韓國經營科學會誌
第31권 第2號
2006年 6月

ASA 군집화를 이용한 군집수 결정 및 다양한 실험*

윤 복 식**

Finding the Number of Clusters and Various Experiments Based on ASA Clustering Method*

Bok Sik Yoon**

■ Abstract ■

In many cases of cluster analysis we are forced to perform clustering without any prior knowledge on the number of clusters. But in some clustering methods such as k-means algorithm it is required to provide the number of clusters beforehand. In this study, we focus on the problem to determine the number of clusters in the given data. We follow the 2 stage approach of ASA clustering algorithm and mainly try to improve the performance of the first stage of the algorithm. We verify the usefulness of the method by applying it for various kinds of simulated data. Also, we apply the method for clustering two kinds of real life qualitative data.

Keyword : Clustering, Hierarchical Clustering, Number of Clusters, Simulated Annealing, ASA Clustering Algorithm

1. 서 론

주어진 데이터를 여러 개 군집으로 분할할 때 대부분 경우 군집수에 대한 사전정보가 없이 군집화를 실시하게 된다. 그러나 적절한 군집수의 결정은 군집화 결과의 타당성 여부에 매우 중요하고 군집화 기법 중에는 k-means 방법 등과 같이 사전에 군집수를 정해 주도록 설계된 것들도 많이 있다. 적절한 군집수의 결정에 대한 연구는 지금까지 많이 진행되어 왔고, 다양한 방법 및 판단기준들이 제시된 바 있으나[2, 6, 10, 11, 15] 아직까지 신뢰할 만한 접근방법이 정립되지 못한 채 난제로 남아 있다. 사

논문접수일 : 2006년 2월 7일 논문게재확정일 : 2006년 4월 28일
* 이 논문은 2004년도 홍익대학교 학술연구조성비에 의하여 연구되었음.
** 홍익대학교 기초과학과 응용수학전공

실 군집수 결정 문제는 구체적으로 군집화를 수행하는 방법 자체에 크게 의존하므로 군집화 결과의 타당성 문제와 별개로 독자적으로 접근하기 곤란한 문제이다. 예를 들면 2장에서 설명될 그래프를 이용하는 방법이나 Mojena 방법[13]은 계층군집화를 전제로 하는 방법들이고, 최적화를 기반으로 하는 방법들에서는 군집화와 군집수 결정이 동시에 이루어진다. 따라서 우선 군집화 방법자체가 신뢰할 만해야 하는데, 윤복식[1]에서 데이터의 구조에 무관하게 최적에 근접한 군집화를 수행해주는 범용의 군집화 방법으로 ASA 군집화 방법이 소개된 바 있다. 본 연구의 주된 목적은 이미 선행연구를 통해 성능이 확인된 ASA 군집화 방법을 기반으로 접근하여 효과적인 군집수 결정방법을 마련하고 그 유효성을 다양한 실험을 검증하는데 있다. 보다 구체적으로 ASA 군집화는 2단계 접근법으로 1단계에서 되도록 우수한 초기해를 마련한 후 2단계에서 모의어닐링을 통해 해를 반복적으로 개선하는 방식으로 구성되어 있는데, 본 연구에서는 주로 1단계 초기해를 마련하는 단계의 개선을 시도한다. 또한 선행논문에서 미비 되었던 정성적인 데이터에서의 ASA 군집화의 군집수 결정기능에 초점을 맞추어 실제 데이터에 대한 적용실험을 시도한다.

본 논문의 구성은 우선 2장에서 군집수 결정에 대한 기존의 접근방법을 간략히 정리하고 3장에서 기존의 방법의 개선 방안을 논한다. 여기에서는 ASA 군집화방법의 2단계 접근법에 맞추어 우선 계층군집화의 Mojena 방법을 개선하여 군집수를 효과적으로 결정하는 방법을 제시하고 그것을 초기해로 삼아 ASA 군집화 알고리즘을 통해 최적의 군집수로 근접하게 하는 방법을 논한다. 4장에서는 제시된 방법의 성능을 검토하기 위해 5개의 샘플 데이터를 만들어 초기에 군집수가 정해지는 과정과 ASA 군집화를 통해 어떻게 개선되는가 하는 것을 실험을 통해 보인다. 5장에서는 군집화하기가 상대적으로 어려운 정성적(qualitative) 데이터를 분석하여 ASA를 통해 군집수를 결정하고 군집화를 수행하는 사례를 제시한다. 정성적 데이터에서는 특히 두 개의 개체사이의 거리를 설정하는 방법이 이슈가 되는데 5.1절에서는 웹 문서들을 키워드의 출현빈도를 이용하여 내용에 따라 분류하는 문제를 다루었고 5.2절에서는 강사들을 소속학회 및 관심분야에 따라 분류하는 문제를 시도하였다.

2. 군집수 결정의 접근 방법

2.1 군집화 문제

각각 m개의 속성을 가진 n개의 개체의 데이터 $\boldsymbol{x_i} \in R^m$, $i=1,\cdots,n$을 $S_1, S_2,\cdots, S_c$의 c개의 군집(cluster)으로 겹치지 않도록 나누되 가장 적합하게 분할(c-분할)하는 것이 군집화(clustering) 문제라고 할 수 있다[12]. 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

$$S=\{S_1, S_2,\cdots, S_c\}, \quad X=\{\boldsymbol{x_1}, \boldsymbol{x_2},\cdots, \boldsymbol{x_n}\}$$

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

(i) 군집수와 군집을 한꺼번에 고려하여 전역최적해(global optimum)을 구한다.

