韓國經營科學會誌
第29卷 第1號
2004年 3月

최적에 가까운 군집화를 위한 이단계 방법*

윤 복 식**

A Two-Stage Method for Near-Optimal Clustering*

Bok Sik Yoon**

■ Abstract ■

The purpose of clustering is to partition a set of objects into several clusters based on some appropriate similarity measure. In most cases, clustering is considered without any prior information on the number of clusters or the structure of the given data, which makes clustering is one example of very complicated combinatorial optimization problems. In this paper we propose a general-purpose clustering method that can determine the proper number of clusters as well as efficiently carry out clustering analysis for various types of data. The method is composed of two stages. In the first stage, two different hierarchical clustering methods are used to get a reasonably good clustering result, which is improved in the second stage by ASA(accelerated simulated annealing) algorithm equipped with specially designed perturbation schemes. Extensive experimental results are given to demonstrate the apparent usefulness of our ASA clustering method.

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

1. 서 론

군집화(clustering)는 군집의 특성에 대한 사전 정보가 주어지지 않은 상태에서 주어진 데이터를 유사한 것 끼리 묶어 군집을 형성하는 기법이다. 군집화가 잘 이루어지면 데이터를 군집별로 묶어 특성을 파악할 수 있게 되어 데이터 규모를 축소시킬 수 있고, 또한 일차적인 데이터가 가지고 있는

논문접수일 : 2003년 11월 25일 논문게재확정일 : 2003년 12월 31일
* 이 논문은 2003년 홍익대학교 학술연구조성비에 의하여 연구되었음.
** 홍익대학교 기초과학과 응용수학전공

복잡한 구조로부터 보다 단순한 구조와 형태를 얻을 수 있어, 군집화는 패턴 분류, 생태학, 데이터 마이닝을 비롯한 매우 다양한 분야에서 응용 되고 있다(Everitt, 1980, 1991 ; Manly, 1994 ; Duda *et al.*, 2001 등 참조). 응용범위가 점차 넓어지고 계산 수단이 발달함에 따라 실제문제에 광범하게 적용할 수 있는 군집화 방법의 개발 및 관련 연구들이 활발히 진행되어 왔다. 그러나 대부분 경우 군집화가 데이터의 구조특성에 대한 사전 정보가 충분하지 않거나 또는 전혀 없는 상태에서 이루어지게 되므로 군집화는 본질적으로 매우 애매하고 또한 복잡한 조합최적화(combinatorial optimization) 문제가 된다(Mirkin, 1996).

군집화의 접근 방법은 크게 계층적 방법(hierarchical approach)과 최적화(optimization) 방법으로 나눌 수 있는데(Hand, 1981), 전자는 개체들 간의 거리가 가장 가까운 것 끼리 묶어 하위군집들을 형성하고 또 하위군집간의 거리를 비교하여 그 다음 계층의 군집을 형성하는 방식으로 진행되는데(결합법, agglomerative method) 계산이 용이해 많은 군집화에서 적용되지만 결과의 신뢰성이 때로는 문제가 된다. 반면에 최적화 방법은 군집화 문제를 조합최적화 문제로 모형화하여 접근하는 방식인데 최적 또는 최적에 가까운 군집화가 가능하나 개체의 수가 많아짐에 따라 계산량이 크게 증가하여 (NP-hard) 실제문제 해결에서 효율적이지 못할 수도 있다. 또 최적화 방법의 경우 군집화의 정확성을 반영하는 목적함수(군집화 기준함수)를 설정하여 이 목적함수를 최적화 하도록 군집 간 개체들을 이동시키게 되는데, 이 기준함수의 적절성이 문제가 될 수도 한다. 최적화 방법의 경우 대개 근사적인 접근을 하여 발견적(heuristic) 기법을 사용하는데 가장 대표적인 예로 k-평균 방법(Selm and Ismail, 1984 ; Kovesi *et al.*, 2001 등 참조)을 들 수 있다. 이 방법은 전체 데이터를 k개의 군집에 임의로 배정한 후 중심을 계산하여 개체들을 가장 가까운 중심의 군집에 재배정하고 또 중심을 수정하는 과정을 되풀이하는 일종의 국부탐색(local search) 방식으로, 사전에 군집수가 k개로 알아야 한다는 점과 최적의 해에 수렴성을 보장할 수 없다는 점을 단점으로 지적할 수 있다. 최근에 Sanghamitra *et al.*(1998, 2001)은 k-평균 방법에 모의어닐링(simulated annealing) 기법을 연결시킨 방법(SAKM)을 소개하고 실험을 통해 k-평균 방법에 비해 개선된 결과를 보여주었다. 그러나 이 방법도 군집수에 대한 사전 정보가 필요하다는 k-평균 방법의 한계를 여전히 벗어나지 못하고 있다.

현재까지 계층적 또는 최적화의 방식을 따라 매우 많은 군집화 방법들이 소개되어 있지만 다양한 형태의 문제에 일관되게 적용될 수 있는 범용의 군집화 방법을 찾기는 매우 힘들다. 본 논문은 군집수에 대한 사전정보가 불충분한 매우 일반적인 상황 하에서도 군집화 분석을 효과적으로 수행할 수 있는 범용의 방법을 개발하는 것을 주목적으로 한다. 이를 위한 첫 번째 착안은 계층적 군집화 방법의 간편성과 최적화 방법의 정확성을 함께 이용하는 것인데, 이에 따라 우선 계산상 효율적이지만 정확성이 다소 떨어지는 계층적 군집화 기법을 이용하여 1단계 군집화를 되도록 최적에 가깝게 수행한 후, 모의어닐링(simulated annealing)을 개선한 ASA(accelerated simulated annealing) 알고리듬(윤복식, 조계연, 1996)을 이용하여 1단계 군집화 결과를 최적에 가깝게 효과적으로 개선해 가는 2단계 기법인 ASA 군집화 방법을 제시한다. 특히 두 번째 단계에서 ASA 알고리듬의 해의 변동과정을 군집화 분석에 맞도록 적절히 설계하여 원하는 군집화가 정확히 또한 신속히 이루어지도록 시도하게 된다. 본 논문의 ASA 군집화는 주어진 기준에 가장 적합한 군집수와 군집구조를 자동적으로 함께 결정하여 주므로, 군집수에 대한 사전 정보가 불충분한 일반적인 문제에 대해서도 간편하게 적용될 수 있다. 일반적으로 모의어닐링 방법은 주어진 목적함수에 대해 최적해 또는 최적에 매우 가까운 해를 주게 되므로 본 논문의 방법은 주어진 목적함수(군집화 기준)에 대해 적어도 최적에 가까운(near-optimum) 군집화를 주게 된다.

