IE Interfaces
Vol. 15, No. 2, pp. 182-188, June 2002.

정량적 자료에 대한 효과적인 군집화 과정 및 사용 후 핵연료의 분류에의 적용

강금석[1] · 윤복식[1†] · 이용주[2]

[1]홍익대학교 기초과학과 응용수학 전공 / [2]이화여대 경영학과

An Effective Clustering Procedure for Quantitative Data and Its Application for the Grouping of the Reusable Nuclear Fuel

Jin-Xi Jiang[1] · Bok-Sik Yoon[1] · Yong-Joo Lee[2]

[1] Department of Science(Applied Math. Group), Hongik University, Seoul, 121-791

[2] School of Business Administration, Ewha Women's University, Seoul, 120-750

Clustering is widely used in various fields in order to investigate structural characteristics of the given data. One of the main tasks of clustering is to partition a set of objects into homogeneous groups for the purpose of data reduction. In this paper a simple but computationally efficient clustering procedure is devised and some statistical techniques to validate its clustered results are discussed. In the given procedure, the proper number of clusters and the clustered groups can be determined simultaneously. The whole procedure is applied to a practical clustering problem for the classification of reusable fuels in nuclear power plants.

Keywords: clustering, hierarchical clustering, validation, reusable fuel classification

1. 서 론

특성이 알려져 있지 않은 자료들에 대한 군집화(clustering)는 고전적인 적용 대상인 사회심리학, 생태학, 농학, 의학 등의 분야에서 뿐만 아니라 경영학, 공학 등 거의 모든 분야에서 자료 분석의 출발점이 되고 있다. 군집화는 주로 통계적인 방법에 의해 이루어지므로 통계적 군집화(statistical clustering)로 부르기도 하는데 군집화의 중심문제는 어떻게 하면 주어진 개체들을 적절하게 그룹별로 나누어 같은 그룹의 개체들은 동질성을 갖게 하고, 서로 다른 그룹 간에는 이질성이 표출되도록 하는가하는 문제이다. 군집화가 잘 이루어지면 데이터를 그룹별로 묶어 특성을 파악하고 분석할 수 있게 되므로 데이터 개체의 수를 줄일 수 있어 분석이 용이해질 수 있다.

군집화 방법은 크게 계층적 방법(hierarchical approach)과 최적화(optimization)를 통한 방법으로 나눌 수 있는데, 그 중 계층적 방법은 방법 자체가 이해하기 쉽고 계산이 용이한 장점을 가지고 있어 군집화 과정에서 많이 적용되고 있다. 그러나 계층적 군집화 과정에서 일단 한 그룹에 배정된 개체는 다른 그룹으로의 재배치가 불가능하므로 최적의 군집화가 이루어지지 않을 가능성이 높다. 반면에 최적화 방법은 최적 또는 최적에 가까운 군집화가 가능하나 개체의 수가 많아짐에 따라 계산량이 크게 증가하는 단점이 있다. 또한 특성이 잘 알려지지 않은 데이터집합에 대해 군집화의 적절성을 잘 반영하는 목적함수를 설정하는 것 자체가 쉽지 않은 문제점이 있다. 본 논문에서는 전체 그룹수가 결정되어 있지 않은 정량적인 자료인 경우에 대해 그룹수 결정 및 군집화를 합리적으로 수행하기 위해 효과적으로 사용할 수 있는 실용적인 방법을 만들고 이것을 원자력 발전소에서 제기되고 있는 사용 후 핵연료의 분류 문제에 적용하여 유효한 분류를 얻는 것을 목적으로 한다. 본 연

† 연락저자 : 윤복식 교수, 121-791 서울시 마포구 상수동 72-1 홍익대학교 기초과학과, Fax : 02-336-8130, e-mail : bsyoon@hongik.ac.kr
2002년 1월 접수, 2회 수정 후 2002년 3월 게재 확정.

구에서의 군집화 방법은 계층적인 방법을 기본으로 하되 적절한 그룹수를 결정하기 위해 최적화 방법에서의 목적함수를 도입한다. 또한 최종적인 군집화의 타당성을 검정하기 위해 통계적인 기법을 도입하여 적용한다.

본 서론에 이어 2절에서 기존의 군집화 방법의 특징 및 용도에 대해 소개하고 3절에서 군집화에서 선결되어야 할 거리, 목적함수 등의 문제를 약술하고 4절에서 본 연구에서 따르고 있는 계층적 방법에 최적화의 개념을 혼합한 군집화 방법 및 결과에 대한 타당성을 검증하는 통계적 분석 방법을 설명하고, 5절에서 사용 후 핵연료에의 적용 예를 제시한 후 6절에서 결론 및 향후 연구 과제를 논한다.