(ii) 군집수 c를 먼저 $1, 2, 3,\cdots, C$ 등으로 고정시키고 최적의 군집화를 수행하여 이들의 기준함수값을 비교하여 최적의 값을 주는 c를 구한다.

(ii)는 주로 계층적 군집화(hierarchical clustering)에서 적용하는 접근법이고 (i)는 정수계획법과

같은 전통적인 조합최적화의 해법으로 접근하든가 아니면 본 논문에서 실험결과를 제시하게 되는 ASA 군집화 방법(윤복식[1])과 같은 근사적 최적화 기법을 따르는 방법이다. (i)의 접근을 위해서는 기준함수를

$$f(c,S) = W^{(1)}(c) + W_c^{(2)}(S) \tag{1}$$

또는

$$f(c,S) = W^{(1)}(c) \cdot W_c^{(2)}(S) \tag{2}$$

과 같이 군집수 c가 명시적으로 포함되는 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([2, 4, 11] 등 참조).

기준함수의 $W_c^{(2)}(S)$의 전형적인 예로는

군집내부 편차

$$W_1(S) = \frac{1}{n}\sum_{i=1}^{c}\sum_{j\in S_i}\|\boldsymbol{x_j} - \overline{\boldsymbol{x_{S_i}}}\|^2 \tag{3}$$

군집간 편차

$$B(S) = \frac{1}{n}\sum_{i=1}^{c} n_i \,\|\overline{\boldsymbol{x_{S_i}}} - \overline{\boldsymbol{x}}\|^2 \tag{4}$$

비율기준

$$R(S) = \frac{B(S)}{W_1(S)} \tag{5}$$

등을 들 수 있다. 여기와 논문의 다른 곳에서 $n_i = |S_i|$ (즉, 원소의 개수)이고 $\|\|$는 거리, $\overline{\boldsymbol{x}}$,$\overline{\boldsymbol{x_{S_i}}}$ 등은 평균을 의미한다.

2.2 군집수 결정을 위한 기존의 기법

본 절에서는 주로 결합형(agglomerative) 계층군집화에 기반하여 군집수를 정하는 기존의 접근방법을 검토한다. 결합형의 계층적 군집화에서는 매 단계에서 군집간의 거리가 가장 짧은 두 군집을 결합하게 된다. 각 결합 단계에서 결합된 두 군집간의 거리를 흔히 결합수준(fusion level)라고 부른다. 군집화가 진행됨에 따라 대개의 경우 결합수준이 점점 증가하게 되는데, 매 결합과정에서 결합수준을 관찰하여 결합수준이 갑자기 커질 때 결합과정을 종료하고 군집수를 결정하는 것이 합리적일 것이다. 이 방식을 통계적으로 객관화한 방법이 Mojena [13]의 방법이다. 즉, Mojena 방법에서는 $\alpha_0, \alpha_1, \cdots\cdots, \alpha_{n-1}$을 군집수 $n, n-1, \cdots, 1$에 대응되는 결합수준이라 하고, $\bar{\alpha}$와 s_α를 각기 결합수준 값들의 평균 및 표준편차라고 할 때

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

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

결합수준 대신에 식 (3)과 같은 군집기준함수를 설정하고 기준값의 변화를 군집수 c에 따른 그래프로 표현하여 변화폭이 급격히 줄어들기 시작하는 시점을 관찰하여 군집수를 정하는 방법도 실용적으로 많이 사용된다.

군집수의 신뢰구간을 통계적으로 설정하여 군집수의 범위를 제시할 수도 있는데 대개의 경우 데이터들의 분포를 알 수 없으므로 어려움이 따른다. Jain and Moreau[7], Peck et al.[15] 등 우선 군집화 기준을 식 (1)의 형태로 설정하고 Bootstrap 기법을 적용하여 군집수의 신뢰구간을 정하는 방법을 제안한 바 있다.

또한 반복적인 가설검정을 통해 군집수를 설정하는 방법도 생각할 수 있다. 즉 적당히 c의 상한 C를 설정한 후, $l = C, C-1, \cdots, 1$ 대해

[1] $H_0 : c = l$ 대 $H_1 : c = l-1$ (즉, 군집수를 하나 줄일 수 있는가?)

또는

[2] $H_0 : c = 1$ 대 $H_1 : c = l$ (즉, 전체를 동질군집($c=1$)으로 보는 것보다 l개의 군집으로 보는 것이 타당한가?)

의 검정을 반복적으로 시행하여 귀무가설을 더 이상 기각할 수 없는 시점에서 군집수를 결정하는 방

식이다. 이런 검정은 검정 통계량의 설정이 무엇보다 중요하지만 보통 검정 통계량에 대한 정보가 부족하여 정확한 분석이 어려워지게 되고(일부 통계량에 대한 예로 Bock[2]참고), bootstrap과 같은 시뮬레이션 기법을 사용하게 되어 계산량이 방대해지는 단점이 있다.

