에서는 증가폭과 평균 간의 편차를 고려할 때 증가폭의 스케일이 비교적 작은 점을 감안해서보통 사용하는 자승합의 평균 대신 절대값의 평균을 사용한다.

<결합기준치 증가폭에 의한 결정방법>

$c_{\max}$ 군집수의 상한이라고 할 때,

단계 1: 층적 결합방법으로 군집화를 실시하여 n개 군집(n개의 단일 개체군집)으로부터 1개의 군집(전체 개체를 한 군집에 묶는다)까지 순차적으로 결합해 가면서 결합기준치 l_c($c = 1, 2, \ldots . n-1$)을 계산한다.

단계2 : $c_{\max}$개 군집에서부터 시작하여 1개 군집으로 결합하는 단계까지의 결합기준치의 증가폭 α_j, $j = 1, \ldots, C_{\max} - 2$ 을 계산한다. 즉

$$\alpha_j = l_{n-C_{\max}-(j+1)} - l_{n-(C_{\max}-j)}$$

단계3 : 아래와 같은 통계량을 계산한다.

$$\bar{\alpha} = \frac{1}{c_{\max}-2} \sum_{j=1}^{c_{\max}-2} \alpha_j$$

$$s_\alpha = \frac{1}{c_{\max}-2} \sum_{j=1}^{c_{\max}-2} |\alpha_j - \bar{\alpha}|$$

그리고 다음과 같이 판별기준치 γ를 계산한다

$$\gamma = \bar{\alpha} + k \cdot s_\alpha$$

(본 연구에서는 실험을 통해 k을 1.5~3.5사이에서 취했을 때 상대적으로 좋은 결과를 얻었다. 본 논문 중 실제 분석에서는 k값을 2.5로 취하였다.)

단계4 (군집수 결정): $j = 1, 2, \ldots\ldots, c_{\max} - 2$ 에 대해 최초로

$$\alpha_j > \gamma$$

을 만족하는 j를 선택하여 $n-j$를 적절한 군집수로 설정한다(본 방법으로는 최대한 $n-1$개 군집으로 분리할 수 있다). 만일 앞에 식을 만족하는 j가 존재하지 않으면 개체 전체를 하나의 군집으로 간주한다.

3.2 ASA 군집화에 의한 결정

ASA 군집화 방법은 고정된 군집수의 분할에 의해 얻어진 결과를 초기해로 입력하여 모의어닐링 기법에 기반한 알고리즘으로 반복적으로 개선하여 최적에 가까운 군집화를 얻을 수 있게 한다. 뿐만 아니라 적절히 설계된 해의 변동과정과 변형된 군집화 기준의 설정을 통해 초기분할에서 주어진 군집수를 자동적으로 변동시키면서 최종적으로 적절한 군집수와 동시에 군집화 결과를 얻을 수 있게 해준다.

일반적으로 2절에서 예시한 표준적인 기준함수들은 최적화 과정에서 상대적으로 군집수를 크게 해주는 경향이 있어서 군집수의 증가를 효과적으로 제어할 수 없게 된다. 본 연구에서는 이런 문제점을 고려하여 다음과 같이 3.1.1과 유사한 변형된 군집화 기준을 정의하여 사용한다.

$$f(c, S) = \beta\, c \sum_{i=1}^{c} \sum_{k=1}^{n(i)} \sum_{l>k}^{n(i)} \frac{\|x_k - x_l\|}{n(i)-1}$$

여기서 β는 기준함수 값의 스케일을 조절하는 적절한 상수값 이다.

ASA 방법은 이 기준함수를 최소화하는 c-분할을 찾아가게 되는데 이 기준을 이용하면 무제한적으로 늘어날 수 있는 군집수와 불합리한 단일 개체의 군집의 증가를 방지할 수 있게 된다.

3.3 실험을 통한 비교

3.3.1 데이터 설명

여기서는 3.1, 3.2절에서 언급한 군집수 결정방법들을 비교하기 위해 [표 3.1]과 같이 의도된 군집구조의 모의데이터에 대한 분석을 실시한다. 여기서 데이터 4와 5는 군집 간에 다소 겹치는 부분이 있는

3차원 구형군집 데이터와 각 군집에 포함된 개체수가 서로 다른 4차원 구형군집 데이터이다.

일반적으로 최종결과에서 얻어진 군집수는 분석과정에 사용된 군집화 방법과 밀접한 관계가 있다. 이를테면 똑같은 군집수 결정방법이라도 어떤 군집화 방법에서 적용되었는가에 따라 결과에서 큰 차이를 보일 수 있어서 그 선택이 매우 중요하다.