2. 군집화 방법의 개요

2.1 군집화의 특징 및 용도

군집화는 최근에 다변량 통계학에서 주된 응용 기법으로 주목받게 되었는데(Everitt, 1991), 이는 데이터의 동질 그룹들을 결정함으로써 복잡한 구조의 1차 데이터로부터 보다 간단한 구조와 형태의 데이터를 얻을 수 있어 많은 실제적 응용 분야를 가지고 있기 때문이다(Krzanowskil, 1995). 몇 가지 예를 들면 미국 어떤 도시의 236명의 자살시도 자들을 14개의 사회심리학적 변수로부터 3개 그룹으로 분류하여 좋은 결과를 얻은 적이 있고, 농업에서는 농작물의 수확량 추정, 병충해 방지 등을 공중 촬영 자료의 군집화를 통해 직접적인 지면 조사보다 훨씬 경제적으로 또한 정확하게 수행한 보고가 있다(Hand, 1981). 또한 많은 지역을 대상으로 새 상품의 판매 정보를 수집분석할 때, 지역들을 적절한 기준에 따라 몇 개의 동질 그룹으로 분류하여 각 그룹의 적절한 대표지역만 선택하여 조사하면 조사대상의 수를 줄일 수 있어 시장 상황을 파악하는 데 큰 도움이 될 수 있다. 이밖에도 의학 연구에서의 조기 진단과 환자의 분류 등 많은 실제 문제에서 군집화 방법을 적용할 수 있다.

2.2 군집화 방법의 종류

군집화 방법은 크게 계층적 방법(hierarchical approach)과 최적화 방법(optimization approach)으로 구분할 수 있다.

2.2.1 계층적 방법

이 방법은 가장 오랜 역사를 가지고 있는 방법이고 또한 보편적으로 사용되고 있는 기법이다. 계층적 군집화 방법은 결합적 방법과 분리적 방법으로 구분할 수 있는데 결합적 방법은 처음에 매 개체를 각기 1개 그룹으로 간주하고 적절한 기준(3절에서 설명)에 의해 매 단계마다 가장 밀접한 두 그룹을 결합해 가는 방법이고, 분리적 방법은 이와 반대로 처음에 전체 개체 모두를 하나의 그룹으로 간주하고 적절한 기준에 의해 매 단계마다 가장 구별이 되는 부분을 떼어내는 방식으로 군집화를 수행하는 방법이다. 일반적으로 결합적 방법이 많이 사용된다.

2.2.2 최적화 방법

최적화 방법은 계층적 접근 방법과 달리 군집화 과정에서 그룹 간의 개체 이동이 가능한 방법이다. 일반적으로 군집화 결과의 최적화를 목적으로 사전에 군집화 목적함수(clustering criterion function)를 설정하여 군집화 과정에서 이 목적함수를 최적화하도록 그룹 간 개체들을 이동시킨다. 이 방법은 개체수가 많아지면 가능한 군집화 경우 수가 너무 많아져서 최적화를 수행하기가 어려워지는 NP-hard 문제가 된다. 따라서 지역탐색법(local search)과 같은 근사적인 최적화 기법을 많이 사용하는데, 대표적인 예로 k-means 방법을 들 수 있는데 k-means 방법을 간략히 소개하면 다음과 같다.

그룹의 수가 k개로 설정되었다고 가정하자. 우선 적절하게 군집화를 하여 k그룹으로 나눈 후, 각 그룹의 평균을 계산한다. 그리고 개체를 검정하여 만약 한 개체와 다른 그룹의 평균 간의 거리가 자신이 포함되어 있는 그룹의 평균 간의 거리보다 작을 때는 그 개체를 가까운 그룹으로 이동시키고 반대 상황이면 계속 그 그룹에 머물게 한다. 다음 단계에서 다시 그룹평균을 계산하고 위의 과정을 반복하는 과정을 안정 상태에 접근할 때까지 계속하는 방법이 k-means 방법이다. 주의할 점은 이 방법은 지역최적해(local optimum)에 머물 수 있고 이 방법을 적용하기 위해서는 사전에 그룹수 k가 결정되어 있어야 하는 것이다.

k-means 방법이 비교적 간편한 근사적 최적화 기법이지만 그룹수(k)가 바뀌면 전체적인 과정을 처음부터 다시 시작해야 한다. 이에 따라 사전에 그룹수에 대한 정보가 없을 때에는 그룹수 결정을 위해 여러번 k-means 과정을 반복해야 하므로 복잡성 더해지는 단점이 있다. 그러나 계층적 방법은 k 이전의 해의 구조가 그대로 유지되면서 다음 과정($k-1$)으로 넘어가기 때문에 그룹수 변동을 포함한 전체 과정이 훨씬 간편하게 수행될 수 있다. 본 연구에서는 전체적으로 계층적 방법을 따르되 최적화의 목적함수를 반영하여 별다른 계산량의 증가없이 적절한 군집수를 결정할 것이다.

3. 군집화의 선결문제

3.1. 적절한 변수의 선택

변수 선택에 따라 군집화 결과가 달라지게 되므로 각 그룹에 포함된 여러 개의 변수들 중에서 당면 목적에 맞는 변수들을 적절히 선택해야 한다. 원시 데이터의 특성을 되도록 잘 파악하기 위해 분석 시작시에는 많은 변수를 고려하고 분석이 진행됨에 따라 주성분분석 방법 등 다변량 통계 분석을 이용하여 변수의 개수를 줄이는 방법을 택할 수 있다. 이 경우 원래

의 데이터의 구조를 파괴할 우려가 있으므로 주의를 요한다. 각 변수의 영향을 동일하게 고려하기 위해 정규화 기법도 군집화 과정에 흔히 사용된다.