이외에도 군집화 알고리듬에 적절한 군집수의 결정과정을 통합시킨 방법들도 있다. Nakamura and Kehtarnavaz[14], Kothari and Pitts[8]에서는 scale-space이론을 이용하여 군집화 과정에 적절한 군집수도 함께 결정할 수 있는 알고리듬을 소개했다. 이런 방법들은 특별히 설계된 알고리듬을 통해 군집화 과정에 군집수의 결정과정을 통합시키므로 군집수에 대한 확실한 사전정보가 필요 없이 실제 문제에 적용될 수 있지만 적용범위의 일반성과 적용과정의 간편성이 문제가 된다. 본 논문에서 제시하게 되는 ASA 군집화 방법에 의한 군집수 결정은 이런 통합적 접근방법을 따르면서 일반적인 군집화 문제에 매우 간편하게 적용될 수 있다는 장점이 있다.

3. ASA 군집화 방법의 개선

ASA 군집화 방법(윤복식[1])은 1단계에서 계층군집화를 수행하여 잠정적인 군집을 얻은 후에 이를 초기해로 취하여 가속화된 모의어닐링에 기반한 ASA 군집화 알고리듬을 통해 최적에 가까운 군집화로 접근시키는 방법이다. 모의어닐링 과정에서 최적해에 보다 빠르게 접근하기 위해서는 초기해의 품질이 더 좋으면 유리하므로 1단계 과정에서 되도록 좋은 군집화 결과를 제공해줄 필요가 있다.

3.1 초기해의 개선

3.1.1 군집간 거리 설정

모의 어닐링에서 일반적으로 초기해가 우수하면 최적해로의 접근시간이 단축되는 것을 실험적으로 확인할 수 있으므로 되도록 우수한 초기해를 만들어낼 필요가 있다. 본 연구에서는 적용이 비교적 간편한 결합형 계층군집화를 사용하여 초기해를 생성하는데, 계층군집화를 수행할 때 중요한 과제 중 하나는 두 군집간의 거리를 어떻게 설정하는가 하는 문제이다. 군집간의 거리 설정에 흔히 사용하는 방법으로는 두 군집에 각각 속하는 개체들 중 가장 가까이 위치한 것들 사이의 거리로 정하는 최단거리(single-link) 방법, 가장 멀리 떨어진 개체들 간의 거리로 정하는 최장거리(complete-link) 방법, 두 군집의 중심을 구하여 중심간의 거리(centriod distance)로 정하는 방법 등이 있다[5]. 군집간의 거리의 차이는 (결합)계층군집화에서 군집들의 결합순서를 다르게 하여 군집화 결과에 큰 차이를 야기시킬 수 있으므로 데이터의 구조에 맞추어 적절하게 설정해야 한다. 특히 주목해야 할 점은 최단거리 방법은 별개의 군집으로 볼 수 있는 두개의 군집을 하나로 연결시키는 연쇄화(chaining) 경향을 보이는 경우가 있고, 최장거리 방법, 중심간의 거리 방법 등은 길쭉하게 늘어선 데이터를 구 모양으로 군집화 시키는 경향이 있다는 점이다([9] 등). 본 연구에서는 상호보완적인 최단거리 방법과 최장거리 방법으로 각각 계층군집화를 수행하여 얻어지는 두 개의 결과 중에서 기준함수 값이 더 우월한 것을 초기해로 설정한다. 이렇게 함으로써 데이터의 구조에 주의를 기울일 필요 없이 일반적으로 보다 우월한 군집화로 알고리듬을 시작할 수 있게 된다.

3.1.2 군집기준함수

본 연구에서는 계산상의 효율성과 군집의 크기 및 군집수의 영향을 보다 효과적으로 반영하기 위해 2.1절에서 예시한 표준적인 기준함수들은 변형하여 사용한다. 우선 각 군집의 개체들 간의 평균거리 $d(S_i)$를

$$d(S_i)=\frac{2}{n(i)\cdot(n(i)-1)}\sum_{k,k\in S_i l\neq k}\|x_k-x_l\| \qquad (6)$$

로 계산한다. 이 값이 작을수록 군집의 동질성은 커질 것으로 기대할 수 있다. 또한 각 군집의 개체수에 따라 가중치를 주어

$$W_2(S)=\sum_{i=1}^{c}\frac{n(i)}{n}d(S_i)$$

와 같이 계산하여 군집의 규모에 따라 영향력을 달리 반영하면 결국

$$W_2(S)=\frac{2}{n}\sum_{i=1}^{c}\frac{1}{n(i)-1}\sum_{k\in S_i l\neq k}\|x_k-x_l\| \qquad (7)$$

와 같은 형태가 얻어지는데 여기에서 상수는 무시하고 식 (2)와 같이 군집수에 따른 벌금 효과를 덧붙여

$$f(c,S)=\beta c\sum_{i=1}^{c}\frac{1}{n(i)-1}\sum_{k\in S_i l\neq k}\|x_k-x_l\| \qquad (8)$$

와 같은 형태의 최종적인 기준함수를 얻는다. 이 함수는 ASA 군집화의 전체 과정에서 최소화의 기준함수로 사용되는데 식 (8)에서 β는 대략[0.3,0.5]의 상수로 적절하게 놓으면 모의어닐링 과정에서 수렴을 잘 하는 것을 관찰할 수 있었다.

