Journal of the Korean
OR/MS Society
Vol. 15, No. 2, Dec., 1990

中心體 目的函數를 이용한 多次元 個體 CLUSTERING 기법에 관한 연구

李 鐵* · 姜錫昊*

A Study on Multi-Dimensional Entity Clustering Using the Objective Function of Centroids

Chul Rhee* · Sukho Kang*

Abstract

A mathematical definition of the cluster is suggested. A nonlinear 0-1 integer programming formulation for the multi-dimensional entity clustering problem is developed. A heuristic method namd MDEC (Multi-Dimensional Entity Clustering) using centroids and the binary partition is developed and the numerical examples are shown. This method has an advantage of providing bottle-neck entity informations.

1. 서 론

생산시스템은 후기 산업화의 과정과 함께 보다 정성적이고 복잡한 시스템으로 변이하고 있다. 이에 따라 GT(Group Technology), 정보시스템 분석 등 시스템의 구조를 파악하는 문제들이 clustering기법을 사용하여 연구되고 있으며 이들 문제들은 통계적 clustering과는 달리 두 가지 유형의

* 서울大學校 産業工業科

개체들을 clustering의 대상으로 삼는다는 특징이 있다. 그러나 GT의 경우에서의 기계류, 부품류, 작업자, 공구류나 정보시스템 분석에서의 업무, 자료, 조직 등과 같이 실제 문제에서는 대상 개체의 유형이 2가지 이상으로 나타나 기존의 2차원 개체 clustering기법들은 적용이 어렵다. 다차원 개체 clustering은 여러 유형의 개체들을 동시에 고려하여 clustering하는 문제이다.

일반적으로 clustering 문제는 준구조화 문제(semi-structured problem)로 간주되어 왔는데 그 이유는 cluster 및 clustering에 대한 만족스러운 수학적 정의가 나타나지 않고 있기 때문이다. 따라서 특히 cluster의 수가 미리 결정되어 있지 않은 경우에는 각 연구자에 따라 나름대로의 clustering개념과 방법을 사용하고 있기 때문에 cluster의 수학적 정의나 문제의 정식화는 이루어지지 않았다.

본 연구에서는 N차원 clustering문제에 대하여 기존의 여러 개념들을 분석하여 cluster의 수학적 정의를 내리고 이에 따른 정식화와 해법을 개발하는데 그 목적이 있다.

2. Clustering Objective와 Contiguity

W. D. Fisher〔06〕이래의 많은 연구들은 주어진 cluster수에 맞추어 cluster내의 동질성이 높아지도록 하는 'grouping problem'을 연구해 왔다. 그러나 대부분의 응용분야에서는 cluster의 수를 알 수 없으므로 cluster의 수가 정해져 있다는 가정은 비합리적이다. 이를 극복하기 위한 노력으로 계층적 clustering〔17〕이나 'stopping rule'〔12〕과 같은 연구가 있었으나 분석자의 임의에 cluster의 수의 결정을 맡긴다는 단점을 극복하지 못하였다.

이는 동질성 뿐만 아니라 이질성을 척도에 반영함으로써 해소시킬 수 있다. 〔01〕〔09〕〔15〕 이러한 clustering의 목적함수의 수립에 있어 다음의 두 가지 개념이 사용되었다.

* 대표값 : cluster의 성격을 대표할 수 있는 양을 정의하여 이를 이용하여 목적함수 수립
* 개체쌍 : cluster 내의 개체쌍들간의 관계(거리 또는 친밀도)를 이용하여 목적함수 수립

의미상으로 대표값을 이용하는 것이 문제의 성격을 잘 반영하는 반면 비선형 함수로 표현되는 단점이 있고 반면 개체쌍을 이용하는 경우 선형으로의 표현이 가능하다. 그러나 개체쌍을 단위로 삼을 경우 개체들이 서로 다른 cluster에 소속되게 하는 논리 구성, 즉 이질성 척도의 개념형성이 어려워진다. 이에 따라 transitivity나 친밀도에 대한 chain의 개념에 대한 연구〔07〕〔10〕〔14〕가 이루어 졌으나 일반적으로 받아들여질 수 있는 척도는 개발되지 못하였다.

이들의 연구를 살펴보면 contiguity의 개념이 매우 중요함을 알 수 있다. 본 연구에서 사용하는 contiguity란 다음 성질을 말한다.

정의 1. (Contiguity) 개체들의 집합$\{e_i\}$이 있다 하자. 함수 d가 거리를 나타낼 때 (1), (2)가 만족될 때 (3)이 만족되면 contiguity를 만족시킨다고 한다.

(1) $d(e_i, e_j) \leqq d(e_i, e_k)$

(2) $d(e_k, e_i) \leqq d(e_k, e_j)$

(3) $d(e_j, e_i) \leqq d(e_j, e_i)$

일반적으로 최적해에서 contiguity가 성립한다면 개체쌍을 기본단위로 목적함수를 수립하는 것이 합리적이다. the weighted sum of square를 목적함수로 삼은 W. D. Fisher〔06〕는 다음의 'the contiguous partition property'를 보였다.

정리 1. (The Contiguous Partition Property) 개체 i, j, k가 속성수치(attribute number)에 의해 $i<j<k$이며 i와 k가 최적해에서 동일 cluster에 속한다면 j도 동일한 cluster에 소속된다.

