다차원 CLUSTERING 문제를 위한 척도에 관한 연구

(A Study on the Measurement for Multidimensional Entity Clustering)

서울대 산업공학과 이 철*

1. 서 론

일반적으로 cluster의 수가 미정인 상황하에서의 clustering 문제는 semistructured 문제로 알려져있다. clustering 문제를 구조화하는데 있어서 해의 품질평가(evaluation of solution quality)가 필수적이나 각 응용분야에 널리 적용될 수 있는 척도는 아직까지 개발되어있지 못한 상태이다. 그 주된 원인은 cluster 해에 대한 개념적 차원에서의 평가기준은 제시되어 있으나 척도의 구현에 있어서는 제시된 개념들이 명확하게 적용될 정도의 수준으로는 구체화되지 못한 데에 기인한다고 할 수 있다.

본 연구의 목적은 개체차원이 다차원으로 확장된 clustering 문제를 대상으로 하는 clustering 문제의 척도 개발에 있다.

2. 다차원 개체 clustering

0-1 다차원 개체 clustering은 최근 이 철과 강 석호[1]에 의해 제시된 문제이다. 기존의 clustering 문제와는 달리 N 가지 type의 개체들을 clustering의 대상으로 삼으며 objective는 cluster 내의 동질성과 cluster 간의 이질성을 높이는데 있다.

대상 개체들을 각 차원(type)에 따라 집합 $E^1, E^2, \ldots, E^N$으로 표시하자. 개체들간의 관계 $\mathcal{R}$ 은 입력자료로 주어진다.

$$\mathcal{R} = \{ r \mid r \in E^1 \times E^2 \times \ldots \times E^N \} \qquad \text{(식-1)}$$

feasible cluster 해를 $\mathcal{C}$ 로 표시하면 제약식은 다음과 같다.

$$\mathcal{C} = \{ C_1, C_2, \ldots, C_k \} \qquad \text{(식-2)}$$

where $C_i = C_i^1 \times C_i^2 \times \ldots \times C_i^N$,

$C_i^j \subseteq E^j$,

$C_i^j \neq \emptyset$,

$\cup_i C_i^j = E^j$.

목적함수를 H라 하면 H 는 개체간 관계들과 해에 의해 결정된다. 이때 다차원 개체 clustering 문제는 다음과 같이 쓸 수 있다.

$$\text{Max. } H(X, \mathcal{R}) \qquad \text{(식-3)}$$
$$\text{s.t. } X \in \mathcal{C}$$

이 문제는 combinatorics 의 성격을 띠고 있다. 따라서 H 함수의 합리적 설정은 조합최적화 기법의 도입을 가능케할 것으로 기대된다.

3. 척도의 연구현황

대부분이 1 차원 clustering을 대상으로한 통계학자들의 여러 연구에서 정량적 척도들이 제시되었다. Fisher[4]는 개체간 관계를 weighted sum of square로 표현하여 이를 최소화하려 하였다. Edward 와 Cavalli-Sforza[3]는 ANOVA를 이용하기 위하여 분산을 척도로 사용하였다. King[6]은 두 그룹의 centroid 사이의 correlation으로 척도를 삼았다. 이들 연구의 공통점은 1 차원 개체들을 대상

* 서울대학교 산업공학과

으로 하였으며 기본적으로 두개의 그룹간의 관계에서 유도된 것이라는데 있다. 또한 두 그룹간의 관계파악의 수단으로 통계적 기법이 적용되었으므로 그룹내의 개체 수가 비교적 큰 경우에 유효하다는 한계를 보이고 있다.

mathematical programming으로 접근하려는 일련의 시도들은 모두 cluster의 수가 이미 알려져 있는 것으로 가정하고 있다. Vinod[11]는 목적함수로서 within group sum of square(WGSS)를 설정하여 IP formulation을 하였다. Jensen[5]은 Euclidean metric function을 목적함수로 설정하고 DP formulation을 제시하였다. Rao[10]는 maximum distance within groups를 척도로 설정하였다. 이들 연구들은 근본적으로 clustering의 성격보다는 partition의 성격이 강하다고 볼 수 있다.