3.1.3 Mojena 방법의 개선

원래의 Mojena 방법에서는 군집간의 거리로부터 얻어지는 결합수준을 비교하여 적절한 군집수를 결정하였는데, 본 논문에서는 이 방법을 약간 수정하여 적용한다. 우선 ASA 군집화의 최적화 기준과 통일시키기 위해 결합수준 대신에 식 (8)과 같은 기준함수를 설정하여 사용한다. 또한 기준함수 자체의 평균, 표준편차를 이용하는 대신에 기준함수 증가폭을 통계적으로 비교하여 적절한 군집수 판별기준을 얻는다. 이 방법은 Mojena의 방법처럼 전체 기준함수를 고려할 필요가 없이 증가폭의 상대적인 크기만을 고려하기 때문에 군집수의 상한에서부터 시작하여 계산과정을 단순화 시킬 수 있고 정확성에서도 더 우월할 것으로 기대된다. 이 방법에서는 증가폭과 그것의 평균 간의 편차의 크기가 비교적 작은 점을 감안해서 보통 사용하는 제곱합의 평균 대신 절대값의 평균을 사용한다. 즉 최대군집수 C에서 1개 군집으로 결합하는 단계까지의 기준함수 값을 각각 $f_j, j=C,C-1,\cdots,1$로부터 증가폭 $\Delta_j=f_{j+1}-f_j,\ j=C-1,\cdots,1$ 을 계산하여 평균과 평균편차를 각각

$$\overline{\Delta}=\frac{1}{C-1}\sum_{j=1}^{C-1}\Delta_j,$$

$$a_\Delta=\frac{1}{C}\sum_{j=1}^{C}|\Delta_j-\overline{\Delta}|$$

과 같이 구한 후 판별기준값

$$\gamma=\overline{\Delta}+k\cdot a_\Delta$$

을 구한다. 이제 $j=C,C-1,\cdots,1$로 줄여가면서 $\Delta_{j+1}>\gamma$를 만족하는 최초의 j를 선택하여 j를 군집수로 설정한다. 만일 그런 j가 존재하지 않으면 개체 전체를 하나의 군집으로 간주한다. 여기서 실험을 통해 k을 1.5~3.5사이에서 취했을 때 상대적으로 좋은 결과를 얻었는데 본 논문 중 실제 분석에서는 k 값을 2.5로 취하였다.

3.2 ASA 군집화를 통한 반복적 개선

ASA 군집화 방법은 적절하게 주어진 초기 군집화 결과를 초기해로 입력받아 모의어닐링 기법에 기반한 ASA 알고리즘으로 이를 반복적으로 개선하여 최적에 가까운 군집화를 얻을 수 있게 한다. 또한 군집수를 변동시키는 해의 변동 과정이 적절히 삽입되어 초기분할에서 주어진 군집수를 자동적으로 변동시키면서 최종적으로 적절한 군집수와 동시에 군집화 결과를 얻을 수 있게 해준다(알고리즘에 대한 보다 상세한 설명은 윤복식[1] 참조).

4. 실험을 통한 검증

4.1 데이터 설명

여기서는 3장의 군집수 결정방법을 검증하기 위해 <표 1>과 같이 특별히 의도된 구조의 모의데이터에 대한 분석을 실시한다. 여기서 데이터 1~3은 2차원 평면상의 군집들이고 데이터 4와 5는 군집간에 다소 겹치는 부분이 있는 3차원 구(球)형군집 데이터와 각 군집에 포함된 개체수가 서로 다른 4차원 구형군집 데이터이다.

일반적으로 최종결과에서 얻어진 군집수는 분석

〈표 1〉 5가지 모의데이터

2군집 정규분포 데이터	모의데이터 1 (개체수 30)		
두 분포 평균사이의 거리 2, 분산은 동일	군집 1 : 평균벡터(1, 1), 분산 0.3 군집 2 : 평균벡터(3, 1), 분산 0.3		
3군집 데이터	모의데이터 2 (개체수 30)		
3개 구역에서 각기 10개 개체를 무작위로 추출	$R_1 = \{(x,y)\|0.2 \le x \le 0.5, 0.1 \le y \le 0.3\}$; $R_2 = \{(x,y)\|0.6 \le x \le 0.9, 0.1 \le y \le 0.3\}$ $R_3 = \{(x,y)\|0.5 \le x \le 0.6, 0.4 \le y \le 0.7\}$		
5군집 데이터	모의데이터 3 (개체수 40)		
5개 구역에서 각기 8개 개체를 무작위로 추출	$R_1 = \{(x,y)\|0 \le x \le 0.5, 0 \le y \le 0.5\}$; $R_2 = \{(x,y)\|1.5 \le x \le 2, 0 \le y \le 0.5\}$ $R_3 = \{(x,y)\|1.5 \le x \le 2, 1.5 \le y \le 2\}$; $R_4 = \{(x,y)\|0 \le x \le 0.5, 1.5 \le y \le 2\}$ $R_5 = \{(x,y)\|0.75 \le x \le 1.25, 0.75 \le y \le 1.25\}$		
3차원 구형군집 데이터	모의데이터 4 (개체수 60)		
군집별 개체수	20	20	20
평균	(10, 10, 10)	(20, 20, 20)	(30, 30, 30)
표준 편차(동일)	3	3	3
4차원 구형군집 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(동일)	2	2	2

과정에 사용된 군집화 방법과 밀접한 관계가 있다. 즉, 똑같은 군집수 결정방법이라도 어떤 군집화 방법에서 적용되었는가에 따라 결과에서 큰 차이를 보일 수 있어서 군집화 방법의 선택이 매우 중요하다. 본 연구에서 언급하는 ASA 군집화 방법은 1단계에서 계층적 군집화 방법인 최단거리와 최장거리 방법을 각각 적용하여 초기해를 결정하기 때문에 이 두 가지 계층적 군집화 방법을 적용한 군집수 결정 결과도 함께 제시한다.