H. D. Vinod〔16〕는 거리 개념을 Euclidean distance로 확장한 the general string property를 제시하였다.

"개체 i와 j의 Euclidean distance를 d_{ij}라 하자. 만일 개체 i와 j가 최적해에서 동일 cluster에 소속된다면 $d_{ij} \leqq d_{kj}$를 만족하는 모든 k개체도 동일 cluster에 소속된다."

그러나 M. R. Rao〔13〕는 반례를 통하여 Vinod의 the general string property가 성립하지 않음을 보이고 다음의 the string condition을 보였다.

정리 2. (The String Condition) 최적해에서 각 cluster에는 적어도 하나의 leader 개체가 존재하여 동일 cluster내의 개체들과 leader와의 거리는 다른 cluster의 개체들과 leader와의 거리보다 크지 않다.

따라서 contiguity는 일반적으로 성립하지 않으며 개체쌍방 단위목적함수는 적절한 것이 될 수 없어 대표값을 기본으로 삼아야 한다. partition problem과 clustering problem의 차이는 이 the string condition의 충족여부에서 비롯된다고 할 수 있다. 이에 따른 cluster의 수학적 정의는 다음과 같다.

정의 2. (cluster) 개체들의 집합 $E=q\{e_1, e_2, \cdots, e_n\}$와 이들의 부분집합들의 집합 $S=\{S_1, S_2, \cdots, S_k\}$, 그리고 개체간의 거리척도 d가 주어졌다 하자. 부분집합 S_i의 leader를 l_i라 할때 다음 조건을 만족하는 S를 cluster 해, 각 S_i를 cluster라 부른다.

$d(e_i, l_j) \leqq d(e_i, l_k)$

where

$e_i \in S_j$

$k \neq j$

3. 다차원 개체 clustering 모형의 수립

3.1 용 어

N : 개체차원의 수

E^i : i차원 개체들의 집합

$E : E^1 \times E^2 \times \cdots \times E^N$

$E/E^i : E^1 \times E^2 \times \cdots E^{i-1} \times E^{i+1} \times \cdots \times E^N$

e^i_j : i차원의 색인이 j인 개체

$e^i_{(j)}$: i차원의 현재의 순서가 j번째인 개체

$R(e^1_{j_1} \times e^2_{j_2} \times \cdots \times e^N_{j_N})$: 개체 $e^1_{j_1}, e^2_{j_2}, \cdots e^N_{j_N}$간의 관계

C : cluster 해

C_j : 색인이 j인 cluster

C_j^i : C_j cluster의 i 차원 개체들의 집합

3.2 가 정

가정 1 : 동일차원의 개체 사이의 직접적 관계는 존재하지 않는다.

가정 2 : 개체간 관계는 0−1으로 주어진다.

가정 3 : 개체들의 중요도는 동일하다.

3.3 개체간 거리

동질성과 이질성은 동전의 양면과 같다. 즉, 이질성과 동질성은 서로 동일 차원상에 측정되어지며 상호상보(Mutually Complementary)의 성질을 갖는다. 대부분의 연구에서 이들 척도는 거리 또는 친밀도의 개념으로 표현되고 있다. L. Hubert〔08〕는 개체간 친밀도에 대하여 symmetry, positivity, nullity의 3가지 성질을 가정할 수 있다고 주장하였다.

이러한 사고는 다차원 개체의 경우에서도 적용될 수 있을 것이다. 2차원 개체 clustering의 경우에는 여러 가지 유형의 유사계수들이 제안되어 있는데 M. P. Chandrsekharan과 R. Rajagopalan〔04〕에 의하면 다음의 Jaccard 유사계수가 가장 합리적이라고 한다.

$$\text{Jaccard 유사계수} = \frac{\text{관계의 값이 모두 1인 관계의 수}}{\text{관계의 총수}}$$

이 Jaccard 유사계수는 Hubert의 3가정을 모두 만족한다. 그러나 Jaccard 유사계수는 metric이 아니다. 거리가 mertic이 아닌 경우에는 중심체의 추정이 어려워지는 문제가 발생한다. 이러한 논의의 연장선상에서 city block distance〔04〕를 확장하여 다차원 개체간의 거리는 다음과 같이 Minkowsky metric 함수의 특별한 형태로서 정의할 수 있다. i차원의 개체 j, k가 주어졌을 때 이들간의 거리 SDBE(e_j^i, e_k^i) (Smoothed Distance Between Entities)는 다음과 같다.

$$SDBE(e_j^i, e_k^i) = \sum_{K \in E/E'} |r(e_j^i \times K) - r(e_k^i \times K)| / |E/E'|$$

이 SDBE 척도는 Hubert의 3가정을 만족하는 metric이다.