2 차원 개체 clustering 문제의 경우 LP의 matrix diagonalization을 시도한 McCormick 등[8]의 연구가 주목할만하다. 이들은 열과 행들의 'bond energy'를 정의하고 이로서 목적함수를 삼았는데 이 bond energy는 matrix의 열과 행들의 상호밀접도를 scalar 값으로 표현함으로써 entropy와 유사한 개념적 기반을 취하고있다. 부품기계그룹 설정 문제는 전형적인 2 차원 개체 clustering 문제로서 대부분의 연구가 목적함수의 제시없이 발견적 기법만을 제공하고 있다. Rajagopalan과 Batra[9]는 개체와 개체의 밀접도를 표현하는 유사계수를 이용하였다. King[7]은 Rank Order Clustering 기법을 제시하였는데 그가 사용한 heuristic은 열과 행을 lexicographically 정렬을 하는 것이다. Chandrasekharan과 Rajagopalan[2]은 cluster 내의 평균밀도와 외부의 평균밀도의 1과의 차를 convex combination한 것을 척도로서 제시하였다. 이들 부품기계그룹 설정문제의 연구들은 방법들간의 비교시 평가기준으로 예외개체수(number of exceptional elements)를 사용하고 있다.

4. 척도의 도출

전술한 바와 같이 clustering 문제의 objective에 대하여 개념적 합의가 이루어져있다.[5] 즉 cluster 내의 동질성과 cluster간의 이질성이 척도라고 볼 수 있다. 이들은 다음의 성질을 갖고 있다.

성질 1. order invariant : cluster나 개체들의 순서가 변경되어도 척도는 동일한 값을 제공해야한다.

성질 2. bounded : cluster나 개체들의 속성이 동일하면 동질성은 최대, 이질성은 최소가 되며 반대의 경우에는 동질성은 최소,이질성은 최대가 된다. 따라서 척도는 bounded되어있다.

성질 3. compatibility with entropy : 주어진 개체간 관계들의 불규칙성을 entropy로 표현하면 이질성은 entropy와 compatible해야하며 동질성은 그 반대이어야한다.

상기한 성질을 반영하기 위하여 entropy 개념을 파악하여보자. 물리학적인 의미에서의 entropy는 상태전이에 소모되는 일(통상 열량)로 측정된다. 이와 유사하게 개체간 관계에서의 entropy는 한 부분관계공간에서 다른 부분관계공간으로 전환시켜주기 위하여 소요되는 일의 양으로 정의될 수 있다. 본 연구에서는 부분관계공간 사이의 기대변화수(expected number of changes)를 일의 양을 나타내는 척도로 제시한다.

4.1 동일차원 개체들간의 동질성

특정 cluster내의 개체들은 N 차원으로 분류된다. 이들 개체들은 동일차원의 개체들간에서 비교가 가능하다. 개체들의 속성을 표현하고 있는 i 차원의 k 개체에 해당하는 부분관계공간들을 고려하여보자. 즉,

$$S_k^i = \{ r \mid r \in E^1 \times E^2 \times \cdots \times E^{i-1} \times e_k \times E^{i+1} \times \cdots \times E^N \} \qquad (식-4)$$

이때 i 차원 개체들의 속성들의 대표값을 갖는 centroid를 설정할 수 있다.

$$Cent^i = (a \mid a \in S.^i) \qquad (식-5)$$

이 centroid는 cluster의 i 차원에서의 속성을 대표하고 있다. 해석적으로는 i 차원의 개체를 임의로 선택하였을 때 각 기타차원의 개체들과 관계를 가질 확률들의 집합이 된다. i 차원의 개체들간의 동질성 hom^i는 다음과 같이 정의된다.

$$hom^i = 1 - (\sum_{r=1}(1-a) + \sum_{r=0} a)/ |C||S.^i| \qquad (식-6)$$
$$0 \le hom^i \le 1$$

즉, 동일차원 개체들간의 동질성은 centroid와 동일할 평균 확률로서 정의되며 이 정의는 상기한 성질들을 만족한다. N 개의 hom^i가 산출된 후 이들의 평균을 구하면 cluster 내의 동일차원 개체간의 동질성을 나타내게 된다. 즉,

$$hom = \sum hom^i/N \qquad (식-7)$$
$$0 \le hom \le 1.$$

hom은 cluster 내에서 임의의 차원의 개체를 선택하였을때 해당 차원의 centroid와 동일할 확률이며 cluster 내의 동일차원 개체들간의 동질성을 표현한다.