하지만 본 연구에서 언급하는 ASA 군집화 방법은 1단계에서 계층적 군집화 방법인 최단거리와 최장거리방법을 각각 적용하여 초기해를 결정하기 때문에 본 실험에서도 이 두 가지 계층적 군집화 방법을 주로 적용하여 군집수를 결정하게 된다.

3.3.2. 실험 결과

우선 그래프에 의한 기준함수의 변화를 [그림 1] - [그림 5]에서 볼 수 있다. 또한 [표 3.2]에서는 모의데이터 1부터 5에 대한 분석결과를 제시하였다.

우선 ASA 군집화 방법이 과연 초기해의 군집수를 변동해 가면서 적절한 군집수(즉, c를 변동해 가면서)를 선택할 수 있는지에 대한 검정을 실시하기 위해 다양한 초기 군집수를 설정하여 이에 대응되는 여러 가지 초기분할로부터 시작하여 군집화를 실시하였다. 그 결과 수렴시간은 초기해에 따라 다소 차이가 났지만 50번 실험에서 대부분이 [표 3.2]의 첫 번째 행의 결과를 제시하였다. 실험으로부터 볼 때 ASA 군집화를 통해 얻은 결과는 내부루프의 반복횟수를 100으로 설정했을 때 평균적으로 93%정도가 데이터의 원래 구조를 찾아갔고, 나머지 실험결과에서도 거의 이런 결과에 접근하는 것을 관찰할 수 있었다. 따라서 ASA 군집화 방법은 적어도 다른 3 가지 방법과 대등하거나 또는 더 적절한 군집수를 자동적으로 선택해 주는 것으로 파악된다.

그리고 나머지 방법들을 살펴보면 군집화 기준함수에 기반을 둔 그래프방법이 비교적 효과적인 것으로 나타났다([그림 3.1] ~ [그림 3.5] 참조). 이는 "좋은" 구조의 데이터에 대해 이 방법은 아주 효과적임을 보여주고, 또 한편으로는 본 연구에서 설정한 군집화 기준은 군집구조를 반영하는데 비교적 효과적임을 설명해 주기도 한다.

그러나 결합기준치를 이용하는 표준 Mojena방법은 k값을 2.5부터 3사이에서 취해 가면서 적용했는데도 불구하고 두 가지 계층적 군집화 방법에서는 데이터의 원래 군집수를 정확히 결정하지 못한 것으로 나타났다.

반면 본 연구에서 제시한 결합기준치 증가폭에 의한 개선된 Mojena방법은 최장거리 방법에서 모두 군집수를 정확히 결정하였다. 최단거리 방법에서도 두 균일분포를 따르는 데이터에서는 적절한 군집수를 선택해 주었다. 비록 나머지 데이터에서 최단거리 방법에 의한 결과가 좋지는 않았지만 (최단거리 방법 자체가 이 데이터에 대한 적합한 분할방법이 아닌 것으로 보여진다), 전체적으로 이 방법이 표준 Mojena방법에 비해 보다 효과적이고 계산과정도 보다 간편하였다. 일반적으로 상대적으로 뚜렷이 구별되는 데이터에 대해 불안정한 방법은 좀더 복잡한 구조를 지닌 데이터에의 적용이 문제가 될 수 있기 때문에 이런 결과들은 의미가 있는 비교라고 보여진다.

참고문헌

[1] Bock, H. H.(1996), Probability models and hypothesis testing in partitioning cluster analysis, *clustering and classification,* world science Publ. 377-453.

[2] Everitt, B. S.(1980), *Cluster analysis,* Halsted Press, London.

[3] Everitt, B. S.(1991), *Applied Multivariate Data Analysis,* Wiley, New York.

[4] Fukuyama, Y. & Sugeno, M.(1989), A new method of choosing the number of clusters for the fuzzy c-means method, In: Proc. Fifth Fuzzy Syst. Symp.(in Japanese), 247-250.

[5] Geva, A. B. & Steinberg,Y., ...(2000), A comparison of cluster validaty criteria for a mixture of normal distributed data, *Pattern Recognition Letters,* **21**, 511-529.

[6] Gordon, A. D.(1996), Hierarchical Classification, *clustering and classification,* world science Publ. 65-121.

[7] Hand, D. J.(1981), *Discrimination and Classification,* John Wiley & Sons, New York.