3.4 중심체(Centroid)

특정한 cluster가 주어졌을 때, 해당 cluster의 대표값을 표현하는 방법은 여러 가지가 있을 수 있다. 일반적으로 cluster가 제공하는 속성들은 개체간 관계의 형식으로 나타나므로 대표값은 개체간 관계의 함수형태를 갖게 된다. 본 연구의 경우 각 차원이 독립임을 가정하고 있으므로 i 차원의 대표값은 다음과 같이 표현하여 평균의 의미를 갖도록 하는 것이 타당할 것이다.

cluster $C_j = \{C_j^1, C_j^2, \cdots, C_j^i\}$가 주어졌을 때 i 차원의 대표값 $cent_j^i = \{cent_j^i(K) \mid K \in E/E^i\}$는 다음과 같다.

$$cent_j^i(K) = \sum_{m \in C_j^i} R(m \times K) / |C_j^i|$$

이 중심체 정의는 특정 cluster의 i 차원의 한 개체를 임의 선정하였을 때 다른 차원의 임의 개체와 관계가 있을 확률을 나타내며 Shannon의 entropy와도 일치한다.

3.5 목적함수

Clustering의 목적으로서 연구자들이 공통적으로 지적하는 것이 cluster 내 개체들의 동질성과 cluster간의 이질성이다〔09〕. 이질성은 cluster

의 대표값들간의 차이로 생각할 수 있다.

3.5.1 Cluster내의 동질성

특정 cluster C_j에서 i차원 개체들의 동질성은 다음과 같이 정의된다.

$$Hom_j^i = \sum_{e_k^i \in C_j} \{1-SDBE(cen_j^i, e_k^i)\} / |C_j^i|$$

즉 i차원의 동질성은 소속 개체들의 중심체로부터의 평균거리로 표현된다. 이에 따라 다음과 같이 평균 차원내 동질성을 정의할 수 있다.

$$Hom_j = \sum Hom_j^i / N$$

1차원에서와는 달리 다차원 개체의 경우 차원내 동질성 척도만으로는 충분치 않다. 왜냐하면 cluster내의 개체들이 서로 다른 차원의 개체들로 이루어져 있으므로 차원간의 연관도가 다시 고려되어야 하기 때문이다. 이러한 관점에서의 척도는 M. P. Chandrasekharan과 R. Rajagopalan의 ultimate efficiency 척도[02][03]가 가장 적합하다. 이를 다차원으로 확장한 cluster내의 개체간 관계의 평균밀도 ρ_j. 즉

$$\rho_j = \frac{\sum_{I \in C_j^1 \times C_j^2 \times \cdots \times C_j^i} R(I)}{|I|}$$

는 cluster내의 차원간 관련도(Interdependency)를 나타낸다. 이 두 척도를 결합하여 특정 cluster의 동질성 척도를 얻을 수 있다.

$$HOM_j = \rho_j \cdot Hom_j$$

cluster해 $C = \{C_1, C_2, \cdots, C_k\}$가 주어졌을 때 전체 동질성은 개체간 관계의 규모로 가중하여 얻을 수 있다.

$$HOM_j = \frac{\sum_j \prod_i |C_j^i| \cdot HOM_j}{\sum_j \prod_i |C_j^i|}$$

3.5.2 Cluster간의 이질성

Cluster간의 이질성 측정을 위해서 전술한 중심체를 이용한다. 즉 중심체는 해당 cluster의 속성을 대변하고 있으므로 이들 중심체들의 차이를 척도로 삼을 수 있는 것이다.

이때 유의해야 할 사항은 중심체들의 대표값을 기준으로 할 수 없다는 점이다. 중심체들의 대표값 개념을 사용할 경우 정보의 손실이 발생하여 기대하는 결과를 가져오지 못하는 경우가 발생하기 때문이다. 이것은 병목개체가 중요한 의미를 가지는 경우 특히 그러하다. 이철과 강석호의 논의[18]에서 알수 있듯이 대부분의 clustering 문제에서 병목개체는 분석자에게 대안을 위한 많은 정보를 제공할 뿐 아니라 우수한 cluster 해의 탐색을 위해서도 중요한 비중을 갖는다.

따라서 본 연구에서는 이질성을 임의의 두 중심체를 선정하였을 때 이들 사이의 기대 거리로 삼는다.

$$Het = \sum_{i(j,k)} \sum_{K} \sum SDBE(cent_j^i(K), cent_k^i(K)) / |K| \cdot N \cdot |C|$$

차원간 이질성도 동질성과 마찬가지로 다음과 같이 수립한다.

$$\pi = \frac{\sum_{I \notin C_j^1 \times C_j^2 \times \cdots \times C_j^i} \{1-R(I)\}}{|I|}$$

이를 결합하여 이질성 척도를 얻는다.

$$HET = \pi \cdot Het$$

3.5.3 종합척도

전술한 동질성과 이질성 척도는 공히 최대화를 요구한다. 이들 양 척도를 결합하는 방법에는 여러가지가 있겠으나 본 연구에서는 다음과 같은

convex combination을 제안한다.

$$Z = w \cdot HOM + (1-w) \cdot HET$$

3.6 정식화

의사결정 변수로서 cluster의 수와 개체의 할당을 표현하도록 다음과 같이 정의한다.