2 차원 관계공간의 예를 살펴보자.(그림-1)

	1	2	3	4	5
1	1	1	0	0	0
2	1	1	0	0	0
3	0	0	1	1	1
4	0	1	1	1	0
5	0	0	0	1	1
6	1	0	1	1	1

cluster 해
{1,2}×{1,2}
{3,4,5}×{3,4,5,6}

(그림-1) 2 차원의 예

편의상 열을 1 차원이라 놓으면 {1,2}×{1,2} cluster의 1 차원 개체들의 centroid는 $(1,1,0,1/2,0,1/2)^t$이다. 2 차원 centroid는 $(1,1,0,0,0)^t$이다. hom은

$$hom^1 = 1- (0+0+0+1/2+0+1/2)/26$$
$$= 11/12$$
$$hom^2 = 1 - (0+0+0+0+0)/25$$
$$= 1$$
$$hom = 23/24$$

로 주어진다. 이로써 1 차원 개체와 2 차원 개체들은 높은 동질성을 보이고 있음을 알 수 있으며 특히 2 차원의 경우 개체들의 속성이 동일함(hom=1)을 알 수 있다.

4.2 차원간 동질성

hom은 동일차원에서의 동질성은 반영하지만 차원간의 관계에 대한 고려가 되어있지 않다. 예를 들어 보자.(그림-2)

	1	2	3	4	5
1	0	0	1	1	1
2	0	0	1	1	1
3	1	1	0	0	0
4	1	1	0	0	0

cluster 해
{1,2}×{1,2}
{3,4,5}×{3,4}

(그림-2)

hom은 두 cluster 모두 1의 값을 보여 동질성이 최대상태임을 알 수 있다. 그러나 이 해는 1 차원 개체들과 2 차원 개체들간에 아무 관계가 존재하지 않으므로 좋은 해라 할 수 없다. 따라서 차원들 간의 상호관계가 동질성의 설정에 있어 반영되어야 한다.

동일 cluster 내의 차원간 관련도를 보여주는 척도로는 cluster의 밀도가 합리적이다. 밀도가 1 이면 해당 cluster의 어떤 개체도 다른 차원의 모든 개체들과 관계가 있으며 밀도가 0 이면 해당 cluster의 어떤 개체도 다른 차원의 개체들과 관계를 가지고 있지 않다.

cluster내의 동질성은 상기한 hom과 밀도를 종합한 형태로 다음과 같이 제시된다.

$$HOM = hom \cdot density$$

4.3 이질성

cluster 간의 이질성은 entropy와 compatibile한 개념으로 생각할 수 있다. 동질성과 마찬가지로 각 cluster의 각 차원에서의 centroid를 구한다. 이들 centroid의 centroid를 구한다. 이를 overall centroid라 부르자. overall centroid는 임의의 cluster에서 임의의 개체를 선정하였을 때의 기대 centroid이다.

이질성은 다음과 같이 정의된다.

$$Ocent^i = \sum Cent^i/|G| = (o_1, o_2, \ldots, o_S)^t \qquad (식-9)$$
$$het^i = \sum \{ o_k \cdot (1-a_k) + (1-o_k) \cdot a_k \} / |G||S^i|$$
$$HET = \sum het^i/N$$