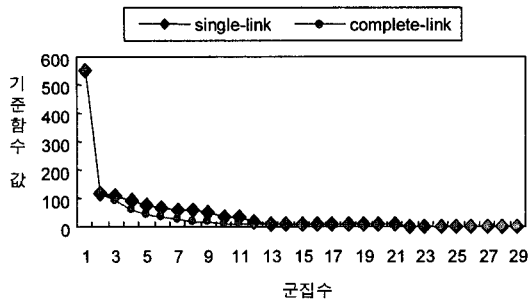
4.2 실험 결과

우선 군집수에 대한 기준함수의 변화를 나타내는 그래프를 [그림 1]~[그림 5]에서 볼 수 있다. 여기서 원래 주어진 군집수까지 기준함수 값이 급격히 감소하다가 그 이후에는 감소폭이 현저하게 줄어드는 것을 관찰할 수 있다. 이것은 우리가 설정한 군집함수가 군집 결과의 적절성을 판단하는 기준으로 적절하다는 것을 보여주는데, 이러한 그래프를 잘 관찰하면 데이터에 내재된 군집수를 추측할 수 있어서 그래프를 이용한 방법은 ASA 군집화의 초기해를 제공하는 대안적 도구로써 유용하게 보인다.

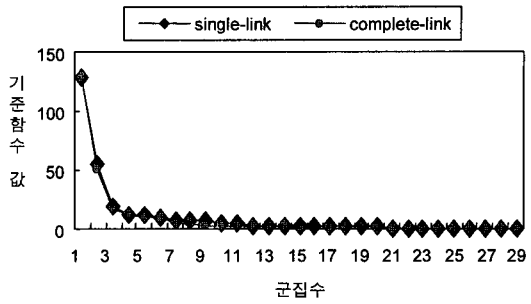
또한 <표 2>에서 군집수 결정방법 들의 타당성을 비교, 검증하기 위해서 모의데이터 1부터 5에 대한 실험결과를 제시하였다.

우선 ASA 군집화 방법이 과연 초기해의 군집수를 변동해 가면서 적절한 군집수를 선택할 수 있는지에 대한 검증을 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 2>의 첫 번째 행의 결과를 보여 주었다. 모의어닐링에서의 내부루프의 반복횟수를 100으로 설정했을 때 ASA 군집화는 평균적으로 93%정도가 데이터의 원래 구조를 찾아갔고, 나머지 실험에서도 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.

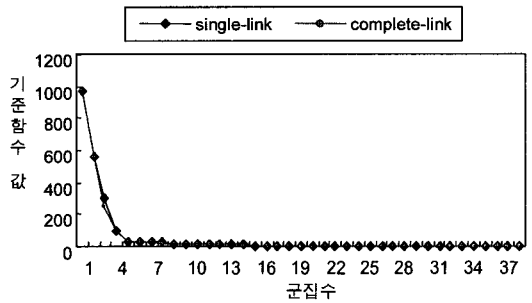
그러나 Mojena 방법은 k값을 2.5부터 3사이에서



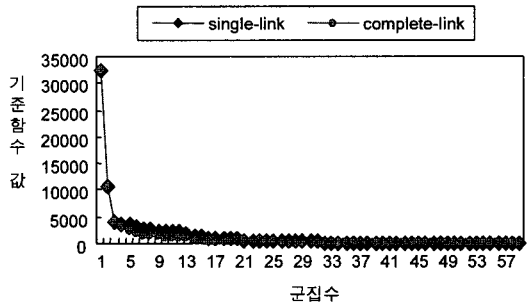
[그림 1] 기준함수 값의 변화(모의데이터 1)



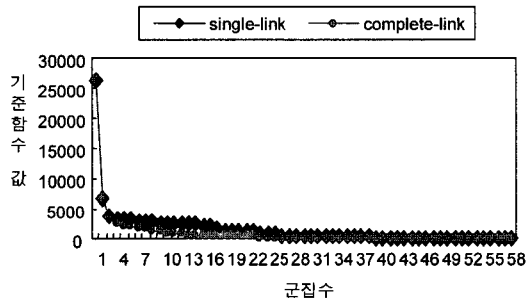
[그림 2] 기준함수 값의 변화(모의데이터 2)



[그림 3] 기준함수 값의 변화(모의데이터 3)



[그림 4] 기준함수 값의 변화(모의데이터 4)



[그림 5] 기준함수 값의 변화(모의데이터 5)

〈표 2〉 모의데이터 1~5에 대한 실험결과 : 군집수

사용된 방법	1	2	3	4	5
ASA 군집화 방법	2개 (95%)	3개 (95%)	5개 (95%)	3개 (90%)	3개 (90%)
Mojena 기준에 의한 방법					
최단거리방법	6개	12개	5개	1개	1개
최장거리방법	2개	5개	5개	1개	1개
기준함수의 증가폭에 의한 방법					
최단거리방법	6개	4개	5개	5개	4개
최장거리방법	2개	3개	5개	3개	3개

취해 가면서 적용했는데도 불구하고 두 가지 계층적 군집화 방법 모두, 특히 최단거리 방법에서 데이터의 원래 군집수를 정확히 결정하지 못한 것으로 나타났다.