$X_{ijk} = 1$, if $e_j^i \in C_k$

$= 0$, otherwise

$Y_k = 1$, if there exists C_k

$= 0$, otherwise

이때 중심체는 다음과 같이 표현된다.

$$cent_k^i(K) = \sum_{K \in E/E^i} R(j \times K) \cdot X_{ijk} / \sum X_{ijk}$$

차원간 동질성 P_k는

$$\rho_k = \sum_i R(I) \cdot \prod_{ij \in I} X_{ijk} / \prod_{ij \in I} X_{ijk}$$

로 나타난다. 차원간 이질성은 다음과 같다.

$$\pi = \sum_i (1-R(I)) \cdot (1 - \sum_k \prod_{ij \in I} X_{ijk}) / (1 - \sum_k \prod_{ij \in I} X_{ijk})$$

따라서 주어진 X, Y로 목적함수 값 Z(X, Y)를 표현할 수 있다.

제약식에 대하여 고려하여 보자. 모든 개체는 cluster 되어야 하므로

$$\sum_k X_{ijk} \geq 1, \ \forall \ i, j, 1 \leq k \leq \min\{|E^i|\} \quad (1.1)$$

이며 만일 overlapping cluster를 허용하지 않는 경우, 즉 병목개체를 고려하지 않을 경우에는 등호를 사용한다. cluster와 개체와의 관계에서

$$Y_k \geq X_{ijk}, \ \forall \ i, \ j, \ k. \quad (1.2)$$

$$Y_k \leq \sum_i \sum_j X_{ijk}, \ \forall \ k. \quad (1.3)$$

이 성립한다.

the string condition으로부터 다음이 만족되어야 한다.

$$\sum_k X_{ijk} \{R(e_j^i \times K) \cdot (1-cen_k^i(K)) + (1-R(e_j^i \times K)) \cdot cent_k^i(K)\} \leq \sum_k X_{ijk} \{R(e_j^i \times K) \cdot (1-cen_n^i(K)) + (1-R(e_j^i \times K) \cdot cent_n^i(K)\}$$

$$\forall \ i, \ j, \ k, \ n \neq k \quad (1.4)$$

이에 따른 정식화는 다음과 같다.

Maximize $Z(X, Y)$ (1.0)

Subject to

$$\sum X_{ijk} \geq 1, \forall \ i, j, 1 \leq k \leq \min\{|E^i|\} \quad (1.1)$$

$$Y_k \geq X_{ijk}, \ \forall \ i, j, k. \quad (1.2)$$

$$Y_k \leq \sum_i \sum_j X_{ijk}, \ \forall \ k. \quad (1.3)$$

$$\sum_k X_{ijk} \{R(e_j^i \times K) \cdot (1-cen_k^i(K)) + (1-R(e_j^i \times K)) \cdot cent_k^i(K)\} \leq \sum_k X_{ijk} \{R(e_j^i \times K) \cdot (1-cen_n^i(K)) + (1-R(e_j^i \times K)) \cdot cent_n^i(K)\}$$

$$\forall \ i, j, k, n \neq k \quad (1.4)$$

$$X, Y \in \{0.1\} \quad (1.5)$$

이 비선형 0−1 정수계획법 정식화는 0−1 정수 조건으로 인하여 일반적인 비선형계획법으로 접근하기 어렵다. 따라서 본 연구에서는 발견적 기법으로 접근한다.

4. 발견적 해법의 개발

4.1 Seed의 추정

Clustering 문제의 해를 구하는 것은 최적의 중심체 또는 leader들을 구하는 것과 동일하다. 최적 중심체 문제는 병목개체 판별문제와 동일한 complexity를 갖으며 NP−complete이다. 그러