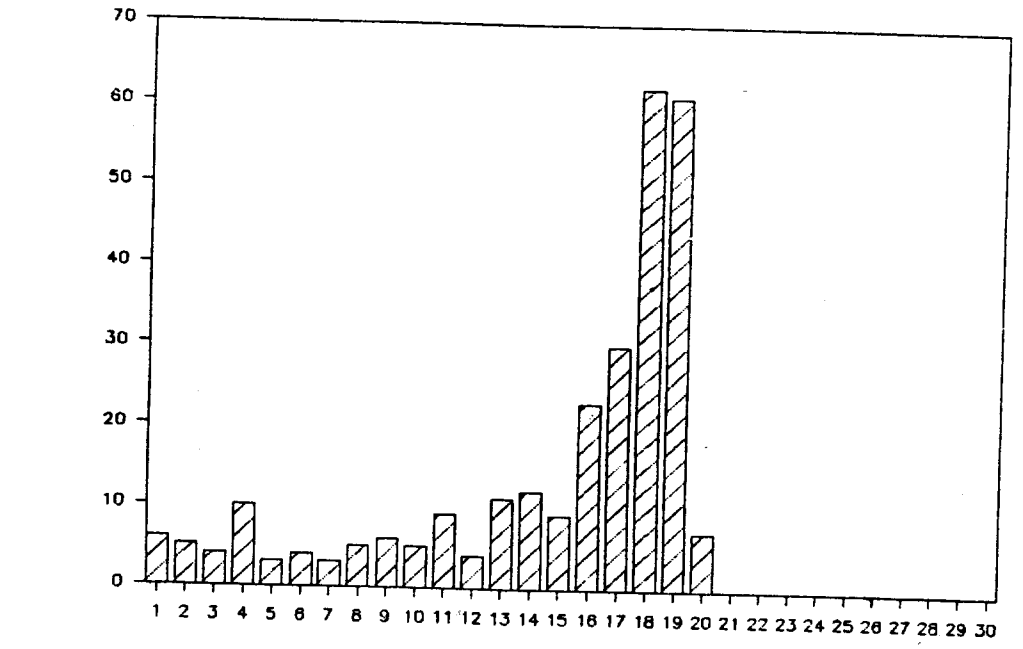
(그림-1)에서의 이질성은 다음과 같다.

$$Ocent^1 = (1/2, 1/2, 1/2, 7/12, 1/3, 3/4)^t$$
$$Ocnet^2 = (5/8, 5/8, 3/8, 1/2, 3/8)^t$$
$$het = 11/24,\ het = 23/40,\ HET = 31/60$$

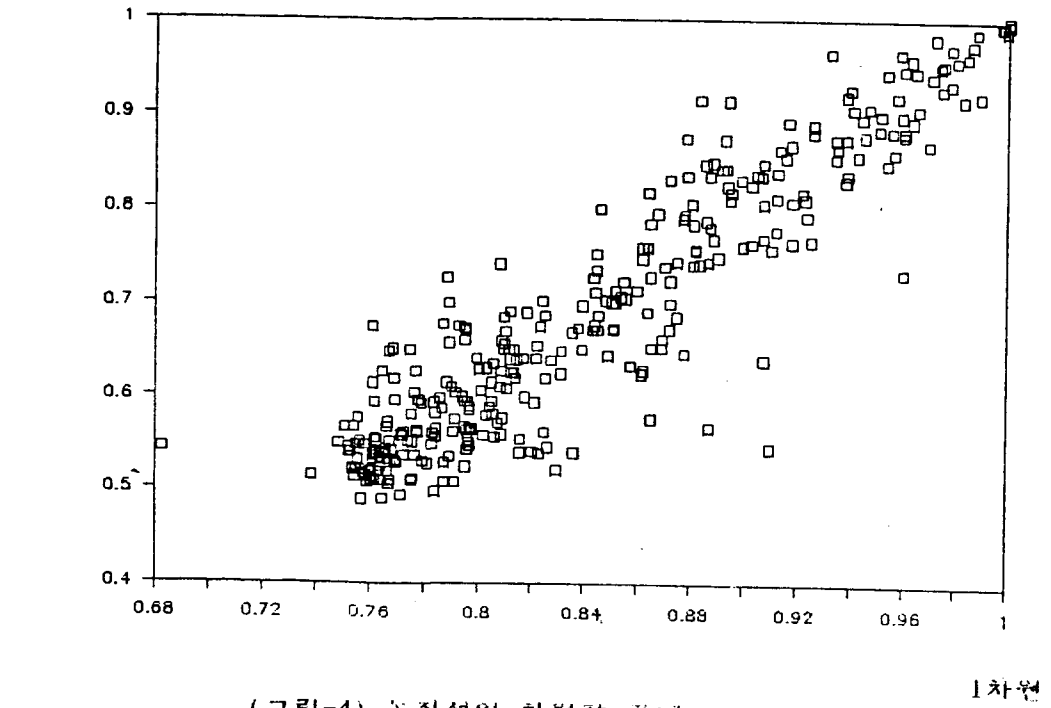
4.4 척도

상기한 동질성 및 이질성 척도는 동일 단위이기는 하지만 동일차원상에 있다고 보기는 어렵다. 그러나 양 척도 모두 maximize의 형태를 취하고 있으므로 다음과 같이 convex combination으로 종합척도를 제시할 수 있다.