본 논문의 나머지 부분은 다음과 같이 구성된다. 우선 서론에 이어 2장에서 군집화 문제와 군집화의 기준이 되는 목적함수를 설명하고 3장에서 ASA 군집화의 과정을 소개한다. 4장에서는 이 방법의 타당성을 검증하기 위해 구조가 이미 알려진 데이터들을 생성하여 군집화를 수행한 결과를 제시하고 5장에서는 기존에 군집화가 수행된 바 있는 실제 데이터들에 대해 군집화를 수행하여 결과를 비교 검증하고 6장에서 결론 및 향후 연구방향을 논한다.

2. 군집화 문제와 군집화의 목적함수

n개의 개체로 이루어진 데이터가 $S=\{ \boldsymbol{x}_1, \boldsymbol{x}_2 \cdots, \boldsymbol{x}_n\}$가 주어질 때 S를 서로 겹치지 않는 c개 부분집합인 $\{S_1, S_2, \cdots, S_c\}$으로 분할($c$-분할)하되 주어진 목적함수를 최적화시키는 c-분할을 찾는 문제를 군집화 문제라고 볼 수 있다. 이것은 전형적인 NP-hard 문제인데, 군집수 c가 주어지지 않으면 더욱 복잡한 문제가 된다(Mirkin, 1996).

군집화의 목적함수는 서로 다른 군집의 개체들 간의 이질성과 같은 군집내의 개체들의 동질성을 합리적으로 반영해야 하는데 군집화의 타당성을 결정하는 아주 중요한 부분이다. 군집화의 목적함수를 설정하가 위해서는 우선 개체간의 유사성을 반영하는 척도인 거리를 설정해야 한다. 주어진 문제의 특성에 따라 적절한 거리를 설정해야 하는데 정량적인 데이터의 경우에는 L_p-norm을 주로 사용한다. 즉 $\boldsymbol{x}, \boldsymbol{y}$을 각기 m개 속성을 갖는 개체라고 할 때(즉 m-차원 벡터), 거리공식은 다음과 같다.

$$\| \boldsymbol{x}, \boldsymbol{y} \| = \left(\sum_{i=1}^{m} |x_i - y_i|^p \right)^{1/p}$$

본 논문의 4장, 5장에서는 $p=2$인 경우인 유클리드 거리를 사용할 것이다.

문제의 특성에 따라 다양한 군집화 목적함수들이 가능한데 전통적으로 많이 사용되고 있는 것은 군집내부 편차(variance of the within-cluster) 기준,

$$w_c = \frac{1}{n} \sum_{i=1}^{c} \sum_{\boldsymbol{x}_j \in S_i} \| \boldsymbol{x}_j - \overline{\boldsymbol{x}}_{S_i} \|$$

과 군집 간 편차(variance of the between-clusters) 기준,

$$b_c = \frac{1}{n} \sum_{i=1}^{c} n_i \| \overline{\boldsymbol{x}}_{S_i} - \overline{\boldsymbol{x}} \|$$

이다. 여기서 $n_i = |S_i|$ (즉, 원소의 개수)이고 $\overline{\boldsymbol{x}}$, $\overline{\boldsymbol{x}_{S_i}}$등은 평균을 의미한다. 이때 군집화 문제는 w_c를 최소화 하거나 또는 b_c를 최대화 하는 분할을 찾는 문제가 된다. w_c을 최소화하는 c-분할은 b_c을 최대화하므로 비율을 고려한 비율기준 $r_c = b_c / w_c$이 사용되기도 한다. 이런 기준들을 적절히 변형하여 실제 문제에서 군집화 기준으로 이용될 수 있다. 그러나 일반적으로 많이 알려진 기존의 군집화 알고리듬(k-평균방법 등)들은 다음과 같은 형태의 목적함수

$$\sum_{i=1}^{c} \sum_{j,k \in S_i, k \neq j} \| \boldsymbol{x}_j - \boldsymbol{x}_k \|$$

을 최소화하는 c-분할을 찾게 된다. 그러나 이런 형태의 목적함수들은 대체로 군집수 c의 단조감소함수이므로 $c = n$일 때 최소값을 갖게 되어, 군집수가 사전에 주어지지 않아 군집수를 동시에 결정해야 하는 경우에는 타당성 척도로 사용하는데 문제점이 있다.

본 연구에서는 군집수에 대한 사전 정보가 불충분할 때에도 최적 군집수의 군집화을 얻도록 하기 위해 군집수 c의 증가에 따르는 페널티(penalty)를 목적함수에 반영하여 다음과 같이 정규화된(nor-

malized)된 군집화 목적함수를 정의한다.

$$W_c = (a \cdot c) \sum_{i=1}^{c} \frac{2}{n(n_i-1)} \sum_{j,k \in S_i, k \neq j} \| x_j - x_k \| \quad (1)$$

여기서 첫 번째 합 내부의 곱은 i군집내의 거리합을 평균한 후 군집 내 개체 수만큼 가중치를 주어 $\frac{2}{n(n_i-1)} = \frac{n_i}{n} / \frac{n_i(n_i-1)}{2}$ 과 같이 계산하여 구하였다. 또한 군집수에 대한 가중치 a의 값은 많은 실험을 통해 1.5와 2.5 사이에서 취하면 적합한 것으로 나타났다. 여기서

$$d(S_i) = \sum_{j,k \in S_i, j \neq k} \frac{2\|x_j - x_k\|}{n_i(n_i-1)}$$

는 군집의 평균지름을 나타내므로 $d(S_i)$가 작을수록 군집내부 개체들의 밀집해 있다고 간주할 수 있다.

3. 군집화 과정

본 연구의 군집화 과정은 두 단계로 이루어진다. 첫 번째 단계에서 적용이 상대적으로 편리한 계층적 군집화 방법으로 c-분할을 얻고, 두 번째 단계에서 가속화된 가상어닐링 기법을 적용하여 군집수와 군집을 변동시켜서 초기분할을 점차 개선하여 최적에 가까운 분할로 접근하게 된다. 가상어닐링 기법에서 초기해를 최적에 가깝게 시작하면 최적해로의 수렴속도가 빨라지는 것을 실험적으로 관찰할 수 있으므로 1단계에서 되도록 좋은 군집화를 얻는 것이 알고리즘의 속도개선에 유리할 것이다.