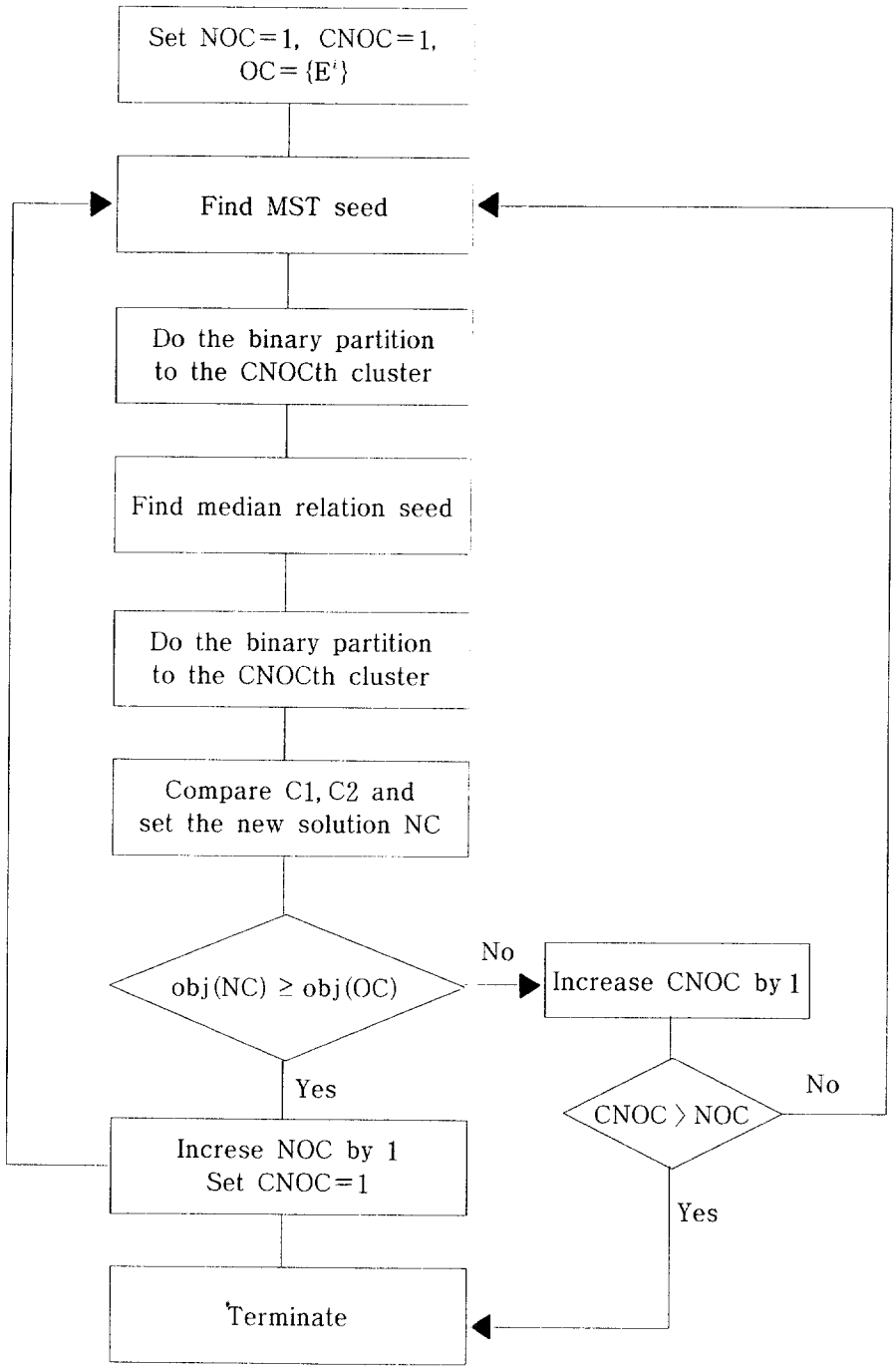

Set NOC=1, CNOC=1,
OC={E^i}
Find MST seed
Do the binary partition
to the CNOCth cluster
Find median relation seed
Do the binary partition
to the CNOCth cluster
Compare C1, C2 and
set the new solution NC
obj(NC) ≥ obj(OC)
No
Increase CNOC by 1
Yes
CNOC > NOC
No
Yes
Increse NOC by 1
Set CNOC=1
Terminate

그림-1 MDEC의 절차

나 기회측면에서 볼때 leader개체의 역할을 수행할 수 있는 개체의 수가 병목개체의 수보다 상대적으로 많으므로 leader 개체의 추정에 기법의 촛점을 맞추게 된다.

이렇게 추정된 leader 개체를 seed라 부른다. [02] seed를 잘못 추정하게 되는 경우 다음의 두가지 경우가 발생한다. [18]

* 병목개체의 선정
* 예외개체간 관계의 선정

이를 회피하기 위한 방법으로 minmum spaning tree의 다음 성질을 이용한다.

1. polynomial complexity
2. sigle-linkage relation

SDBE를 이용한 minmum spanning tree상에서 node로 표현된 각 개체는 연결된 edge의 수가 적을 수록 병목개체가 될 기회는 줄어들며 연결된 edge의 수가 많을 수록 예외 개체간 관계가 선택될 기회가 감소하는 성질을 보인다. 이를 이용화여 seed의 선정에 있어 다음 규칙을 사용할 수 있다.

규칙 1. 1차원에서 branch node로부터의 SDBE가 가장 큰 leaf node를 선정한다.

규칙 2. 2−N 차원까지 선정된 개체와 관계가 가장 많은 개체를 선정한다.

4.2 The binary partition

Seed의 추정이 동시에 적정 개수가 이루어질 경우에는 문제가 없으나 동시에 복수의 seed를 추정하기가 매우 어려우므로 개발된 목적함수를 이용하여 binary partition을 수행함으로써 단계적으로 seed를 추정하여 나간다.

하나의 seed가 선정되었을 때 이에 가까운 개체들을 선택하여야 하는데 이를 위하여서는 sorting을 하여야 하나 여기에서 contiguity의 문제가 발생한다. 즉, 만일 contiguity가 성립할 경우 ordeing후 순차적인 목적함수 값 test만으로 cluster를 결정할 수 있으나 contiguity가 성립하지 않을 경우 추정된 cluster에 소속되어야 할 개체가 누락될 수 있다.

정리 3. 다차원 개체 clustering에 있어 SDBE 거리 척도하에서는 contiguity가 성립하지 않는다.

증명 4. 2차원 개체 clustering의 (그림−1)과 같은 경우를 고려하여 보자. 개체 c, d를 제외하면 이상적인 cluster의 모양이다. 각 cluster의 개체 수가 충분히 많다고 할 때 c와 d를 추가하는 경우 cluster해는 변화하지 않는다. 이때 leader 개체인 a, b와 개체 c, d사이의 SDBE는 contiguity가 성립하지 않음을 보여준다.

(그림−1)에서 볼 수 있 듯이 최적해에서도 contiguity는 성립하지 않는다. 그러나 최적해에서 나타나는 예외개체간 관계의 수가 상대적으로 작은 경우에는 contiguity가 성립한다.