3.2 개체 간 거리의 설정

군집화를 위해서는 개체 간의 유사성을 반영하는 척도인 거리를 설정해야 한다. 정량적인 데이터의 경우 포함된 개체들은 m차원 공간상의 벡터로 볼 수 있는데 일반적으로 많이 쓰이는 거리척도로 Minkowski 거리가 있다. x, y를 m개 변수를 갖는 두 개체라고 할 때(즉 m차원 벡터)

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

를 Minkowski 거리(즉 L_p-norm)라고 하는데, 특히 $p=1$일 때는 Manhattan 거리, $p=2$일 때는 Euclidean 거리를 얻을 수 있다. 그 외에도 공분산을 고려한 Mahalanobis 거리, 상관계수를 이용한 거리 등 여러 가지 거리들이 용도에 따라 쓰이고 있는데 본 연구에서는 Euclidean 거리를 사용한다.

3.3 그룹 간 거리의 설정

계층적 군집화 과정에서는 현 단계에서 c개의 그룹이 얻어져 있다면 다음 단계에서는 가장 유사성이 높은 두 개의 그룹을 묶어, $c-1$개의 그룹으로 만들어야 한다. 이때 그룹 간의 유사성을 반영하는 척도를 어떻게 정하는가에 따라 결합되는 그룹들이 달라져 군집화 결과가 달라진다. 그룹 간 거리의 선택은 군집화 결과와 직접적인 관계가 있는 매우 중요한 과정이다. 많이 사용되고 있는 그룹 간 거리는 다음과 같다.

(1) 최단거리(nearest neighbor(single link)) : 두 그룹 간의 거리를 두 그룹 간 개체들 사이의 거리 중 제일 작은 것으로 설정한다. 이 방법은 아주 멀리 떨어진 개체들이 하나의 그룹으로 연결되어 가는 chaining 현상이 일어날 수 있는 단점이 있다.

(2) 최장거리(farthest neighbor(complete link)) : 두 그룹 간의 거리를 두 그룹 간 개체들 사이의 거리 중 제일 큰 것으로 설정한다.

(3) 중심거리(centroid distant) : 두 그룹 간의 거리를 두 그룹의 중심(centroids)사이의 거리로 설정한다. 이 방법은 작은 규모의 그룹과 큰 규모의 그룹이 합할 때 결합 후 중심이 큰 그룹 쪽에 편중되어 작은 그룹이 군집화 과정에서 자체의 특성을 상실하게 되는 단점이 있다.

(4) 중간값(median) : 두 그룹 간의 거리를 두 그룹의 중간값 사이의 거리로 설정한다.

(5) 그룹평균(group average) : 두 그룹 간의 거리를 두 그룹 개체들 사이의 거리들의 평균으로 설정한다.

(6) 편차자승합(sum of squared deviations) : 두 그룹 간의 거리를 개체들과 혼합평균 간 거리의 자승 합에서 전체 개체와 자신을 포함한 그룹평균 간 거리의 자승 합을 뺀 것으로 설정한다.

본 연구에서는 일단 이들 모든 거리를 모두 적용하여 군집화를 시도한 후에 그룹수 결정 과정에서 주어진 자료에 대해 가장 무난한 거리를 택하게 된다.

3.4 최적화 방법에서 목적함수의 결정 문제

군집화 결과에 대한 평가기준은 군집화가 된 후 서로 다른 그룹의 개체들 간의 이질성과 같은 그룹 내의 개체들의 동질성이 어느 정도 합리적으로 반영되었는가 하는 데 있다. 이렇게 서로 다른 그룹 간의 이질성과 같은 그룹 내의 개체들의 동질성을 반영하여 군집화의 타당성 정도를 표현하는 척도인 목적함수의 정의와 사용은 군집화 과정에 아주 중요한 부분이다.

m차원 실공간에서의 n개의 데이터 $X^n=\{x_1, \cdots, x_n\}$에 대해 c개의 그룹으로 군집화가 되었다고 할 때, i번째 그룹은 $X_i=\{x_j \mid x_j \in$그룹 $i\}$이다. 만일 n_i가 그룹 i에 포함된 데이터의 개수이면, $\bar{x}_i = \sum_{x \in X_i} x/n_i$, $\bar{x} = \sum_{x \in X^n} x/n$이 되고, 다음과 같은 $n \times n$ 행렬들이 계산될 수 있다.