3.1 1단계 과정 : 초기해의 결정

본 연구에서는 계층적 결합법을 1단계에서 이용하여 ASA에 필요한 초기해를 얻게 된다. 계층적 결합법은 초기의 n개의 단일개체 군집으로부터 각 단계마다 제일 유사한 두 군집을 결합해 전체 개체가 하나의 군집으로 될 때까지 묶어 가는 과정이다 (사전에 군집수가 주어졌다면 중간과정에서 종료할 수도 있다). 이때 군집간의 유사성 척도인 거리를 어떻게 정의하는가에 따라 다양한 계층적 결합법을 얻게 된다. 많은 경우에 군집간의 거리에 따라 전혀 다른 군집화가 이루어지므로 이것의 선택도 신중해야 한다. 본 연구에서는 최단거리(single link)와 최장거리(complete link) 방법을 상호보완적으로 사용한다. 최단거리 방법에서는 두 군집간의 거리를 각각 다른 군집에 속한 개체들 중 가장 가까운 두 개체 사이의 거리로 정의하고, 반대로 최장거리 방법에서는 가장 멀리 떨어진 두 개체 사이의 거리로 정의한다. 이 두 가지 방법은 거리계산이 상대적으로 단순하여 적용이 편리하다는 장점이 있는 반면, 군집화에서 서로 다른 특성을 보이게 된다. 전형적으로 최단거리 방법은 가까운 개체들부터 연결해 주게 되므로 경우에 따라서는 원래 구조를 찾지 못하고 길게 연결된 형태의 군집을 만들어 주는 경향(chaining)을 보일 수 있고, 반면에 최장거리 방법은 길게 연결된 구조보다는 구형으로 응집된 군집을 만들어 가는 것을 관찰할 수 있다(Everitt, 1991 ; Hardy, 1996 등). 따라서 데이터에 구조에 따라 보다 적합한 방법이 다른데 서로 상반된 결과들 보이는 이 두 방법을 모두 사용하여 초기군집화를 시행하여 얻어 보다 좋은 목적함수 값을 주는 군집화 결과를 시발점으로 삼으면 효과적일 것 이다. 이 과정을 보다 구체적으로 표현하면 다음과 같다.

● **초기해의 결정**

단계 1. 계층 결합법의 최단거리와 최장거리 방법을 각각 적용하여 두 개의 c-분할을 얻는다. 이때 군집수가 사전에 알려져 있지 않으면 적절한 군집수 결정 방법(윤복식, 강금석, 2003 참조)을 적용하여 군집수를 결정한다. 얻어진 군집수를 c라고 하고 두 분할을 각각 $X_1 = \{S_1^{(1)}, \cdots, S_c^{(1)}\}$, $X_2 = \{S_1^{(2)}, \cdots, S_c^{(2)}\}$ 라고 하자.

단계 2. 대응되는 기준함수의 값을 비교하여,
① $f(X_1) \leq f(X_2)$이면, $X_0 = X_1$로 설정하고,
② $f(X_2) < f(X_1)$이면, $X_0 = X_2$ 로 설정한다.

여기서 f는 기준함수이고, X_0는 2단계 과정의 초기해이다.

3.2 2단계 과정 : ASA 군집화 방법

3.2.1 ASA 알고리듬

2단계 과정의 중심이 되는 ASA(accelerated simulated annealing)은 Kirkpatrick et al.(1983)에 의해 제안된 모의어닐링(SA, simulated annealing)을 개선한 방법이다. SA는 조합최적화에서 기존의 반복적인 개선에 근거한 국부탐색 방법들이 국부최소점(local minimal point)에서 빠져 나오지 못하여 전체최소점(global minimal point)에 수렴하지 못할 수도 있다는 단점이 있음을 보완하여, 열등해로의 상승이동을 확률적으로 허용하여 최적해에 수렴하도록 고안한 반복적 탐색기법의 일종이다. 이 기법은 유사한 기능을 수행하는 유전알고리듬이나 터부서치등의 다른 메타휴리스틱(metaheuristics) 기법에 비해 알고리듬의 구조가 단순하고 전체 최적해로의 수렴성이 이론적으로 증명될 수 있기 때문에 효과적인 최적화방법으로 폭넓게 주목받아 왔다(김여근 외, 1997). 그러나 SA도 다른 메타휴리스틱 기법과 마찬가지로 수렴속도가 느리다는 약점을 가지고 있는데 수렴속도를 증진시키기 위한 목적으로 다양한 변형 및 개선이 시도되었다. 윤복식, 조계연(1996)에 의해 제안된 ASA도 그 한 예인데, ASA는 알고리듬의 수렴성을 손상하지 않도록 표준 SA의 기본 구조를 유지하면서, 필요할 경우에만 내부루프를 반복시켜주는 방식으로 수렴시간을 크게 단축시켜 준다. 더욱이 사전에 기본적

```
INITIALIZE( X, T, L) ;
X_best= X ; counter1 = 0 ; counter2 = 0 ;
   repeat
      cost_old = f(X) ;  check = 0 ;
       for  i=1 to  L do
            X* = PERTURB( X ) ;
            if  f(X*) ≤ f(X)  or  exp[ -(f(X*) - f(X)) / T ] > random(0, 1))
            then   X = X* ;   /* 새로운 해를 받아들인다 */
            end if  ;
            if (f(X) < f(X)_best)) then
             X_best = X ;  counter2 = 0 ; check = 1 ; /* 만일 1이면 ;  T 을 조절한다 */
            else    counter2 = counter2+1 ;
            end if
       end for
       cost_new = f(X) ;
       UPDATE( T,  cost_new, cost_old, check) ;
       if (cost_new = cost_old)  then      counter1 = counter1 +1 ;
      else  counter1 = 0 ;
      end if ;
   until (counter1 > M or counter2 > N) ;
   UPDATE( T,  cost_new, cost_old , check)
        if (check1 = 1 or cost_new < cost_old )  then  T = aT, (0 < a < 1) ;
        end if
```

[그림 1] ASA 알고리듬

으로 설정해 주어야 하는 알고리듬 파라미터들을 거의 자동적으로 설정해 주기 때문에 다양한 문제들에 매우 효과적이고 간편하게 적용시킬 수 있다. 아래 표에는 ASA의 전체 알고리듬이 나와 있다.

3.2.2 ASA 군집화 알고리듬

ASA 알고리듬을 군집화 문제에 적용시키기 위해서 목적함수(군집화 기준함수)와 적절한 초기해, 해이동(PERTURB) 방법을 설정해 주어야 한다. 우선 목적함수는 식 (1)의 군집화 기준함수 W_c로 설정하되 a의 값을 2로 설정하였다. 이것은 문제의 성격에 따라 달라질 수 있다.

ASA의 경우 초기해 보다 우수한 해를 계속 탐색해가므로 초기해가 최적해에 가까울수록 알고리듬의 수렴속도가 빨라지리라고 기대할 수 있다. 따라서 단순한 방법으로 되도록 최적에 가까운 초기해를 선택하여 계산량을 줄이는 것이 바람직하다. 본 연구에서는 계층적 방법 중에서 상호 보완적인 최단거리(single-link)와 최장거리(complete-link) 방법으로 각기 군집화을 시행하여, 목적함수 값을 비교하여 우월한 것을 ASA의 초기해를 선택하는 방법을 택하였다.