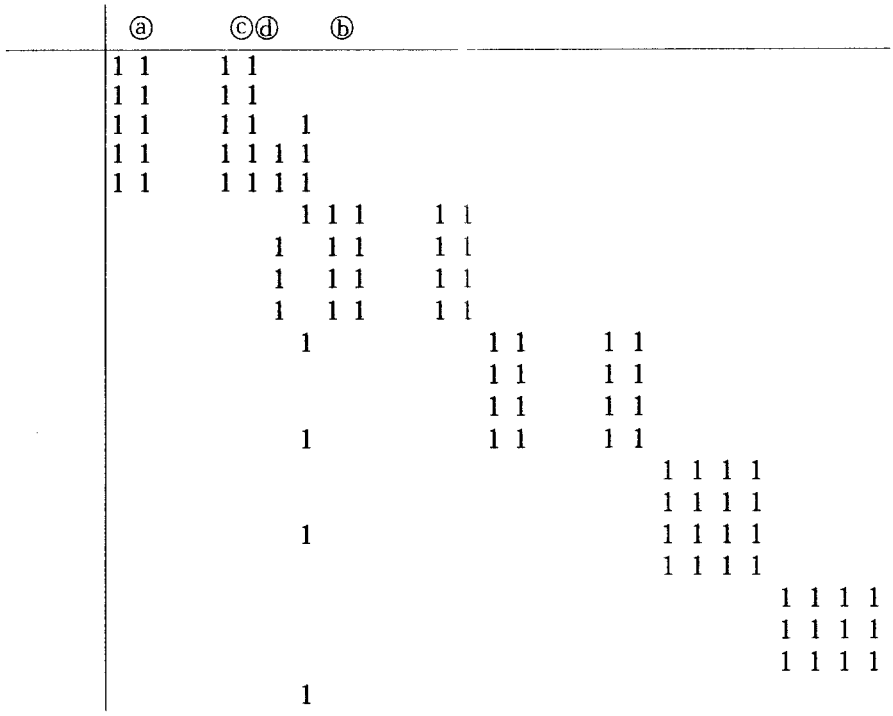
따름정리 3.1 i차원의 개체 a, b, c, d를 상정하여 보자. 다음 (1), (2), (3)이 충분히 작은 ε에 대하여 성립하면 (4)도 성립한다.

(1) $SDBE(a, c) \leq SDBE(a, d)$

(2) $SDBE(a, d) \leq SDBE(b, d)$

(3) $\sum_{K \in E/E'} |(c \times K) - R(d \times K)| (1 - R(a \times K)) (1 - R(b \times K)) \leq \varepsilon$

(4) $SDBE(a, c) \leq SDBE(b, c)$



SDBE(a, c) ≤ SDBE(a, d)
SDBE(a, d) ≤ SDBE(b, d)
SDBE(b, c) < SDBE(a, c)

그림-2 Non Contiguous Example

증명 다음과 같이 E'의 색인집합을 분리하자.

$I=\{K \mid K \in E/E',\ R(a\times K) \neq R(b\times K)\}$

$I^a=\{K \mid K \in I,\ R(a\times K)=1\}$

$I_b=\{K \mid K \in I,\ R(b\times K)=1\}$

(4)가 성립하지 않는다고 가정하자. 그러면

$SDBE(b, c) \leqq SDBE(b, d)$

이고 ε가 충분히 작은 경우이므로

$$\sum_{I^b}(1-R(d\times K))+\sum_{I^a}R(d\times K)>\sum_{I^b}(1-R(c\times K))+\sum_{I^a}R(c\times K)-\sum_{I^b}R(d\times K)+\sum_{I^a}R(d\times K)>-\sum_{I^b}R(c\times K)+\sum_{I^a}R(c\times K)\sum_{I^b}R(c\times K)-\sum_{I^a}R(c\times K)>\sum_{I^b}R(d\times K)-\sum_{I^a}R(d\times K)\sum_{I^a}(1-R(c\times K))+\sum_{I^b}R(c\times K)>\sum_{I^a}(1-R(d\times K))+\sum_{I^b}R(d\times K)$$

그러므로

$SDBE(a, c) > SDBE(a, d)$

이므로 모순이다.

따라서 instance의 최적해가 높은 동질성 및 이질성을 보이는 경우에만 contiguity가 성립함을 알 수 있으며 일반적으로 성립하지 않음을 알 수 있다. 이런 측면에서 볼 때 많은 연구자들이 사용하고 있는 seeding and searching방식의 기법은 최적해의 목적함수 값이 상대적으로 낮은 경우 좋은 해를 제공하지 못함을 알 수 있다.

본 연구에서는 seeding and searching을 수행하되 contiguity를 만족시키도록 재배열하는 절차가 추가된다. 이에 따른 binary partition의 절차는 다음과 같다.

Step 1. Seed에 대한 SDBE의 오름차순으로 개체를 정렬한다.

Step 2. 모든 i에 대하여 $B=e^{i}_{(j(i))}$, $j(i)=1$로 놓는다.

Step 3. B에 포함시킬 경우 목적함수 값이 감소하지 않는 $e^{i}_{(j(i)+1)}$가 존재하면 B에 포함시키고 $j(i)=j(i)+1$로 놓는다. 해당 개체가 존재하지 않을 때까지 반복한다.