$$T = \sum_{x \in X^n} (x-\bar{x})(x-\bar{x})^T$$

$$W = \sum_{i=1}^{c} \sum_{x \in X_i} (x-\bar{x}_i)(x-\bar{x}_i)^T$$

$$B = \sum_{i=1}^{c} n_i (\bar{x}_i-\bar{x})(\bar{x}_i-\bar{x})^T$$

T는 전체 평균에 대한 개체들의 흩어짐의 정도이고, W는 각 그룹 내에서의 흩어짐의 정도이며(동질성), B는 각 그룹의 평균과 전체 평균과의 차이에 각 그룹에 속하는 개체들의 수만큼 가중을 두어 흩어짐의 정도(이질성)를 나타낸 것이다. $T=W+B$임을 알 수 있는데, 좋은 군집화를 한다는 것은 W를 최소화시키거나 B를 최대화시키는 문제로 볼 수 있다. 그러나 이들은 행렬이므로 이들로부터 적절한 1차원 척도를 얻어서 목적함수로 사용할 필요가 있다. 가능한 척도들은 W의 trace(최소화), W의 행렬식(최소화), $W^{-1}B$의 trace(최대화), $T^{-1}W$의 trace(최소화) 등을 들 수 있다(Hand, 1981). W의 trace의 경우 좌표축의 스케일 변화에 따라 값이 달라지므로 주의를 요한다. 일차원 데이터 집합이 분석대상일 때($m=1$)는 W와 B가 스칼라값이 되므로 이들을 직접 목적함수로 사용할 수 있다.

4. 본 연구에서의 군집화 과정

4.1 군집수의 결정 및 군집화

군집화 문제에서 많은 경우에 사전에 그룹수에 관한 정보가

없이 군집화를 시도하게 되는데 보다 합리적인 군집화를 위해서는 그룹수가 적절하게 결정되어야 한다. 최적화의 관점에서는 그룹을 나누는 것과 그룹수를 결정하는 것을 한꺼번에 수행하는 것이 바람직할 것이다. 본 연구에서는 전체적으로 우선 결합 계층적 방법을 이용하여 군집수를 점점 줄여가면서 동시에 최적화 방법의 목적함수 개념을 적용하여 목적함수가 최소에 근접하는 군집수를 결정하는 방식으로 최적에 가까운 군집수와 군집화가 동시에 얻어지는 과정을 제안한다.

계층적 방법의 경우 목적함수의 값은 그룹수에 따라 대체로 단조 증가 또는 단조 감소의 경향을 보이는데, 본 연구에서는 목적함수값의 변화를 관찰하여 그룹수에 따른 목적함수의 그래프로부터 적절한 그룹수를 설정하는 방식을 따른다. 즉, 증가 또는 감소분이 매우 적어지면서 안정상태에 이르는 시점을 파악하여 그룹수를 결정하는 방법이다. 데이터의 분석에서 그룹수에 관한 기술적인 요인에 의한 상한이 주어지면 이 방법은 매우 효과적으로 사용될 수 있지만, 정량화를 시도하기 위해서는 통계적 분석이 필요할 수도 있다.

R. Peck (1989) 등은 손실함수(loss function)를 설정하여 이를 목적함수로 하여 최적화하는 군집화 수행을 전제로 하여 시뮬레이션 방법을 통해 bootstrap 신뢰 구간을 계산하는 방법을 제시하였다. 손실함수는 그룹수와의 관계를 반영하는 부분과 각 그룹의 특성에 그 그룹에 포함된 개체들이 기여하는 정도를 반영하는 부분으로 구성되어 있다. 이 손실함수를 최적화하는 군집화 결과에서 얻어진 그룹수를 적절한 그룹수로 파악하였다.

Chao (1992)는 그룹수를 사전에 모르고 그룹별 확률 분포가 서로 다를 수 있는 모집단에서 하나의 무작위 표본을 뽑았을 때 표본 coverage를 이용하여 그룹수를 추정하는 비모수적(nonparametric) 추정 기법을 제시하였다. 이 기법의 아이디어는 표본 coverage 기대값을 쉽게 추정할 수 있으므로 이들을 적절히 그룹수 추정에 이용하는 것인데 많은 시뮬레이션 실험을 통해 여러 가지 형태의 데이터 집합에 대해 실험을 하여 유용성을 보인 바 있다.

위와 같은 정량적인 분석법은 Peck의 방법같이 계산량이 막대 하든가 아니면 Chao 방법과 같이 그룹수를 결정하는 문제와 군집화 문제를 별개로 취급하는 데 따른 정확성의 결여와 같은 문제점을 보이게 되므로 실제적인 적용에 있어서는 본 연구에서 제시하는 목적함수의 그래프를 이용하는 방법이 보다 유용하리라고 생각된다.

4.2 군집화 결과의 통계적 타당성 검정 방법

군집화 기법의 무비판적인 적용은 경우에 따라 전혀 엉뚱한 그룹들을 만들어 낼 수 있기 때문에 군집화 결과의 검정이 특히 중요하다. 그러나 이를 위한 적절한 통계적 기법은 그리 많이 개발되지 못하고 있는데, 예를 들면 Baker(1975)는 power를 이용하여 계층적 방법에 의한 결과에 대한 분석을 시도하였고, Jolliffe(1995)는 계층적 군집화 분석 방법으로 다변수 데이터집합에 대해 군집화를 하는 과정에서 유별나게 불균형적으로 커다란 영향을 미치는 개체들의 결정 문제에 대해 연구한 바 있다. 본 논문에서는 Gordon (1994)과 유사한 시뮬레이션에 의한 방법을 제시한다. 우선 Gordon의 U-통계량을 계산의 효율성을 높이기 위해 아래와 같이 조정하여 정의한다.