◈ **근거리 이동과정** (확률 $a(T)$로 일어난다)
군집수가 c이고, 대응되는 분할 $S=\{S_1, S_2, \cdots, S_c\}$이 이라고 할 때,
(1) 주어진 개체집합 $S=\{x_1, x_2, \cdots, x_n\}$에서 무작위로 개체 x_i을 추출한다.
(2) 정수집합 $\{1, \cdots, c, c+1\}$에서 무작위로 정수 I를 추출한다.
만일 $I=c+1$이면, $c=c+1$이라고 놓는다 (즉 군집수를 하나 늘린다).
(3) 추출된 개체 x_i를 군집 I에 배정한다.

◈ **원거리 이동과정** (확률1- $a(T)$로 일어난다)
(1) 군집간의 거리가 제일 작은 두 군집을 통합한다.
(2) 원래의 군집의 개수를 c에서 $c-1$로 줄인다.

[그림 2] 해 이동과정[PERTURB]

해의 이동은 [그림 2]에서 보듯이 두 종류로 설정하였다. 하나는 군집수를 바꾸지 않고 개체들의 소속을 바꾸는 근거리이동(near movement)으로 발생빈도가 높도록 설계되었고, 다른 하나는 군집을 통합시켜 주는 원거리이동(far movement)으로 드물게 일어나도록 하였다. 근거리 이동 중에 군집이 늘어날 수도 있게 하여 해가 이동하는 가운데 군집수도 늘거나 줄 수 있으므로 적절한 군집수를 계속 탐색해 나가게 된다.

4. 모의 데이터에 대한 실험

4.1 5가지 모의 데이터

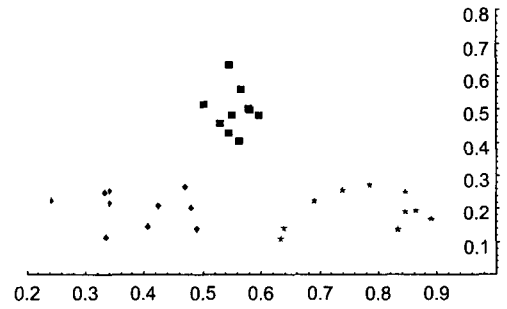
ASA 군집화 방법의 평가를 위해 정량적인 데이터에 대한 적용사례들을 살펴본다. 우선 의도된 군집구조를 가진 5가지 모의 데이터에 대해 ASA 군집화 방법이 데이터의 고유구조를 어느 정도 잘 찾아낼 수 있는지를 검정해 본다. 데이터 1-3은 시각적으로 구조를 알아보기 쉬운 2-차원 데이터 4,5는 고차원 데이터이다.

4.1.1 모의 데이터 1

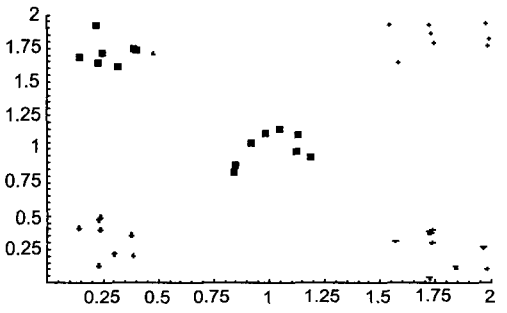
첫 번째 데이터는 3개의 서로 겹치지 않는 구역에서 추출된 표본인데, 개체들은 균질분포를 따르고, 각기 10개의 개체로 구성된다. 실제 추출된 개체들의 분포를 살펴보면 6번 개체와 12, 13번 개체가 자신들을 포함하고 있는 군집의 다른 개체들과 좀 떨어져 있지만 전체적으로 세 개의 군집으로 볼 수 있다([그림 3]). 그리고 설명의 편리를 위해 1~10번, 11~20번, 21~30번 개체들을 각기 순서대로 1, 2, 3번 군집과 대응 시켰다.

4.1.2 모의 데이터 2

군집수를 좀 늘려서 실행 효과를 살펴보기 위해 40개 개체들을 8개씩 각기 서로 분리된 5개 영역에서 무작위로 추출하였다([그림 4]). 개체들을 첫 번째 데이터와 유사한 순서로 각 군집에 배정하였다.



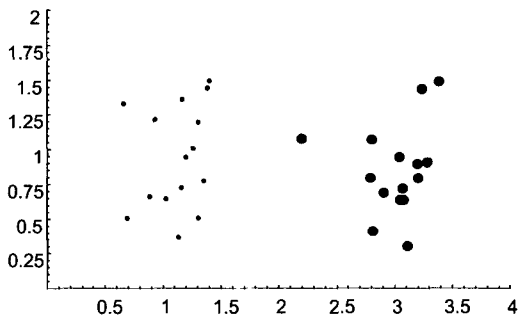
[그림 3] 모의 데이터 1



[그림 4] 모의 데이터 2

4.1.3 모의 데이터 3

세 번째 모의 데이터는 평균간의 거리가 2.0이고 표준편차가 0.3인 두 개의 각기 2-차원 정규분포를 따르는 개체들로 구성되었다(각 변수는 서로 독립이라고 가정한다). 그중 19, 30번과 24번 개체를 제외하고는 비교적 집중된 개체들로 구성되어 기본적으로 두 군집으로 나타내고 있다([그림 5]).



[그림 5] 모의 데이터 3

4.1.4 모의 데이터 4, 5

네 번째와 다섯 번째 모의 데이터는 <표 1>에서 주어진 각기 3-차원과 4-차원 정규분포를 따르는 두 개의 데이터인데, 각기 3개 군집의 총 60개 개체들로 구성되었다. 첫 번째 데이터는 군집간의 개체들이 일부분이 겹치는 경우가 된다.