반면 본 연구에서 제시한 군집기준치 증가폭에 의한 개선된 Mojena방법은 적어도 최장거리 방법에서는 모두 군집수를 정확히 결정하였다. 최단거리 방법에서도 균일분포를 따르는 3번 데이터에서는 적절한 군집수를 선택해 주었다. 비록 나머지 데이터에서 최단거리 방법에 의한 결과가 좋지는 않았지만(최단거리 방법 자체가 이 데이터들에 대한 적합한 분할방법이 아닌 것으로 보인다), 전체적으로 이 방법이 Mojena방법에 비해 보다 효과적인 것으로 보이고 계산과정도 보다 간편하였다. 따라서 ASA 군집화의 초기해를 제공하는 도구로써 무난해 보인다.

5. 실제 문제에의 적용 사례

5.1 웹 문서의 분류

본 절에서는 웹 문서들을 유사한 특성을 가진 것끼리 분류하는 사례를 보인다. 분류를 위해서는 우선 각 웹 문서가 가진 특성을 정량화해야 하는데, 본 연구에서는 미리 키워드를 적절한 갯수만큼 설정하여 각 키워드의 출현빈도수로 그 문서의 특징을 표현하기로 한다. 즉, n개의 문서(개체)가 주어지고 m개의 키워드가 설정되었을 때, $x_{ij} \in \{0,1,2...,\}$를 j번째 키워드가 문서 i에 나타나는 빈도수라고 하면 i번째 문서의 특성은 속성벡터 $\boldsymbol{x_i}=(x_{i1},x_{i2},\cdots,x_{im})$로 표현되고 전체 데이터는 집합 $S=\{\boldsymbol{x_1},\boldsymbol{x_2},\cdots,\boldsymbol{x_n}\}$로 나타낼 수 있다.

이제 개체간의 거리를 설정해야 하는데, 적어도 다음 두 가지 요건을 만족해야 할 것이다.

(i) 두 문서에서 동일한 키워드의 빈도수의 차이가 크면 두 개체간의 거리가 크다.

(ii) 빈도수의 차이가 동일한 경우에는 빈도수 자체의 크기가 작은 경우가 거리가 더 클 가능성이 많다. 극단적으로 어떤 키워드가 아예 나오지 않는 경우($x_{ij}=0$)와 한번 나오는 경우의 차이는 10번 나오는 경우와 11번 나오는 차이보다 훨씬 클 것이다.

이런 요건에 따라 키워드들이 상호간에 독립적이라는 가정 아래 두 문서간의 거리를

$$\|\boldsymbol{x_i},\boldsymbol{x_k}\| = \sum_{j:\max\{x_{ij},x_{kj}\}\neq 0} \frac{|x_{ij}-x_{kj}|}{\max\{x_{ij},x_{kj}\}} \tag{9}$$

$$\|\boldsymbol{x_i},\boldsymbol{x_k}\| = \sum_{j} \frac{|x_{ij}-x_{kj}|}{\min\{x_{ij},x_{kj}\}+1} \tag{10}$$

등과 같이 정의할 수 있는데 본 논문에서는 식 (9)를 사용하여 실험한다.

주어진 웹 문서 데이터는 <표 3>과 같은 형태인데, 여기에는 빈도수가 낮은 키워드들은 생략되어 있다.

이 데이터에 식 (9)를 적용하여 개체들간의 거리를 계산하여 거리행렬을 만들면 <표 4>가 얻어진다.

〈표 3〉 웹 문서 데이터

문서	키워드수	키워드 및 빈도수 (속성변수)				
1	75	8 web	4 site	3 page	2 directed	.. ...
2	60	5 web	3 graph	2 corresponds	2 directed	.. ...
3	77	9 web	5 links	4 multiple	4 node	.. ...
4	66	11 web	4 visualization	3 site	3 this	.. ...
5	62	14 web	5 links	5 page	5 2	.. ...
6	61	10 web	4 page	4 represents	4 set	.. ...
7	134	18 web	8 link	7 page	6 site	.. ...
8	47	4 web	3 visualization	2 overview	2 conclusion	.. ...

〈표 4〉 거리행렬

	문 서 번 호							
	1	2	3	4	5	6	7	8
1	0.0	87.45	123.36	123.27	119.58	113.78	155.84	108.42
2		0.0	114.83	98.72	98.36	100.17	109.43	70.03
3			0.0	131.15	57.42	122.10	172.87	117.22
4				0.0	108.60	110.01	153.39	45.22
5					0.0	100.39	158.64	100.71
6						0.0	162.29	99.10
7							0.0	147.78
8								0.0

이제 3장의 군집수 결정방법을 통해 초기 군집수를 3이나 4로 선택하면 상대적으로 적절한 것으로 나타나 군집수를 3으로 선택한 후 두 가지 계층적 군집화 방법과 ASA 군집화 방법으로 얻은 결과들을 <표 5>와 <표 6>에 제시하였다. <표 5>에서 보듯이 이 데이터의 경우는 최장거리와 최단거리에 의한 계층군집화가 동일한 군집화 결과를 보이고 있고 <표 6>에서 ASA 방법으로 얻은 대부분 결과가 비교적 짧은 시간 내에 기준함수의 값을 크게 줄여 주는 것을 관찰할 수 있다. <표 6>은 ASA 군집화를 50번 실시해서 얻은 결과 중에서 가장 우수한 결과를 제시한 것인데 대부분의 실험결과가 유사한 결과를 보여주었다. 비록 초기해(초기 3-분할을 서로 달리 취했을 때)에 따라 수렴시간은 다소 차이를 보였지만 초기분할과 거의 관계없이 대부분 최적의 결과에 접근하였다.