데이터 집합의 개체 i, j 사이의 거리를 d_{ij} 라고 놓고 서로 같은 그룹에 속하는 개체 쌍들의 집합을 W_e, 서로 다른 그룹에 속하는 개체 쌍들의 집합을 B_e 라고 놓았을 때, $(i, j) \in W_e$ 와 $(k, l) \in B_e$ 에 대해

$$U_{ijkl} = \begin{cases} 0 & \text{if } d_{ij} < d_{kl} \\ \frac{1}{2} & \text{if } d_{ij} = d_{kl} \\ 1 & \text{if } d_{ij} > d_{kl} \end{cases}$$

와 같이 값을 부여한 후

$$U = \sum_{(i,j)\in W_e} \sum_{(k,l)\in B_e} U_{ijkl}$$

와 같은 통계량을 얻을 수 있는데, 이것을 U-통계량이라고 한다. U-통계량의 정의로부터 군집화 결과에서 얻어진 U-통계량이 작을수록 군집화 결과가 그룹 내의 동질성과 그룹간의 이질성을 더 잘 반영하였다고 볼 수 있다. 따라서 아래와 같은 시뮬레이션과정을 통해 U-통계량에 의한 통계적 검정을 행할 수 있다.

[통계적 검정 과정]

Step 1. 원래 데이터에 대해 적절한 군집화를 수행하고, U를 계산하여 U_0 로 놓는다.

Step 2. Poisson Model(무작위성) 또는 Unimodal Model에 의해 원래 데이터와 똑같은 개수의 개체를 생성하여 Step 1과 같은 방법으로 군집화를 수행하고 U를 계산한다. 이 과정을 M회 반복하여 얻어진 값을 각각 $U_1, \cdots, U_M$으로 놓는다.

Step 3. $U_1, \cdots, U_M$을 크기가 작은 것부터 배열하여 i번째 작은 것을 $U_{(i)}$로 표시한 후 적절한 j에 대해 $U_{(j-1)} < U_0 \le U_{(j)}$을 만족하면 군집화 결과가 $100(j/M)\%$ 의 수준에서 유의하다고 본다.

이 방법의 문제점은 계산량이 방대하여 큰 규모의 데이터에 대한 분석에는 적합하지 않다는 것인데, 각기 그룹별로 동질성을 검정하는 것도 하나의 방법이 될 수 있지만, U을 수정하여 보다 효율적인 통계량을 만들 수 있다면 시뮬레이션 방법은 타당성 검정에 아주 효과적으로 적용할 수 있을 것으로 보인다.

5. 원자핵연료 재사용을 위한 적용 예

본 절에서는 4절에서 설명된 군집화 및 타당성 검정 과정을 사

용 후 핵연료를 분류하는 문제에 적용한다. 농축우라늄을 사용하는 경수형 원자로에서 사용된 사용 후 핵연료를 천연우라늄을 사용하는 중수형 원자로에서 재사용하는 것은 이론적으로는 가능하나 사용 후 핵연료의 조성의 매우 편차가 커서 세심한 사전 장전 계획이 없이 사용하면 매우 위험하게 된다. 장전 계획에 앞서 우선 사용 후 핵연료를 되도록 핵연료적 특성의 편차가 적은 것끼리 그룹을 만들어 놓을 필요가 있는데, 다양한 핵연료의 특성을 어떻게 고려하여 군집화를 할 것인가 하는 문제가 생긴다.

5.1 사용 후 핵연료의 특성

경 · 중수로 연계핵발전(DUPIC)에서는 경수로에서 사용 된 핵연료를 사용하게 되는데, 그 조성이 경수로 핵연료의 초기 농축도(initial enrichment)와 방출 연소도(discharge burnup)에 따라 달라지게 되어 조성 비균질도(inhomogeneity)가 매우 크다. 핵연료 다발의 조성을 정확히 파악하지 못한다면 핵연료 장전 시 반응도 삽입을 예측할 수 없고, 만약 이것이 중수로의 반응도 제어 장치인 지역 조절 장치(zone controller)의 조절기능을 초과할 경우 원자로의 안전 운전을 보장할 수 없게 된다. 실제로 본 연구의 대상이 되고 있는 2,910개의 경수로 사용 후 핵연료 집합체(PWR spent fuel assembly)의 초기 농축도는 1.6～3.6 w/o, 방출 연소도 또한 7,000～45,000 MWD/T 같이 넓게 분포하고 있어 그 조성이 각기 다르다. 각기 조성이 다른 2,910개의 핵연료 다발에 대하여 일일이 조성 계산을 기반으로 한 격자계산을 수행하여 반응 단면적(reaction cross section)을 생산하고, 이를 이용하여 각 장전모형에 대한 노심 계산을 수행하는 것은 많은 계산 시간을 요하게 되어 사실상 불가능하다. 또한 중수로 노심 계산에 사용되는 RFSP 전산코드도 40가지 이상의 핵연료를 동시에 취급할 수 없으므로, 2,910개의 기초 자료를 고유한 핵적 특성을 유지한 채 40개 이하로 적절히 분류해야 할 필요가 있게 된다.