4.2 실험 결과

본 연구에서는 ASA 알고리듬을 PC에서 실행시켰고, 외부루프의 반복횟수는 5, 내부루프의 반복횟수는 100으로 취했고, 초기온도 T는 5로 설정하고, $a = 0.95$, $T_{\min} = 0.0001$로 각기 설정하였다. 군집수는 미정으로 놓고 다양한 초기분할(상이한 군집수 및 주어진 군집수에 대응되는 여러 가지 초기분할)을 초기해로 설정하고 ASA 군집화를 수행하였다. 군집화 기준은 식 (1)의 $W_c(\alpha=2)$를 사용하였다. 우선 <표 2>에서는 모의 데이터 1에

〈표 1〉 두 가지 구형 군집 모의 데이터

군집 별 → 통계적 모수↓	군집 1	군집 2	군집 3
3-차원 데이터			
군집별 개체수	20	20	20
평균 벡터	(10, 10, 10)	(20, 20, 20)	(30, 30, 30)
표준편차(각 변수가 일치함)	3	3	3
4-차원 데이터			
군집별 개체수	30	20	10
평균 벡터	(5, 5, 5, 5)	(13, 13, 13, 13)	(20, 20, 20, 20)
표준편차(각 변수가 일치함)	2	2	2

대한 실험결과를 볼 수 있는데, 초기 군집수를 각각 2부터 10사이에서 취하고 이에 대응되는 분할도 다양하게 취했을 때, 50가지 서로 다른 초기해에 대한 실험에서 약 96%의 실험결과가 원래의 군집구조를 나타냈고, 그리고 평균 소요시간은 16(sec)이였다. 초기해에 따라 수렴속도는 약간의 차이를 보였지만 최종 결과에서는 거의 초기해와 상관없이 원래 구조를 찾아갔다.

〈표 2〉 모의 데이터 1에 대한 최적결과

군집번호	개체의수	군집별 개체번호
1	10	1~10
2	10	11~20
3	10	21~30
기준함수 값 : 0.039727 평균 소요시간 : 16(sec)		

또한 <표 3>, <표 4>에서는 모의데이터 2, 모의데이터 3에 대한 실험결과를 제시하였는데, 모의데이터 1에 대한 실험결과와 비슷한 결론을 얻을 수 있었다. 대부분 결과가 초기해와 상관없이 데이터의 원 구조를 잘 나타냈고, 초기군집수의 증가가 알고리듬의 실행에 별 다른 영향이 없는 것으로 나타났다.

〈표 3〉 모의 데이터 2에 대한 최적결과

군집번호	개체의수	군집별 개체번호
1	8	1~8
2	8	9~16
3	8	17~24
4	8	25~32
5	8	33~40
기준함수 값 : 0.124207 평균시간 : 12(sec)		

〈표 4〉 모의 데이터 3에 대한 최적결과

군집 번호	개체의 수	군집별 개체번호
1	15	1~15
2	15	16~30
기준함수 값 : 0.128447 평균 소요시간 : 11(sec)		

모의 데이터 4와 5에 대한 실험에서 사전에 군집수가 정해지지 않은 상태에서 ASA를 통한 최적의 결과가 각기 <표 5>, <표 6>과 같이 얻어졌다. 모의 데이터 4의 경우 원래의 3개 그룹보다 더 세분화된 것을 볼 수 있는데, 이는 원래 데이터가 겹치는 영역을 가지고 있기 때문에 사전에 데이터 구조에 대한 지식을 전혀 가지고 있지 않는 상황에서는 오히려 타당한 결과라고 생각된다. 더욱 고무적

〈표 5〉 모의 데이터 4에 대한 최적결과

군집 별→ 방법 별 ↓	군집 1	군집 2	군집 3	군집 4	군집 5	군집 6	기준함수 값↓
최단거리방법	1~20	21~40	41~60				28874.80
최장거리방법	(1, 3, 7, 9, 10, 12, 14, 15, 17, 20)	(2, 4, 5, 6, 8, 11, 13, 16, 18, 19)	(21~24, 28, 29, 31, 34~37, 39)	(25~27, 30, 32, 33, 38, 40)	(41~43, 45, 52, 53, 55)	(44, 46~51, 54, 56~60)	19357.57
ASA 군집화	(1, 3, 7, 9, 10, 12, 14, 15, 17, 20)	(2, 4, 5, 6, 8, 11, 13, 16, 18, 19)	(21~24, 28, 29, 31, 34~37, 39)	(25~27, 30, 32, 33, 38, 40)	(41~43, 45, 52, 53, 55)	(44, 46~51, 54, 56~60)	19357.57

〈표 6〉 모의 데이터 5에 대한 최적결과

군집 별→ 방법 별 ↓	군집 1	군집 2	군집 3	군집 4	기준함수 값↓
최단거리방법	1~30	31~50	51~60		21486.16
최장거리방법	(1, 5, 6, 9, 12, 15, 17, 21, 22, 28, 30)	(2~4, 7, 8, 10, 11, 13, 14, 16, 18~20, 23~27, 29)	31~50	51~60	18739.84
ASA 군집화	(1, 5, 6, 9, 11~13, 15, 17, 18, 21, 22, 24, 28, 30)	(2~4, 7, 8, 10, 14, 16, 19, 20, 23, 25~27, 29)	31~50	51~60	18254.98

인 것은 목적함수를 최소화하여 최종적으로 얻은 군집들이 비록 세 개의 군집으로 주지는 않았지만 원래 주어진 구조를 뒤섞은 것이 아니라 원래의 군집내부에서 더 세분화한 것이라는 점이다. 모의데이터 5의 경우 표준편차를 좀 줄였기 때문에 서로 다른 군집사이의 데이터들이 비교적 잘 구분된 상태이므로 ASA의 결과도 데이터의 원 구조와 좀더 근접하게 나타난 것으로 보인다. 초기해에서 얻은 군집수와 상관없이 대부분 실험에서 유사한 결과를 보여 주었다.

5. 실제 데이터에 대한 적용 예

5.1 붓꽃 표본의 분류

Fisher(1936)의 고전적인 붓꽃 분류 데이터는 3종류 붓꽃(iris setosa, iris versicolor, iris virginica)의 총 150개 표본으로 구성되었는데, 매 개체는 4개의 생태학 변수(꽃받침과 꽃잎의 길이와 넓이)의 특징을 가졌고, 또 각 종류는 50개 개체표본으로 구성되었다. 기존의 연구결과로부터 두 번째와 세 번째 종류중 일부분 개체들이 주어진 4개의 특성변수에 의해 구분이 잘 안 된다는 사실이 알려져 있다. 여기서는 원 데이터 중에서 잘 구분되지 않았던 일부 개체들을 포함시킨 20개 개체(총 60개 개체)만을 추출하여 군집화를 예시한다. <표 7>의 데이터에서 왼쪽 위에서부터 개체 번호를 부여하여 군집 1은 1~20, 군집 2는 21~40, 군집 3은 41~60에 해당하고 군집 2에 있는 24, 29, 40번 개체들과 군집 3에 있는 44, 52, 59번 개체들이 잘 구분되지 않는 예가 된다.