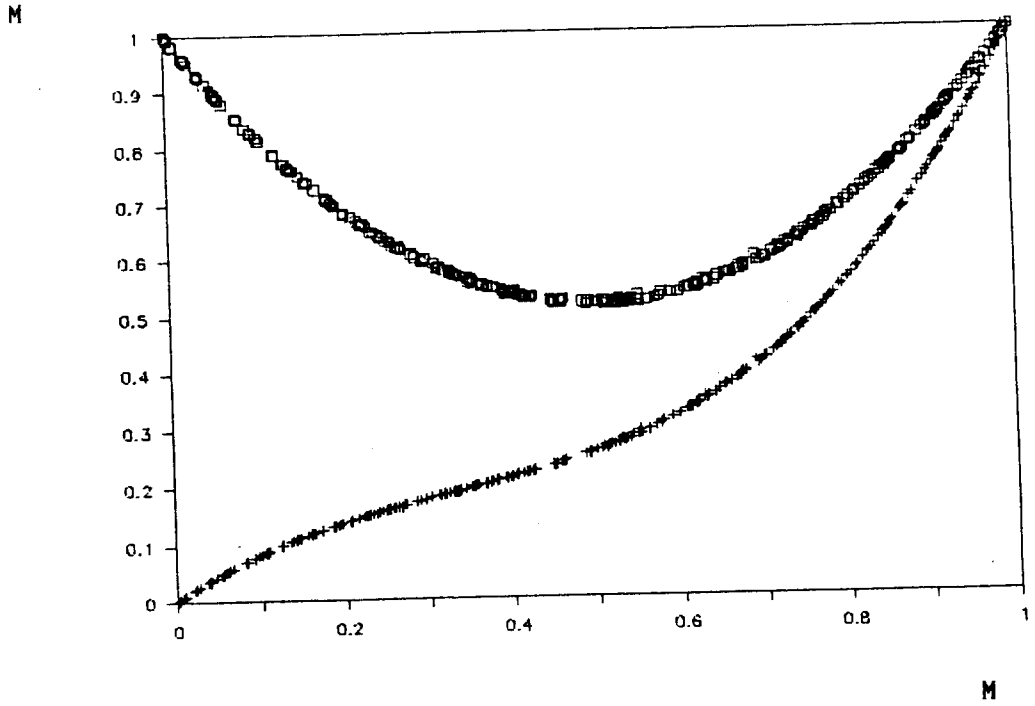
$$M = w \cdot HOM + (1-w) \cdot HET \qquad (식-10)$$