군집화를 위해서는 우선 적절한 변수를 특성 변수로 선정하여야 하는데, 특성치 중에는 특히 출력에 영향이 큰 초기농축도(initial enrichment)와 방출연소도(discharge burnup)를 동시에 고려하여 군집화를 실행해 볼 수 있는데, 이는 의미있는 군집을 주지 못한다. 이는 이들 두 변수 값을 0에서 1사이로 정규화했을 때 2,910개의 데이터가 분포하는 양태를 아래 <그림 1>과 그것의 등위곡선인 <그림 2>에서 살펴봄으로써 보다 명확히 알 수 있다.

<그림 1>에서 x, y축은 각각 초기농축도와 방출연소도를 나타내고 수직축은 도수를 나타내는데 주어진 핵연료의 데이터가 일부 영역에 편중되어 있음을 볼 수 있다. 또한 <그림 2>의 등위곡선을 통해보면 두 변수가 매우 깊은 연관성이 있다는 것을 알 수 있다. 실제로 이들 두 변수 값을 이용한 2차원 군집 결과는 군집내의 연료들 간의 출력변동이 매우 심하여 결과에 따라 그룹을 만들어 그룹별로 조절하려는 원래 의도와

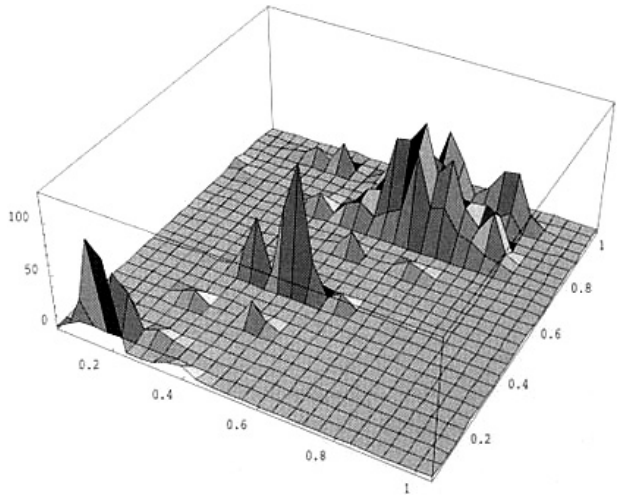

그림 1. 정규화된 초기농축도와 방출연소도의 분포.

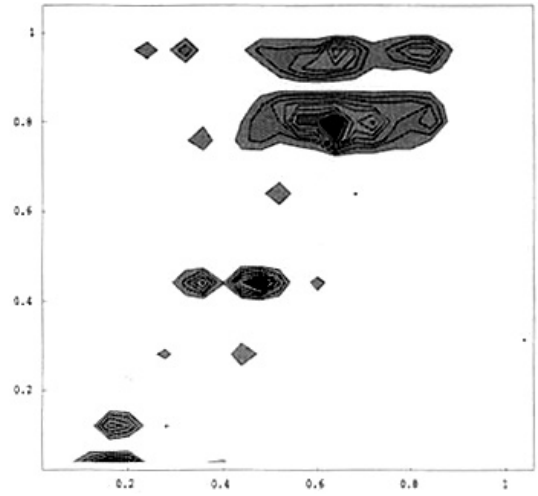

그림 2. <그림 1>의 등위곡선.

무관한 결과를 주는 것을 예비분석을 통해 확인하였다. 따라서 이 두 변수를 기본으로 하여 출력 결과를 근사화할 수 있는 새로운 변수를 생성할 필요가 있게 되었다. 이에 따라 본 분석에서는 앞의 두 변수로부터 핵연료 연소계산 도구인 CASMO를 이용하여 무한증배계수 계산하여 이를 군집화의 기본 변수로 삼는 과정을 밟게 되었다. 전체 데이터의 무한증배계수의 범위는 0.832660～1.383270으로 계산되었다.

5.2 그룹수의 결정 과정

군집화 문제에서는 적절한 그룹수를 결정하는 것이 우선이 된다. 물론 그룹수가 작을수록 이후에 분석이 용이해지므로 그룹수를 되도록 줄일 필요가 있다. 그러나 이것도 군집화 결과의 타당성과 실제 문제에서의 그룹수 제한을 전제로 한다. 이런 목적에서 본 논문에서는 그룹수의 기술적인 제한을 35로 설정하고 이에 따라 동질성과 이질성을 반영하는 목적함수를 설정하였다. 프로그램을 통해 계층적 군집화를 수행하여 그룹

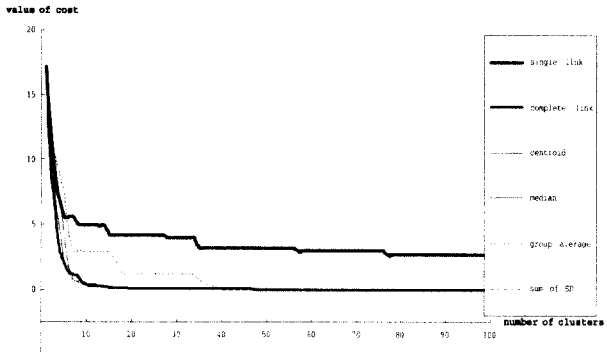

그림 3. 각 방법별 W값의 변화.