사전에 군집수를 정해주지 않은 상태에서 세 가지 방법으로 군집화를 실시한 결과를 <표 8>에서

〈표 7〉 붓꽃 표본데이터(Fisher 1936)에서 추출한 부분 데이터

종류 별 →	첫 번째 종류				두 번째 종류				세 번째 종류			
속성변수→ 개체번호↓	v1	v2	v3	v4	v1	v2	v3	v4	v1	v2	v3	v4
1	5.1	3.5	1.4	0.3	6.3	2.3	4.4	1.3	5.9	3.0	5.1	1.8
2	4.4	3.2	1.3	0.2	6.1	3.0	4.6	1.4	6.3	3.4	5.6	2.4
3	4.4	3.0	1.3	0.2	5.9	3.0	4.2	1.5	5.8	2.7	5.1	1.9
4	5.0	3.5	1.6	0.6	6.0	2.7	5.1	1.6	6.3	2.7	4.9	1.8
5	5.1	3.8	1.6	0.2	5.6	2.5	3.9	1.1	6.0	3.0	4.8	1.8
6	4.9	3.1	1.5	0.2	6.7	3.1	4.7	1.5	7.2	3.2	6.0	1.8
7	5.0	3.2	1.2	0.2	6.2	2.2	4.5	1.5	6.2	2.8	4.8	1.8
8	4.6	3.2	1.4	0.2	5.9	3.2	4.8	1.8	6.9	3.1	5.4	2.1
9	5.0	3.3	1.4	0.2	6.3	2.5	4.9	1.5	6.7	3.1	5.6	2.4
10	4.8	3.4	1.9	0.2	6.0	2.9	4.5	1.5	6.4	3.1	5.5	1.8
11	4.8	3.0	1.4	0.1	5.6	2.7	4.2	1.3	5.8	2.7	5.1	1.9
12	5.0	3.5	1.3	0.3	6.2	2.9	4.3	1.3	6.1	3.0	4.9	1.8
13	5.1	3.3	1.7	0.5	6.0	3.4	4.5	1.6	6.0	2.2	5.0	1.5
14	5.0	3.4	1.5	0.2	6.5	2.8	4.6	1.5	6.4	3.2	5.3	2.3
15	5.1	3.8	1.9	0.4	5.7	2.8	4.5	1.3	5.8	2.8	5.1	2.4
16	4.9	3.0	1.4	0.2	6.1	2.9	4.7	1.4	6.9	3.2	5.7	2.3
17	5.3	3.7	1.5	0.2	5.5	2.5	4.0	1.3	6.7	3.0	5.2	2.3
18	4.3	3.0	1.1	0.1	5.5	2.6	4.4	1.2	7.7	2.6	6.9	2.3
19	5.5	3.5	1.3	0.2	5.4	3.0	4.5	1.5	6.3	2.8	5.1	1.5
20	4.5	3.4	1.6	0.2	6.3	3.3	4.7	1.6	6.5	3.0	5.2	2.0

주) v1 = 꽃받침 길이, v2 = 꽃받침 넓이, v3 = 꽃잎 길이, v4 = 꽃잎 넓이.

제시하였다. 여기서 두 가지 계층적 군집화 방법에 의한 결과는 군집수를 강금석, 윤복식(2003)에서 제시된 방법에 따라 구한 최우수 결과들을 선택하여 제시하였다. ASA 군집화 방법에서는 초기 군

〈표 8〉 실제데이터 1에 대한 실험결과 2

군집 별→ 군집화 방법 ↓	군집 1	군집 2	군집 3	군집 4	군집 5	기준함수 값
최단거리방법	1~20	21~57, 59, 60	58			7752.86
최장거리방법	1~20	21, 25, 27, 31, 35, 37~39, 53	22~24, 26, 28~30, 32~34, 36, 40, 41, 43~45, 47, 51, 52, 55, 59	42, 46, 48~50, 54, 56, 57, 60	58	3726.82
ASA 군집화 방법	1~20	33, 42, 45, 46, 48~50, 54, 56~58, 60	나머지 개체들			3656.39 11.0(sec)

〈표 9〉 유럽 26개국 9개 산업별 종사자 백분비를 표준화한 데이터

번호	국가별 ↓/속성 변수→	AGR	MIN	MAN	PS	CON	SER	FIN	SPS	TC
1	Belgium	-1.02	-0.36	0.09	-0.02	0.02	1.34	0.78	0.96	0.47
2	Denmark	-0.64	-1.19	-0.74	-0.81	0.08	0.36	0.89	1.78	0.40
3	France	-0.54	-0.47	0.07	-0.02	0.45	0.84	0.71	0.38	-0.61
4	West Germany	-0.80	0.05	1.25	-0.02	-0.53	0.32	0.36	0.33	-0.32
5	Ireland	0.26	-0.26	-0.90	1.04	-0.40	0.84	-0.43	0.11	-0.32
6	Italy	-0.21	-0.67	0.08	-1.08	1.12	1.12	-0.85	0.01	-0.61
7	Luxembourg	-0.73	1.90	0.54	-0.29	0.63	1.21	0.21	-0.12	-0.25
8	Netherlands	-0.82	-1.19	-0.64	0.24	1.05	1.10	1.00	1.24	0.18
9	UK	-1.06	0.15	0.46	1.31	-0.77	0.86	0.61	1.21	-0.10
10	Austria	-0.41	-0.16	0.46	1.31	0.51	0.84	0.32	-0.47	0.33
11	Finland	-0.39	-0.88	-0.16	1.04	-0.46	0.38	0.54	0.63	0.76
12	Greece	1.43	-0.67	-1.34	-0.82	-0.04	-0.32	-0.57	-1.32	0.11
13	Norway	-0.65	-0.78	-0.66	-0.29	0.26	0.86	0.25	1.11	2.05
14	Portugal	0.56	-0.98	-0.36	-0.82	0.14	0.07	-0.46	-0.49	-0.61
15	Spain	0.24	-0.47	0.21	-0.55	2.03	-0.71	1.60	-1.20	-0.75
16	Sweden	-0.84	-0.88	-0.16	-0.29	-0.59	0.31	0.71	1.81	0.18
17	Switzerland	-0.73	-1.09	1.54	-0.29	0.81	0.99	0.46	-0.68	-0.61
18	Turkey	3.07	-0.57	-2.73	-2.15	-3.26	-1.70	-1.03	-1.19	-2.40
19	Bulgaria	0.29	0.67	0.76	-0.82	-0.16	-1.08	-1.18	-0.27	0.11
20	Czechosolvakia	0.17	1.70	1.21	0.78	0.32	-0.82	-1.11	-0.31	0.33
21	East Germany	-0.96	1.70	2.03	1.04	-0.34	-0.38	-1.00	0.30	1.33
22	Hungary	0.16	1.90	0.37	2.64	0.02	-0.78	-1.11	-0.41	1.04
23	Poland	0.77	1.28	-0.19	-0.02	0.14	-1.19	-1.11	-0.57	0.25
24	Romania	1.00	0.87	0.44	-0.82	0.32	-1.54	-0.96	-1.22	-1.11
25	USSR	0.29	0.15	-0.17	-0.82	0.63	-1.50	-1.25	0.52	1.98
26	Yugoslavia	1.90	0.25	-1.46	0.51	-1.98	-1.43	2.60	-2.16	-1.83