[8] Hardy, A.(1996), On the number of clusters, *computational statistics & Data Analysis* ,**23**, 83-96.

[9] Jain, A. K. & Moreau, J. N.(1987), Bootstrap techniques in cluster analysis. *Pattern Recognition,* **20**, 547-568.

[10] Kothari, R. & Pitts, D.(1999), On finding the number of clusters, *Pattern Recognition Letters* , **20**, 405-416.

[11] Manly, B. J.(1994), *Multivariate Statistical Methods* (Second Ed.), Chapman & Hall, London.

[12] Milligan. G. W. (1980), Application and Implementation: A note on procedures for testing the quality of a clustering of a set of objects, *Decision Sciences,* **11**, 669-677.

[13] Milligan. G. W. (1981), A Monte Carlo Study of Thirty Internal Criterion Measures For Cluster Analysis, *Psychometrika,* **46** (2), 187-199.

[14] Milligan, G. W. & Cooper, M. C.(1985), An examination of procedures for determining the number of clusters in a data set, *Psychometrika,* **50** (2), 159-179.

[15] Milligan, G. W.(1996), Clustering Validation: Results and Implications for Applied Analysis, *clustering and classification,* world science Publ. 341-375.

[16] Mirkin, B.(1996), *Mathematical Classification and Clustering,* Kluwer Academic Publishers.

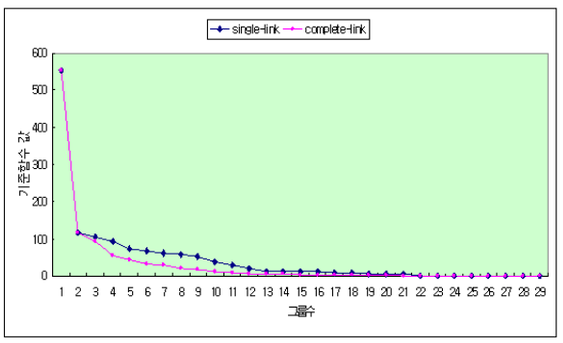
[17] Mojena, R.(1977), Hierarchical grouping methods and stopping rules: an evaluation, *Computer Journal,* **20**, 359-363.

[18] Nakamura, N. & Kehtarnavaz, N.(1998), Determining number of clusters and prototype locations via multi-scale clustering, *Pattern Recognition Letters,* **19**, 1265-1283.

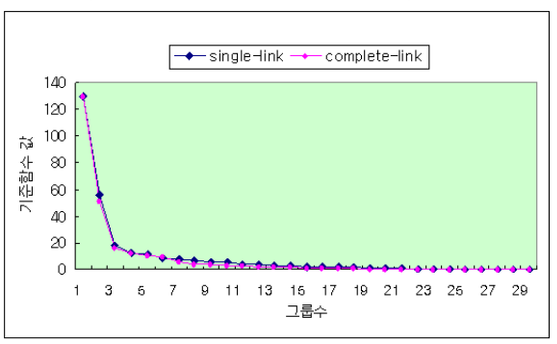
[19] Peck, R., Fisher, L. and Ness, J. V.(1989), Approximate confidence intervals for the number of clusters, *Journal of the American statistical Association,* **84** (405), 184-191,

[20] Sun, Y., Zhu, Q. & Chen, Z. X.(2001), An

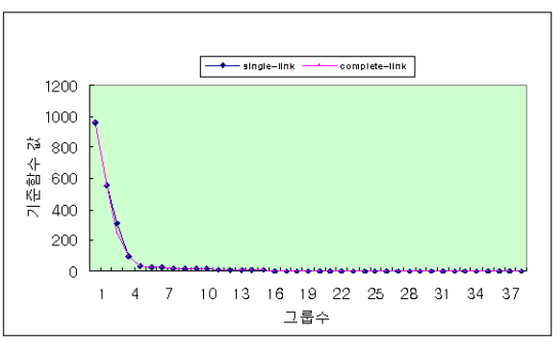
iterative initial-points refinement algorithm for categorical data clustering, *Pattern Recognitions Letters* (article in press).
[21] Xie, X. L. & Beni, G. A.(1991), Validity measure for fuzzy clustering., *IEEE Trans. Pattern Anal. Machine Intell.* **3** (8), 841-846.
[22] 윤복식(2004), 최적에 가까운 군집화를 위한 이단계 방법, 한국경영과학회지.