〈표 5〉 최단거리 및 최장거리 계층군집화에 의한 결과

군집 번호	군집 크기	군집별 개체
1	5	1, 2, 4, 8, 6
2	2	3, 5
3	1	7
기준함수 값 : 7514.78		

〈표 6〉 ASA 군집화 방법 의 결과

군집 번호	군집 크기	군집별 개체
1	3	2, 4, 8
2	3	3, 5, 6
3	2	7,1
기준함수 값 : 5333.83	실행 시간 : 0.94(sec)	

5.2 대학강사들의 분류

정성적 데이터에 대한 두 번째 적용사례로 <표 7>의 Dorndorf and Pesch[3]에 주어진 데이터의 일부)에 대한 분할문제를 다룬다. <표 7>은 50명의 대학강사들의 소속된 학회 및 관심분야를 보여주고 있는데 이것에 따라 대학강사들을 적절한 군집으로 분류하는 것이 목적이다. 우선 매개체(강사)마다 10개의 속성변수 v_{ij}를

〈표 7〉 대학강사 데이터 표

Specification of all 10 attributes

column	attributes
1/2	Member of the commission/Interest in International Management
3/4	Member of the commission/Interest in Information Management
5/6	Member of the commission/Interest in Operations Research
7/8	Member of the commission/Interest in Production Management
9/10	Member of the commission/Interest in Marketing

No.	attributes	No.	attributes	No.	attributes	No.	attributes	No.	attributes
1	1 2 9 10	11	3 4	21	6 8	31	5 6 7 8	41	6 7 9
2	3 4 6	12	3 4 5 6 7 8	22	10	32	8	42	9 10
3	5 6 8	13	4 8	23	2	33	3 5	43	4
4	1 2	14	2 9 10	24	3 5 7 9	34	4 5 6	44	6
5	6	15	4 6 8	25	9	35	4 8	45	6 7 8
6	7 8	16	5 6 7 8	26	6	36	5 6 7 8	46	3 4 5 6 9 10
7	3 4	17	3 4 9 10	27	4 6	37	6	47	9
8	1 3 9	18	1 3 5 7 9	28	4 6 8	38	7 8	48	9
9	5 6	19	4 5 6 7 8	29	3 4 6 8	39	1 2	49	1 2
10	6 8 9	20	2 4 6	30	7 8	40	9	50	6 8

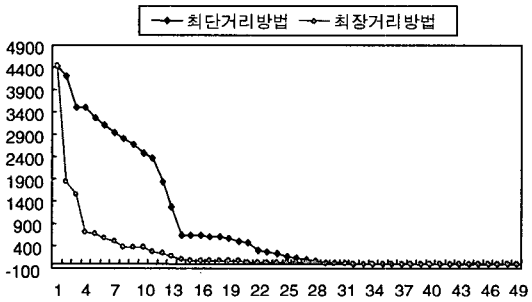
$$v_{ij}=\begin{cases}1 & i\text{번째 사람이 } j\text{학회의 소속,}\\ & \text{또는 } j\text{분야에 관심이 있을경우}\\ 0 & \text{그외의 경우}\end{cases}$$

과 같이 설정하고 데이터행렬을 $V=(v_{ik})_{50\times 10}$로 표기한 후, 서로 다른 개체간의 거리를

$$d_{ik}=10-|\{j\,|\,v_{ij}=v_{kj}, j=1,...,10\}|,\quad (1\le i<k\le 50)$$

과 같이 정의한다.

우선 군집수 결정방법을 통해 초기해의 군집수를 선택한 결과 초기해의 군집수를 14로 선택하면 적절할 것으로 나타났는데 군집수에 따른 기준함수의 그래프를 그림에서 볼 수 있다.



[그림 6] 군집수에 따른 기준함수값의 변화

〈표 8〉 최단거리방법에 의한 결과

군집 번호	군집 크기	각 군집에 포함된 개체
1	7	1, 14, 25, 40, 42, 47, 48
2	23	5-7, 9, 11, 13, 15, 21, 26-30,2, 32,34, 35, 37, 38, 43-45, 50
3	5	3, 16, 19, 31, 36
4	4	4, 23, 39, 49
5	1	8
6	1	10
7	1	12
8	1	17
9	2	18, 24
10	1	20
11	1	22
12	1	33
13	1	41
14	1	46
기준함수 값 : 624.0		

〈표 9〉 최장거리방법에 의한 결과

군집 번호	군집 크기	각 군집에 포함된 개체
1	6	1, 4, 14, 23, 39, 49
2	4	2, 7, 11, 29
3	8	3, 5, 9, 21, 26, 37, 44, 50
4	4	6, 30, 32, 38
5	5	8, 25, 40, 47, 48
6	2	10, 41
7	2	12, 19
8	5	13, 15, 28, 35, 43
9	4	16, 31, 36, 45
10	2	17, 46
11	2	18, 24
12	3	20, 27, 34
13	2	22, 42
14	1	33
기준함수 값 : 93.0		