주) AGR = agriculture, MIN=mining, MAN=manufacturing, PS=power supplies, CON=construction, SER=service industries, FIN=finance, SPS=social and personal service, TC=transport and cmmunication

집수를 달리하면서($c=2,\cdots,7$) 군집화를 수행해 보았는데 50번 실험에서 대부분 군집수를 3으로 주었고, 약 90%의 결과가 기준함수의 값을 3700 이하로 감소 시켰다(평균 소요시간 11초). 또한 초기 군집수 각기 3과 4로 주었을 때 최적결과에 수렴하는 속도도 빨라진다는 것을 관찰할 수 있었다. 따라서 큰 데이터에 대한 분석에서는 초기분할을 되도록 최적에 가깝게 선택하는 것이 알고리듬의 수렴속도를 높이는데 효과적이라고 볼 수 있다. 여기서 군집화 기준은 내부 편차합 기준에 군집수 증가에 따르는 벌금을 반영한 변형기준인 식 (1)을 $\alpha=2.5$으로 놓고 사용하였다. <표 8>의 ASA 군집화 결과는 가장 적은 목적함수를 준 것을 나타내었다.

결국 군집수를 사전에 정해주지 않았을 때도 초기해의 선택과 거의 상관없이 적절한 군집수의 분할결과를 주게 되므로 군집화 기준만 당면문제에 알맞게 설정해 주면 ASA 군집화 방법은 매우 효과적인 기법이 될 것이다. 이 예의 경우 기존 연구에서 보여준 결론과 비슷하게 본 실험에서도 두 번째와 세 번째 군집이 주어진 속성변수에 의해 잘 구분되지 않는 것으로 나타났다.

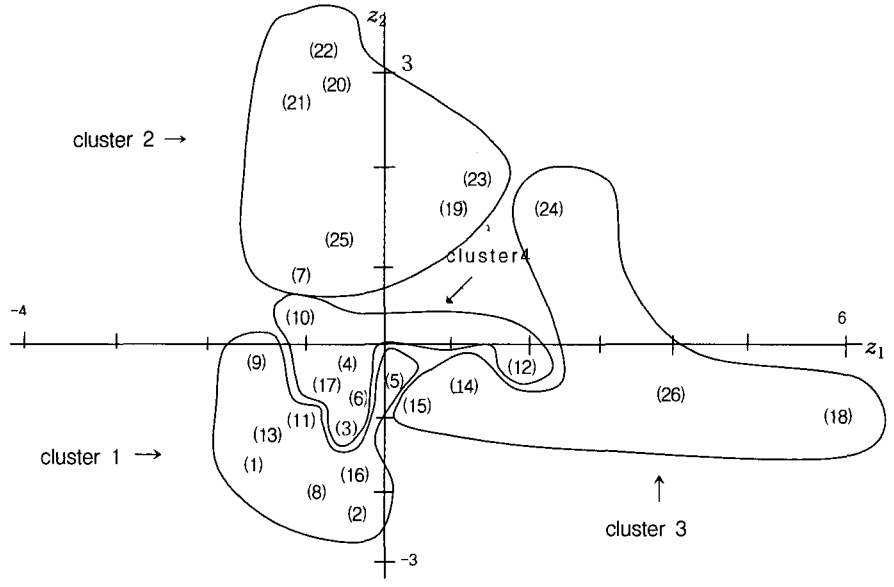
5.2. 유럽 국가의 분류

<표 9>의 데이터는 유럽 26개국을 대상으로 9개 특정 산업분야에 종사하는 노동자의 백분비를 표준화를 시켜 얻은 데이터이다(즉 평균은 0이고 분산은 1이다)(Manly(1994)). 여기서의 분할 목적

〈표 10〉 유럽 국가 데이터 대한 실험결과 1

◈ 군집수 4일 때

군집 별 → 방법 별 ↓	군집 1	군집 2	군집 3	군집 4	기준함수 값↓
최단거리방법	(1~14, 16, 17)	(18, 26)	(15)	(19~25)	4034.70
최장거리방법	(1~11, 13~17)	(12, 19, 23~25)	(18, 26)	(20, 21, 22)	3057.66
ASA 방법	(1, 2, 5, 8, 9, 11, 13, 16)	(7, 19-23, 25)	(14, 15,18, 24, 26)	(3, 4, 6, 10, 12, 17)	1751.40



[그림 6] 두 주성분에 의한 유럽 26개국 분산도 1

은 유럽 26개 국가들을 주어진 산업별 종사자의 백분비를 속성변수로 하여 비슷한 양상을 띤 국가 그룹으로 분류하는 것이다.

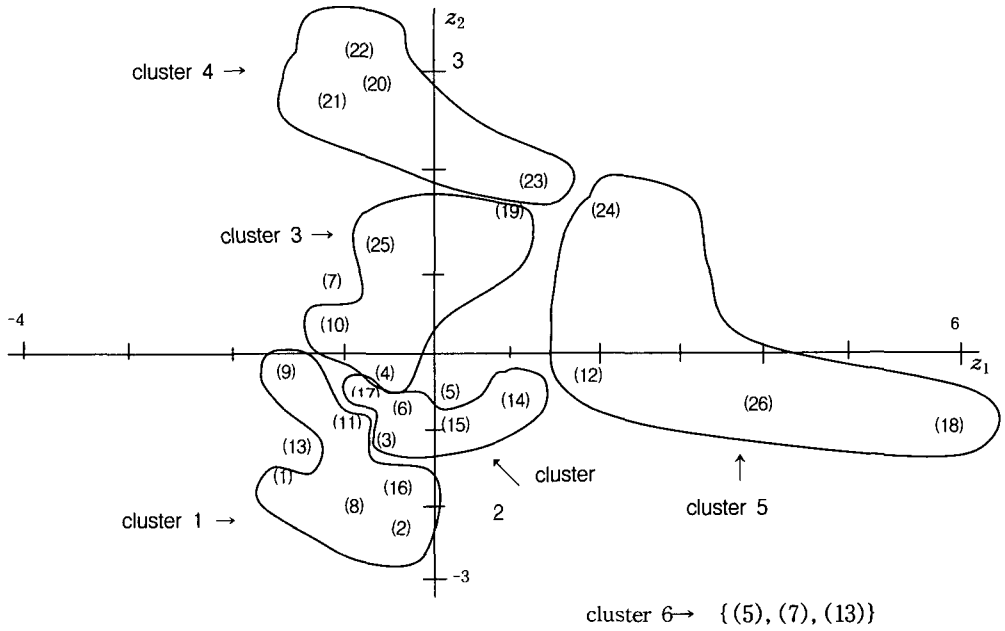
<표 10>과 <표 11>에서는 사전에 군집수를 정해주지 않은 상태에서의 실험결과를 볼 수 있다. 각각 다른 군집수의 계층분할을 초기해로 설정하여 수행한 20번의 ASA 군집화 결과 중에서 최적의 결과를 선택하여 제시했다. 우선 최대 군집수를 5로 제한하였을 때 최적의 결과는 4개 군집으로 나타났다(<표 10>). [그림 6]과 [그림 7]은 주성분 분석을 통해 두 주성분 z_1, z_2(데이터의 변동을 약 61%를 반영한다)에 의해 분류된 26개 유럽국가별 분산도를 나타난다. 비록 전체 요인들을 전부 고려한 것이 아니기에 완전히 객관적이지는 못할 수도 있지만 데이터의 특성을 대부분 반영한 결과이므로 비교를 위해 제시하였다(Manly, 1994)). 그리고 실선으로 분리된 각 구역은 본 실험에서의 분할결과를 나타낸 것이다. 실험에서 4개 군집으로 분할했을 때 몇몇 국가를 제외하고는 공통적인 경향을 볼 수 있는데, 같은 군집에 배정된 국가들이 상대