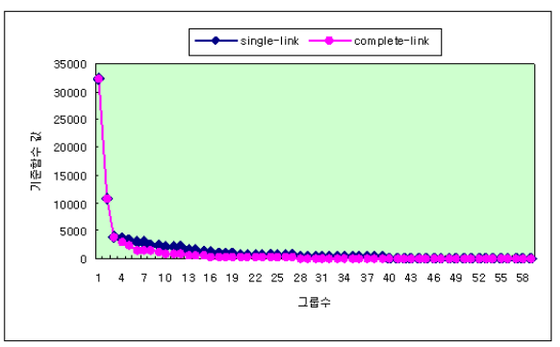
[그림 3.1] 데이터 1에 대한 기준값의 비교



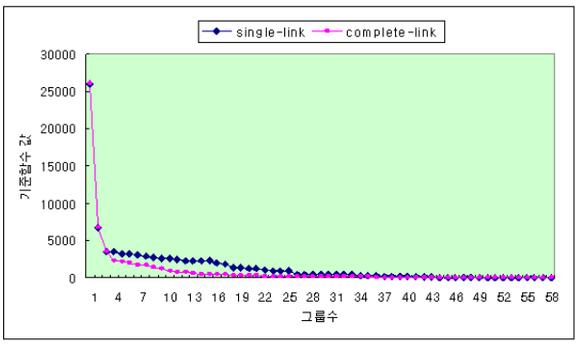
[그림 3.2] 데이터 2에 대한 기준함수 값 비교



[그림 3.3] 데이터 3에 대한 기준함수 값 비교



[그림 3.4] 데이터 4에 대한 기준함수 값 비교



[그림 3.5] 데이터 5에 대한 기준함수 값 비교

[표 3.1] 5 가지 모의데이터

◈ 모든 변수들은 서로 독립이라고 가정

2-군집 정규분포 데이터	모의데이터 1 (개체수 30)
두 분포 평균사이의 거리 2 분산을 일치 함	군집 1 : 평균벡터(1, 1), 분산 각기 0.3 군집 2 : 평균벡터(3, 1), 분산 각기 0.3
3-군집 데이터	모의데이터 2 (개체수 30)
3개 구역에서 각기 10개 개체를 무작위로 추출	$R_1=\{(x,y)\|0.2\le x\le 0.5,\ 0.1\le y\le 0.3\}$; $R_2=\{(x,y)\|0.6\le x\le 0.9,\ 0.1\le y\le 0.3\}$ $R_3=\{(x,y)\|0.5\le x\le 0.6,\ 0.4\le y\le 0.7\}$
5-군집 데이터	모의데이터 3 (개체수 40)
5개 구역에서 각기 8개 개체를 무작위로 추출	$R_1=\{(x,y)\|0\le x\le 0.5,\ 0\le y\le 0.5\}$; $R_2=\{(x,y)\|1.5\le x\le 2,\ 0\le y\le 0.5\}$ $R_3=\{(x,y)\|1.5\le x\le 2,\ 1.5\le y\le 2\}$; $R_4=\{(x,y)\|0\le x\le 0.5,\ 1.5\le y\le 2\}$ $R_5=\{(x,y)\|0.75\le x\le 1.25,\ 0.75\le y\le 1.25\}$

3-차원 구형-군집 데이터	모의데이터 4 (개체수 60)		
	군집 1	군집 2	군집 3
군집별 개체수	20	20	20
평균	(10, 10, 10)	(20, 20, 20)	(30, 30, 30)
표준 편차(각 변수 일치)	3	3	3

4-차원 구형-군집 데이터	모의데이터 5 (개체수 60)		
군집별 개체수	30	20	10
평균	(5, 5, 5, 5)	(13, 13, 13, 13)	(20, 20, 20, 20)
표준 편차(각 변수 일치)	2	2	2

[표 3.2] 모의데이터 1-5 에 대한 실험결과

사용된 방법 ↓ 데이터 종류 별 번호→	1	2	3	4	5
ASA 군집화 방법	2개(95%)	3개(95%)	5개(95%)	3개(90%)	3개(90%)
군집화 기준에 의한 그래프방법					
최단거리방법	2개	3개	5개	3개	3개
최장거리방법	2개	3개	5개	3개	3개
Mojena기준에 의한 방법 (k=2.5, 3)					
최단거리방법	6개	12개	5개	1개	1개
최장거리방법	2개	5개	5개	1개	1개
결합기준치 증가폭에 의한 방법 (k=2.5 , 3)					
최단거리방법	6개	4개	5개	5개	4개
최장거리방법	2개	3개	5개	3개	3개