〈표 10〉 ASA 방법에 의한 결과

군집 번호	군집 크기	각 군집에 포함된 개체
1	4	4, 23, 39, 49
2	4	2, 7, 11, 33
3	5	5, 9, 26, 37, 44
4	4	6, 30, 38, 45
5	5	25, 40, 42, 47, 48
6	3	20, 27, 34
7	4	3, 16, 31, 36
8	4	13, 32, 35, 43
9	3	12, 19, 29
10	4	15, 21, 28, 50
11	3	8, 18, 24
12	2	17, 46
13	3	1, 14, 22
14	2	10, 41
기준함수 값 : 62.0	실행 시간 : 1.81(sec)	

<표 8>~<표 10>에서는 3가지 방법으로 얻은 결과를 제시했는데, <표 10>에 보듯이 ASA 군집화 방법은 두 계층적 군집화 방법으로 얻은 결과의 기준함수의 값을 크게 줄여 주었다. ASA의 실험조건 및 파라미터는 5.1절에서와 유사하고 내부루프의 반복횟수는 50으로 주었다.

6. 결 론

본 논문에서는 ASA 군집화를 기반으로 하여 군집수를 결정하는 방법이 소개되었다. 물론 ASA 군집화는 최적화의 관점에서 군집수를 자동 결정해주는 특성이 있지만 알고리즘의 효율성을 위해 초기해로 되도록 우수한 해가 제공될 필요가 있다. ASA 군집화의 초기해는 어떠한 방법으로 얻어지든 상관없으나 계산이 비교적 간편한 계층군집화를 이용하면 좋은데 3장에서는 계층군집화에서 보다 우수한 군집수와 군집결과를 얻기 위한 방안들이 소개되었다. 제시된 방법들은 다양한 모의데이터에 적용되어 타당성이 검증되었고, 웹 문서 분류와 강사의 전공별 분류의 두 종류의 실제 정성적인 데이터에 적용되는 사례를 보였다. 군집기준의 군집수에 대한 그래프를 통하여 최종적으로 얻어진 군집수가 실제 데이터의 특성을 잘 반영함을 또한 확인하였다.

윤복식[1]에서 ASA 군집화가 정량적인 데이터의 군집화에 매우 효과적으로 적용될 수 있음을 보인 바 있는데 본 연구를 통하여 정성적인 데이터에도 무리없이 적용될 수 있음을 알 수 있었고 군집수를 결정하는 문제에도 유효함을 확인하였다. 앞으로 ASA 군집화 방법을 현실에서 대두되는 다양한 군집화에 적용하는 연구가 계속될 것이다.

참 고 문 헌

[1] 윤복식, "최적에 가까운 군집화를 위한 이단계 방법", 「한국경영과학회지」, 제29권, 제1호 (2004), pp.43-56.

[2] Bock, H.H., "Probability Models and Hypothesis Testing in Partitioning Cluster Analysis," in P. Arabie, L.J. Hubert, and G. De Soete (Eds), *Clustering and Classification*, World Scientific, Singapore, (1996), pp.377-453.

[3] Dorndorf, U. and E. Pesch, "Fast Clustering Algorithms," *ORSA Journal on Computing*, Vol.6, No.2(1994), pp.141-153.

[4] Geva, A.B. and Y. Steinberg, "A Comparison of Cluster Validity Criteria for a Mixture of Normal Distributed Data," *Pattern Recognition Letters*, Vol.21(2000), pp.511-529.

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

[6] Hardy, A., "On the Number of Clusters," *Computational Statistics & Data Analysis*, Vol.23(1996), pp.83-96.

[7] Jain, A.K. and J.N. Moreau, "Bootstrap Techniques in Cluster Analysis," *Pattern Recognition*, Vol.20(1987), pp.547-568.

[8] Kothari, R. and D. Pitts, "On Finding the Number of Clusters," *Pattern Recognition Letters*, Vol.20(1999), pp.405-416.

[9] Manly, B.J., *Multivariate Statistical Methods* (2nd ed.), Chapman & Hall, London, 1994.

[10] Milligan, G.W. and M.C. Cooper, "An Examination of Procedures for Determining the Number of Clusters in a Data Set," *Psychometrika*, Vol.50, No.2(1985), pp.159-179.

[11] Milligan, G.W., "Clustering Validation : Results and Implications for Applied Analysis," in *Clustering and Classification*,P. Arabie, L.J. Hubert, and G. De Soete (Eds), World Scientific, Singapore, (1996), pp. 341-375.

[12] Mirkin, B., *Mathematical Classification and Clustering*, Kluwer Academic Publishers, 1996.

[13] Mojena, R., "Hierarchical Grouping Methods and Stopping Rules : An Evaluation," *Computer Journal*, Vol.20(1977), pp.359-363.

[14] Nakamura, N. and N. Kehtarnavaz, "Determining Number of Clusters and Prototype

Locations Via Multi-scale Clustering," *Pattern Recognition Letters*, Vol.19(1998), pp.1265-1283.

[15] Peck, R., L. Fisher, L. and J.V. Ness, "Approximate Confidence Intervals for the Number of Clusters," *Journal of the American Statistical Association*, Vol.84 No.405 (1989), pp.184-191.