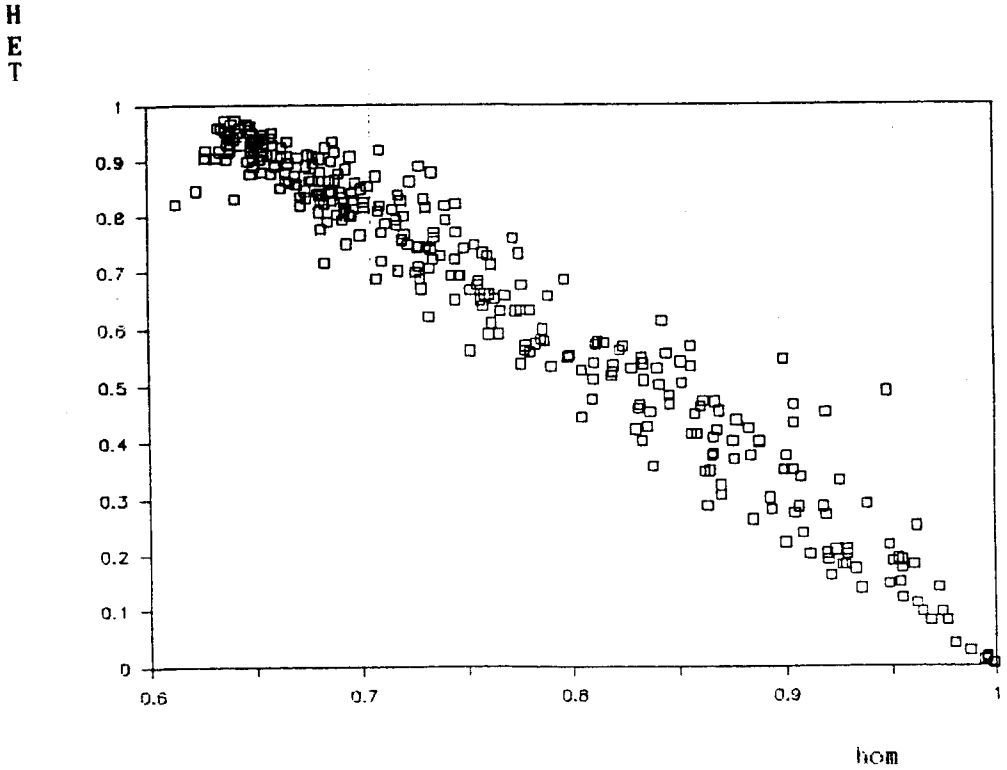
(그림-3) 종합척도의 발생분포



(그림-4) 응집성의 차원간 관계

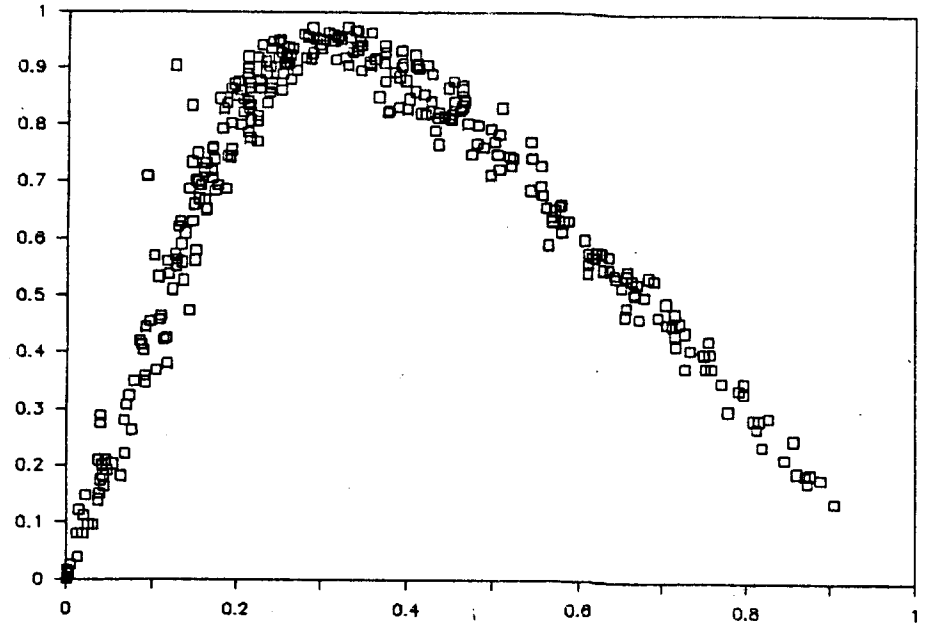


(그림-5) 종합척도와 동질성의 관계

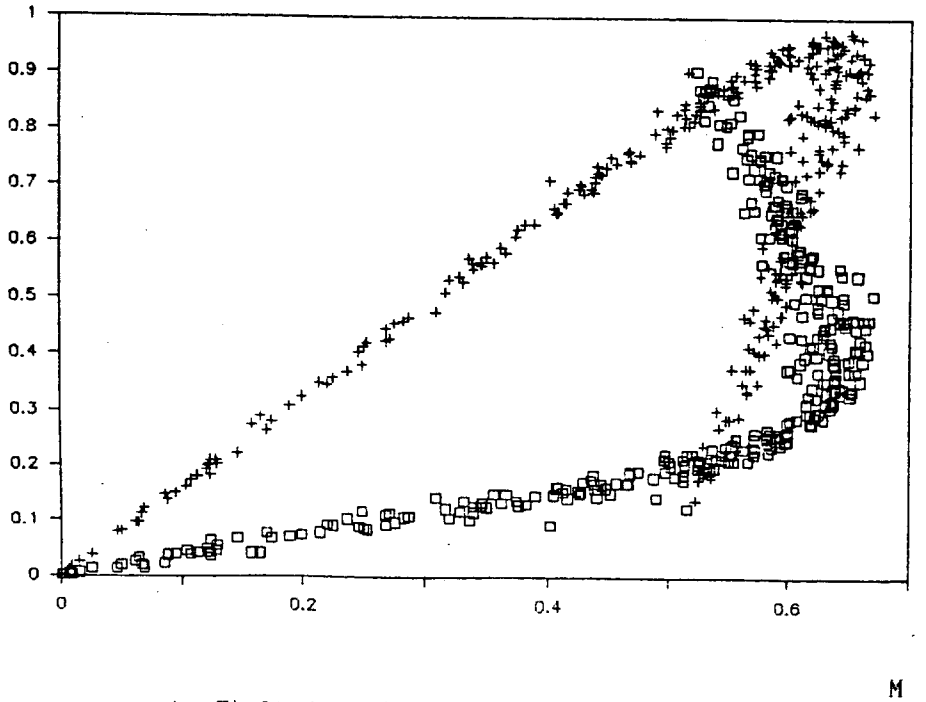


(그림-6) 평균동질성과 이질성의 관계

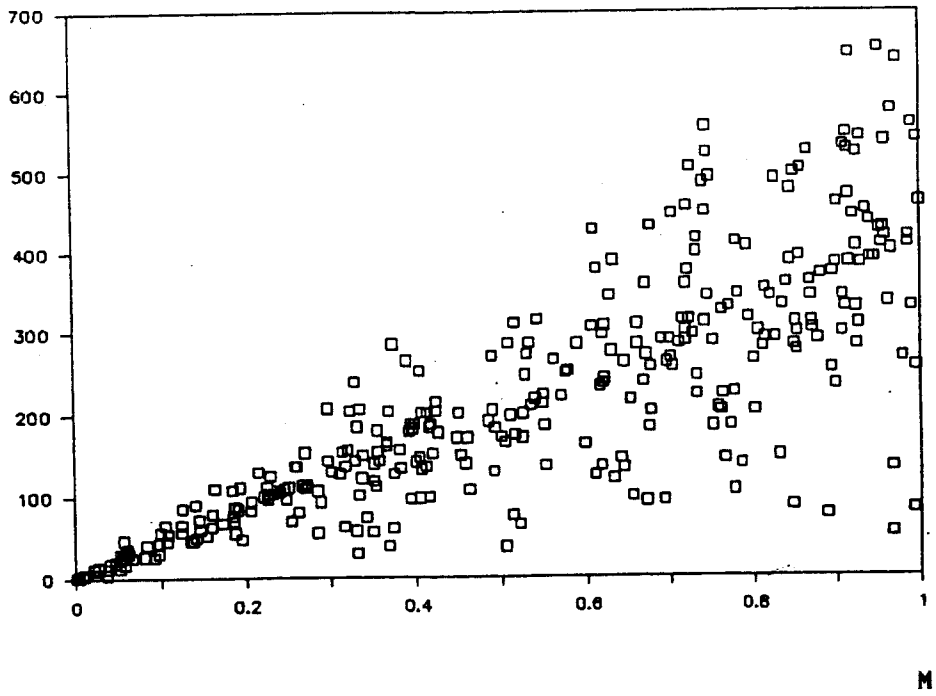
H
E
T



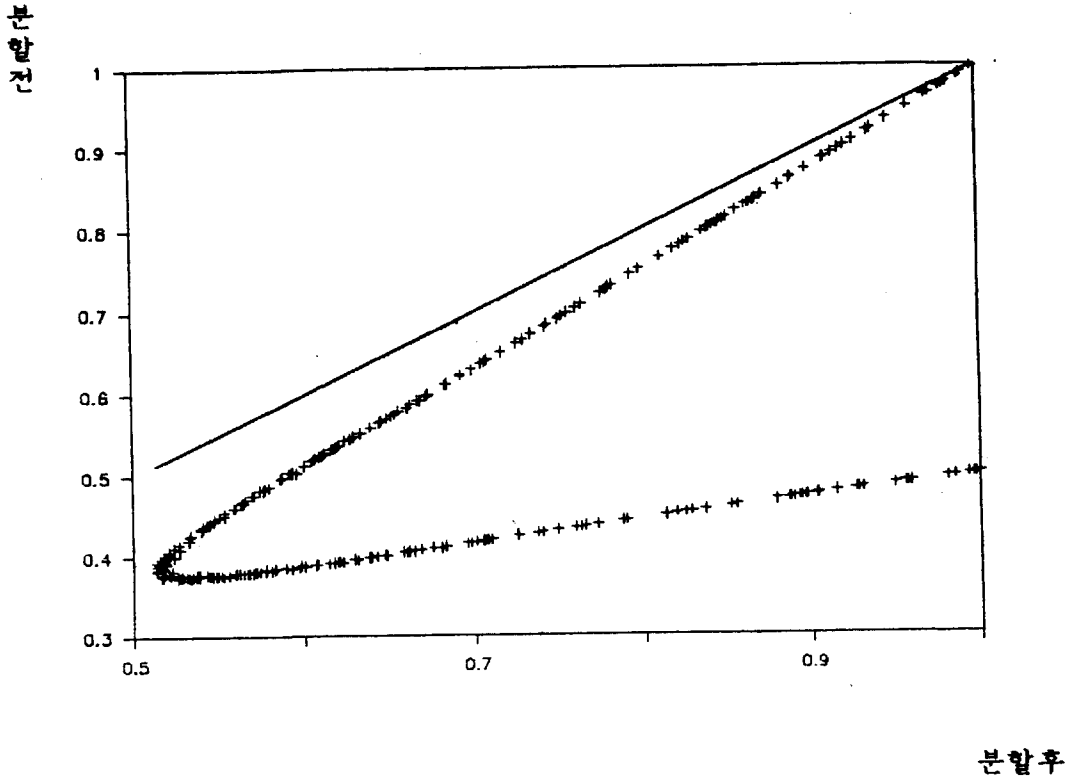
(그림-7) 동질성과 이질성의 관계



(그림-8) 동질성,이질성과 종합적도와의 관계



(그림-9) 종합척도와 밀도척도



(그림-10) 분할전후의 동질성

5. 척도의 분석

제시된 척도의 분석을 위해 2 차원 개체의 경우에 대한 simulation 을 수행하였다. simulation 내용은 30×30의 행렬을 임의로 발생시키고 행렬의 밀도, bond energy, 동질성 등을 측정하였다. 