Step 4. 해당 개체가 존재하지 않으면 나머지 개체들을 B^c에 넣는다.

Binary partion을 수행하여 얻은 cluster는 서로 이질성이 높게 형성이 되므로 더 이상의 목적함수의 개선이 일어나지 않을 때까지 계속하여 cluster해를 얻을 수 있다. 그러나 이 해는 overall optimum을 보장하지 못한다.

4.3 중간 개체간 관계 seed 선정(Median Relation Seed)

전술하였듯이 병목개체의 회피의 complexity는 NP-complete이다. 따라서 추정된 seed는 병목개체를 포함하고 있을 수 있다. 이를 회피하기 위한 방법으로 중간 개체간 관계를 seed로 삼는 방법이 알려져 있다. [18]

이를 이용하여 C_j를 partition 하여 얻어진 (C_{j_1}, C_{j_2})와 중간 개체간 관계 선정 seed를 이용한 partition (C'_{j_1}, C'_{j_2})의 목적함수를 비교하여 우수한 partition을 선택한다.

4.4 Algorithm (MDEC; Multi-Dimensional Entity Clustering)

Step 1. $noc=1$, $n=1$, $OC=\{E\}$로 놓는다.

Step 2. n번째 cluster에 대한 MST seed를 구한다.

Step 3. Binary parition을 수행하여 그 결과를 C1으로 놓는다.

Step 4. 중간 개체간 관계 seed를 구한다.

Step 5. Binary partition을 수행하여 그 결과를 C2로 놓는다.

Step 6. C1, C2 중 목적함수 값이 우수한 해를 선택하여 NC로 놓는다.

Step 7. 만일 NC가 OC보다 우수하면 $noc=noc+1$, $n=1$, $OC=NC$로 놓고 step 2로 간다. 아니면 $n=n+1$로 놓는다.

Step 8. $n>noc$이면 terminate, 아니면 step 2로 간다.

5. 예 제

다른 기법들과의 비교를 위하여 F. F. Chen[05]의 예제에 대하여 본 기법을 적용하였다. (그림-3)의 기계-부품 관계 행렬에 대한 Chen의 해를 cluster의 수가 2, 3인 경우에 대하여 (그림-4), (그림-5)에 보였다. 이에 대하여 본 기법의 해인 (그림-6)는 예외개체간 관계가 1, cluster의 수가 4로 Chen의 해보다 개선된 결과임을 알 수 있다.

기존의 GT기법에서 도입하지 못하였던 공정의 선후관계를 고려한 3차원 개체 clustering을 고려하여 보자. 〈표-1〉에 공정의 유통경로를 보였다. 이 경우 흐름의 방향을 고려하여야 하므로 이 시스템에 존재하는 유통 ARC의 집합을 또 하나의 개체 차원으로 도입한다. 〈표-2〉 이에 따른 개체

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1			1	1				1	1	1			1		1
2			1					1	1			1	1		
3	1						1								
4					1	1					1				
5		1		1				1	1	1			1		1
6						1								1	
7					1						1				
8					1						1				
9	1						1								

그림-3 입 력 행 렬

	2	3	4	8	9	10	12	13	15	1	5	6	7	11	14
1		1	1	1	1	1		1	1						
2		1		1	1		1	1							
5	1		1	1	1	1		1	1						
3										1			1		
4											1	1		1	
6												1			1
7											1			1	
8											1			1	
9										1			1		

그림-4 Chen의 2 cluster solution

	2	3	4	8	9	10	12	13	15	5	6	11	1	7	14
1		1	1	1	1	1		1	1						
2		1		1	1		1	1							
5	1		1	1	1	1		1	1						
4										1	1	1			
7										1		1			
8										1		1			
3											1		1	1	
6															1
9													1	1	

그림-5 Chen의 3 cluster solution

	14	6	1	7	5	11	3	12	2	15	13	8	9	4	10
6	1	1													
3			1	1											
9			1	1											
4		1			1	1									
7					1	1									
8					1	1									
2							1	1			1	1	1		
1							1			1	1	1	1	1	1
5									1	1	1	1	1	1	1

그림-6 발견적 기법 MDEC의 적용해

간 관계를 〈표-3〉에 보였다. 이 3차원 개체 clust ering을 푼 결과가 〈표-4〉이다. 이 예제에서 M 4는 병목개체로서 만일 M4를 하나 더 도입한다면 두 cluster가 독립적으로 구성될 수 있음을 알려 주고 있다.

w값은 동질성과 이질성의 비중이라는 의미를 갖는바 w의 값에 대한 sensitvity를 분석하기 위하여 w를 0.4 - 0.6까지 변화시켜본 결과 0.4에서 cluster의 수가 3으로 감소하였다. 예제의 수를 증가시킨다 하여도 w값의 통계적 유의성은 보장되기 어렵겠으나 Chen의 3 예제에 대한 실험결과 0.5에서 가장 우수한 해를 얻을 수 있었으며 0.45 - 0.6 사이에서는 해의 값의 변화가 없었으며 따라서 일반적인 문제의 경우 w=0.5의 값이 적절하리라 생각된다.