〈표 11〉 유럽 국가 데이터에 대한 실험결과 2

◈ 군집수 6일 때

군집별 →	군집 1	군집 2	군집 3	군집 4	군집 5	군집 6
군집별 국가명칭 →	(1) Belgium	(3) France	(4) W.Germany	(20) Czechos.	(12) Greece	(5) Ireland
	(2) Denmark	(6) Italy	(10) Austria	(21) E.Germany	(18) Turkey	(7) Luxemb.
	(8) Netherlands	(14) Portugal	(19) Bulgaria	(22) Hungary	(24) Romania	(13) Norway
	(9) UK	(15) Spain	(25) USSR	(23) Poland	(26)Yugoslavia	
	(11) Finland	(17) Switzerland				
	(16) Sweden					

기준함수 값 : 1643.24



[그림 7] 두 주성분에 의한 유럽 26개국 분산도 2

적으로 근접한 두 주성분 값을 갖고 있음을 관찰할 수 있다.

그리고 <표 11>에서는 군집수의 상한을 8로 주었을 때의 결과를 보여주었다. 최적의 군집화는 군집수가 6일 때의 분할결과로 나타났다. 여기서도 3개 국가((5), (7), (13)번 개체)를 제외하고는 거의 비슷한 양상을 띠고 있음을 볼 수 있다.

본절의 실험에서도 초기 군집수의 선택과 거의 상관없이, 기준함수의 값을 줄여 가는 과정에서 군집수를 증가 또는 감소시켜 가면서 최적 근처의 결과에 수렴하는 것을 관찰할 수 있었다. 각기 다른 20번 실험에서 기준함수의 평균값은 1870.22이였고, 평균소요시간은 9.8초였다. 초기군집화가 최적에 가까우면 더 빨리 "좋은" 결과에 접근하는 것을 관찰할 수 있다.

6. 결론 및 토의

본 논문에서는 변형된 가상어닐링의 일종인 ASA 알고리듬을 군집화 분석에 적용하여 사전정보가 불충분한 일반적인 데이터 집합의 분할에 적용할 수 있는 군집화 방법을 소개하고 몇 가지 전형적인 데이터 및 실제 데이터에 대해 적용 실험을 수행하여 타당성을 보였다. 이 방법은 다음과 같은 두 가지 측면에서 매우 효과적이다. 첫째로, 기존의 다른 방법으로(예를 들면 계층적 방법) 얻은 결과들을 개선할 수 있게 해준다. 둘째로, 군집수에 대한 사전정보가 충분하지 못한 경우에도 초기의 군집수와 상관없이 기준함수를 최적화하는 과정을 통해 적절한 군집수와 최적에 가까운 군집화 결과를 제공해준다. 이것은 4, 5장의 실험에서 초기에 군집수를 어떻게 주든 상관없이 거의 동일한 결과를 주는 것으로 확인할 수 있었다. 물론 현실의 문제에서는 대개 군집수에 적절한 상한을 줄 수 있으므로 본 방법을 더욱 효과적으로 적용할 수 있을 것으로 본다. 단지 데이터의 특성에 따라 적절한 거리 및 군집기준함수를 설정하는 것이 중요한데 향후 본 논문에서 제시한 군집화 방법을 다양한 형태의 정량적, 정성적인 데이터의 군집화에 적용하는 연구가 계속될 것이다. 또한 군집화 결과의 타당성을 검정하는 통계적인 기법에 대한 연구가 필요할 것이다.

참 고 문 헌

[1] 강금석, 윤복식, 적절한 군집수 결정 방법(준비중), 2003.

[2] 김여근, 윤복식, 이상복, 「메타 휴리스틱」, 영지문화사, 1997.

[3] 윤복식, 조계연, "Simulated Annealing의 가속화와 ATM 망에서의 가상경로 설정", 「한국경영과학회지」, 제20권, 제2호(1996), pp.125-140.

[4] Duda, R.O, P.E. Hart and D.G. Hart, *Pattern Classification*, Wiley, New York, 2001.

[5] Everitt, B.S., *Cluster analysis,* Halsted Press, London, 1980.

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

[7] Fisher, R.A. "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, Vol.2(1936), pp.179-188.

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

[9] Kirkpatrick, S., C.D. Gelatt and M.P. Vecchi, "Optimization by simulated annealing," *Science*, Vol.220(1983), pp.671-680.

[10] Kovesi, B., Boucher, J.M. and Saoudi, S.c, "Stochastic K-means algorithm for vector quantization," *Pattern Recognition Letters*, Vol.22(2001), pp.603-610.

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

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

1996

[13] Sanghamitra, B., S.K Pal and C.A. Murthy, "Simulated annealing based pattern classification," *Journal of Information Sciences,* Vol.109(1998), pp.165-184.

[14] Sanghamitra, B., M. Ujjwal and K.P. Malay, "Clustering using simulated annealing with probabilistic redistribution," *International Journal of Pattern Recognition and Artificial Intelligence*, Vol.15, No.2(2001), pp.269-285.

[15] Selim, S.Z. and M.A. Ismail, "K-mean type algorithms : a generalized convergence theorem and characterization of local optimality," *IEEE Tans. Patt. Anal. Mach. Intell.,* Vol.6(1984), pp.81-87.