수를 하나씩 늘려가면서 2.3절에서 정의된 W의 감소를 관찰하여 그룹수를 선택하였다. <그림 3>에서는 서로 다른 그룹간 거리기준을 적용한 6가지 방법에 대하여 그룹수를 100에서부터 1까지 변화시키며 계산한 W의 변화를 볼 수 있다. 그림에서 보는 것과 같이 그룹수를 증가시키는 과정에서 W는 처음에는 급격히 감소하다가 그룹수가 32개 정도가 되면 거의 모든 방법(최단거리, 그룹평균방법 제외하고)에서 변화가 없으므로 기술적 요구(35 그룹 이하)에 부합되는 32를 적절한 그룹수로 결정하였다. 또한 클러스터 결과는 가장 정규적인 양태를 보이는 최장거리 방법에 의한 것으로 결정하였다. 물론 그룹수 20개 이전에서 급격한 감소로부터 안정상태로 접어들게 되므로 그룹수를 보다 작게 잡을 수 있으나 보다 동질의 핵연료들로 군집화하여 핵연료 장전시의 위험성을 줄이기 위해 여기서는 되도록 크게 잡았다.

5.3 군집화 결과의 타당성 검정

4.2절에서 설명된 U 통계량은 비교적 작은 규모의 표본데이터에 대해 유효한 결과를 얻을 수 있다. 앞에서 소개된 6가지 방법으로 32개 그룹으로 분류된 결과를 보면 대부분 방법이 무한증배계수가 중간 정도인 개체들에 대해서는 군집화 결과에 큰 차이를 보이지 않은 반면, 무한증배계수가 상대적으로 작은 개체들에 대해 다소간의 차이를 보였고 그룹 내에서의 분포도 정규분포 성질을 상실하는 경향이 있었다. 예를 들면 최장거리방법을 이용하여 분류한 결과 무한증배계수 1.00586~1.09409 사이의 6개 그룹에 포함한 개체를 관찰하였을 때 데이터에 대한 뚜렷한 경계가 잘 구분이 되지 않았다. 이런 측면을 고려하여 일부 군집화가 어렵게 느껴지는 구간을 추출하여 Poisson 모형(무작위성)을 가정하여 시뮬레이션을 실시하여 U 통계량을 계산, 비교하였다. <표 1>은 <그림 3>에서 가장 바람직한 결과를 보여주는 최장거리 방법과 가장 좋지 못한 결과를 보이는 최단거리 방법에 대해 이 실험을 100번 수행하여 얻은 결과를 요약해서 보여주고 있다.

<표 1>에서 보는 바와 같이 최장거리 방법으로 얻은 결과는 그 부분 데이터들의 동질성과 이질성을 잘 반영하였다고 본다. 반면 최단거리 방법으로 얻은 결과는 그 자체의 U값은 50% 유의 수준도 초과하는 값을 보이고 있다. 이것은 W 곡선을 분석해 본 결과(<그림 1>)와도 일치하다. 따라서 U통계량은 비교적 작은 규모의 데이터집합에 대한 군집화 결과의 타당성 평가에 있어 유효한 방법이라고 볼 수 있다. 하지만 이 방법은 계산량이 방대하여 데이터집합이 아주 클 때 전체 그룹들에 대한 검정이 어려울 것이다. 따라서 위와 같이 분석하거나 중간부분에서는 정규분포나 혹은 데이터 특성에 따라 다른 모델을 적용하여 일부분씩 분석하는 것이 바람직하다. 또 계산량을 줄일 수 있도록 서로 다른 그룹 간 개체 사이 이질성과 동일 그룹 내의 개체의 동질성을 더 잘 반영하면서도 간편한 다른 통계량을 설정하는 연구가 필요하다.

표 1. U의 유의수준 10%, 5%, 1%의 값과 U_0의 비교

U의 유의수준 값 / 방법 및 데이터 범위	U_1	U_5	U_{10}	U_0
최단거리 방법 (3~12그룹, 129개 데이터) 범위: 0.99255~1.06782	1.1×10^5	1.6×10^5	2.0×10^5	4.7×10^5
최장거리 방법 (3~8그룹, 204개 데이터) 범위: 1.00586~1.09409	19.2×10^5	20.6×10^5	22.5×10^5	15.5×10^5

5.4 군집화 결과 및 대표값 설정

우선, 최단거리와 그룹평균을 이용한 방법은 다른 방법들에 비해 특이한 결과를 나타내고 있어 제외시키고, 다른 기준과 유사한 양태를 보이면서 계산이 비교적 용이한 기준인 최장거리방법을 표준기준으로 설정하고 그룹수를 32개로 정하여 이 기준에 의한 32 그룹으로 된 군집화를 채택하였다. <표 2>는 이 방법을 이용하여 32개 그룹으로 군집화를 하여 얻은 결과이다. 즉 각 그룹별 데이터의 개수, 무한증배계수의 평균 및 표준편차를 계산하여 <표 2>를 작성하였다.