〈표-1〉 공정 가공경로

P1	M1-M2-M3-M4
P2	M2-M4-M1
P3	M2-M3-M4
P4	M5-M6-M4-M7
P5	M5-M6-M4
P6	M7-M5-M6

〈표-2〉 공정간 유통 ARC개체

A1	M1-M2
A2	M2-M3
A3	M3-M4
A4	M2-M4
A5	M4-M1
A6	M5-M6
A7	M6-M4
A8	M4-M7
A9	M7-M5

〈표-3〉 R= 1인 개체간 관계

P1	A1	M1 M2
	A2	M2 M3
	A3	M3 M4
P2	A4	M2 M4
	A5	M1 M4
P3	A2	M2 M4
	A3	M3 M4
P4	A6	M5 M6
	A7	M4 M6
	A8	M4 M7
P5	A6	M5 M6
	A7	M4 M6
P6	A9	M5 M7
	A6	M5 M6

〈표-4〉 Cluster 해

Cluster 1

P1, P2, P3
M1, M2, M3, M4
A1, A2, A3, A4, A5

Cluster 2

P4, P5, P6
M4, M5, M6, M7
A6, A7, A8, A9
병목개체 : M4

6. 결 론

다차원 개체 clustering문제에 대한 정식화와 이에 따른 발견적 기법이 제시되었다. 이 발견적 기법은 non-contiguity, 병목개체 등을 고려한다는 측면에서 기존의 seeding and searching, graph theoretic method 등의 단점을 극복할 수 있으리라 기대된다. 또한 본 연구에서 제시된 cluster의 수학적 정의는 clustering에 대한 앞으로의 연구에 도움이 될 수 있을 것으로 기대된다. 본 기법은 GT, 정보시스템 분석등에 응용될 수 있다.

모듈화 문제, part type selection, GT문제와 같은 여러 생산 시스팀 관련 문제에서는 용량 제약, alternative route과 같은 선형 제약이 존재하므로 추후 이의 해결을 위한 연구가 필요하다 하겠다.

References

[01] Arnold, S. J., "A test for clusters", Journal of Marketing Research, 1979, Vol. XVI, November, pp. 545~551.

[02] Chandrasekharan, M. P. and Rajagopalan, R., "An ideal seed non-hierarchical clustering algorithm for cellular manufacturing", International Journal of Production Reseach, 1986, Vol. 24, No. 2, pp. 451~464.

[03] Chandrasekharan, M. P. and Rajagopalan, R., "ZODIAC-an algorithm for concurrent formation of part-families and machine-cells", International Journal of Production Research, 1987, Vol. 25, No. 6, pp. 835~850.

[04] Chandrasekharan, M. P. and Rajagopalan, R., "Groupability : an analysis of the properties of binary data matrices for group technology", International Journal of Production Research, 1989, Vol. 27, No. 6, pp. 1035~1052.

[05] Chen, F. F., "An integrated production planning system for flexible manufacturing", Ph. D dissertation, University of Missouri-Columbia, May, 1988.

[06] Fisher, W. D. and College, K. S., "On grouping for maximum homogeneity", American Statistical Association Journal, December, 1958, pp. 789~798.

[07] Frank, O. and Harary, F., "Cluster inference by using transitivity indices in empirical graphs", Journal of the American Statistical Association, December, 1982, vol. 77, No. 380, pp. 835~840.

〔08〕 Hubert, L., "Approximate evaluation techniques for the single-link and complete-link hierarchical clustering procecures", Journal of the American Statistical Association, September, 1974, Vol. 69, No. 347, pp. 698~704.

〔09〕 Jensen, R. E., "A dynamic programming algorithm for cluster analysis", Operations Research, 1969, Vol. 17, No. 6, pp. 1034~1057.

〔10〕 Ling, R. F., "A probability theory of cluster analysis", Journal of the American Statistical Association, March, 1973, Vol. 68, No. 341, pp. 159~164.

〔11〕 Ling, R. F., and Killough, G. G., "Probability tables for cluster analysis based on a theory of random graphs", Journal of the American Statistical Association", 1976, Vol. 71, No. 354, June, pp. 293~300.

〔12〕 Mojena, R., "Hierarchical grouping methods and stopping rules : an evaluation", The Computer Journal, 1977, Vol. 20, No. 4, pp. 359~363.

〔13〕 Rao, M. R., "Cluster analysis and mathematical programming", Journal of the American Statistical Association, 1971, Vol. 66, No. 335, September, pp. 622~626.

〔14〕 Robinson, D. and Duckstein, L., "Polyhedral dynamics as a tool for machine-part group formation", International Journal of Production Research, 1986, Vol. 24, No. 5, pp. 1255~1266.

〔15〕 Stanfel, L. E., "A recursive Lagrangian method for clustering problems", European Journal of Operational Research, 1986, Vol. 27, pp. 332~342.

〔16〕 Vinod, H. D., "Integer programming and the theory of grouping", American Statistical Association Journal, June, 1969, pp. 506~519.

〔17〕 Ward, Jr., J. H., "Hierarchical grouping to optimize an objective function", American Statistical Association Journal, March, 1963, pp. 236~244.

〔18〕 李鐵, 姜錫昊, "다차원 개체를 위한 차이등급 clustering", 韓國經營科學會誌, 1989年 6月, 第14券, 第1號, pp. 108~118.