추후 분석을 위해 이미 나누어진 개별 그룹에 대해 군집화의 타당성과 통계적 정규성 등을 분석할 필요가 있는데, 실제로 그룹별 도수분포를 분석한 결과 개체들을 극히 적게 포함한 그룹을 제외하고는 대부분 그룹이 근사하게 정규분포를 따랐다. 이로부터 최장거리방법으로 군집화 하였을 때 대부분 그룹은 그룹평균을 대표값으로 설정할 수 있을 것으로 생각된다. 이제 각 그룹에 속한 연료들의 평균값을 기본으로 장전 시뮬레이션을 하는 과정이 뒤따르게 되는데 이 과정은 본 논문의 범위에서 벗어나므로 생략한다.

6. 결론

본 논문에서는 계층적 군집화를 기본으로 하는 실용적인 군

표 2. 각 그룹의 크기, 그룹평균 및 표준편차

그룹번호	그룹의 크기	평균	표준편차
1	48	0.838069	0.00230224
2	34	0.999487	0.00406833
3	28	1.01076	0.004026
4	23	1.0348	0.00308331
5	20	1.04823	0.00484371
6	23	1.06298	0.00188103
7	31	1.0755	0.00377078
8	79	1.08656	0.00333493
9	51	1.0991	0.0021584
10	110	1.10911	0.00280065
11	188	1.12015	0.00349963
12	123	1.12971	0.00215393
13	242	1.13824	0.0028039
14	236	1.14996	0.00383787
15	382	1.16444	0.00348817
16	147	1.17701	0.00348927
17	121	1.18641	0.00238808
18	209	1.19484	0.00259339
19	182	1.20332	0.00235045
20	194	1.2143	0.00419683
21	117	1.22749	0.00298443
22	73	1.24026	0.00392564
23	35	1.25302	0.00251308
24	18	1.2636	0.00171664
25	24	1.27071	0.00245802
26	31	1.29006	0.00206145
27	26	1.29814	0.00301455
28	15	1.31169	0.00449166
29	42	1.3398	0.00346436
30	28	1.35739	0.00290455
31	22	1.3713	0.00280141
32	8	1.38036	0.0016239

집화 과정을 제시하고 적용사례를 보였다. 군집화 분석에서 난점으로 여겨지는 그룹수의 결정은 목적함수를 도입하여 그래프를 이용하는 방법으로 접근하였고, 군집화 결과의 타당성에 대한 통계적 검정도 시도하였다. 특히 원자핵연료 분류의 실제 사례를 통해 그룹수 결정 및 군집화 결과에 대한 타당성 검정을 실시하는 전반적인 과정을 제시하였다.

계층적 군집화 기법의 경우 그룹 간의 거리를 정의하는 방법에 따라 서로 다른 군집화 결과를 주는데 주어진 예의 경우에는 최장거리 와 편차자승합 방법은 비교적 효과적인 방법이라고 볼 수 있는 반면 일반적으로 많이 쓰이는 최단거리방법은 좋지 못한 결과를 보이고 있다. 이를 통해 군집화방법의 선택과 사용은 반드시 주어진 문제에 따라 결정되어야 함을 알 수 있다.

군집화의 타당성 검정을 위한 U-통계량 방법은 규모가 너무 크지 않는 표본에 대해 아주 효과적인 것으로 보인다. 그러나 표본규모가 크면 계산량이 너무 커져서 전체 데이터에 대한 분석이 어려워진다. 보다 적은 계산량을 요구하는 새로운 통계량이 개발되면 유용할 것이다.

참고문헌

Baker, F. B. (1975), Measuring the Power of Hierarchical Cluster Analysis, *Journal of the American Statistical Association*, 70(349), 31-38.

Chao, A. (1992), Estimating the Number of Classes Via Sample Coverage, *Journal of the American Statistical Association*, 87(417), 210-217.

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

Gordon, A. D. (1994), Identifying genuine clusters in a classification, *Computational Statistics & Data Analysis*, 18, 561-581.

Hand, D. J. (1981), *Discrimination and Classification*, Wiley, New York.

Jolliffe, I. T. (1995), Identifying influential observations in hierarchical cluster analysis, *Journal of Applied Statistics*, 22(1), 61-80.

Krzanowskil, W. J. (1995), *Recent Advanced in Descriptive Multivariate Analysis*, Oxford.

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.

강 금 석

중국 연변대 수학과 학사

홍익대학교 응용수학 석사

현재: 연변대 수학과 교수, 홍익대 수학과 박사과정

관심분야: 확률모형 분석, 클러스터링

윤 복 식

서울대학교 산업공학과 학사

서울대학교 산업공학과 석사

미국 U. C. Berkeley 대학 박사

현재: 홍익대학교 기초과학과 응용수학 분야 교수

관심분야: 응용확률론, 통신시스템 분석, 메타휴리스틱스, 시뮬레이션

이 용 주

서울대학교 경영학과 학사

KAIST 경영과학과 석사

미국 Columbia대학교 경영학과 박사

현재: 이화여자대학교 경영대학 교수

관심분야: 선형 및 정수계획법의 응용, 대기행렬모형 분석