다시 임의의 cluster 해를 발생시켜 1 차원 동질성, 2 차원 동질성, 동질성, 이질성, Chandrasekharan과 Rajagopalan의 평균밀도 척도값, 예외개체수, 종합척도값 등을 측정하였다.

328회의 simulation 결과 [0,1] 구간을 30 등분한 구간에서 종합척도의 발생빈도는 (그림-3)과 같았다. uniform distribution을 보이는 것이 이상적이지만 본 연구에서 제시한 척도는 특정 구간에 밀집된 현상을 보이고 있는데 이에 대한 보다 자세한 조사가 요망된다.

1 차원 동질성과 2 차원 동질성의 경우 (그림-4)와 같은 분포를 보였는데 2 차원의 경우 1 차원의 동질성이 높을수록 2 차원의 부분관계공간이 유사해지므로 매우 자연스러운 결과이다.

행렬밀도와 평균동질성, 동질성의 관계를 (그림-5)에 보였다. 밀도가 낮거나 높을수록 부분관계공간들은 유사해지므로 평균동질성은 convex의 형태를 보인다. 반면 밀도가 낮을수록 차원간 관계가 낮아지므로 동질성은 밀도에 따라 증가하는 형태를 보인다.

평균동질성과 이질성간의 관계는 (그림-6)에 보였다. 동질성과 이질성이 상보관계를 보이는 것은 자연스러운 귀결이다. 반면 동질성과 이질성과의 관계는 차원간 동질성의 영향으로 (그림-7)과 같이 concave 형태를 보인다.

종합척도에 대한 동질성과 이질성의 관계를 살펴보면 duality라고 부를만한 모양을 보이고 있다.(그림-8) 이 그림에서 동질성 및 이질성의 각각의 관점에서 볼 때 각각 최고 종합척도를 주는 값은 서로 다르며 최적해는 그들이 만나는 지점에서 찾을 수 있을 것이 기대된다.

제안된 종합척도와 Chandrasekharan과 Rajagopalan이 제시한 밀도척도의 관계는 (그림-9)와 같이 나타나 상당히 관계가 없음을 보인다. 또 종합척도와 예외개체수의 관계 또한 관련이 없어 보인다.(그림-10)

행렬을 분할하기 전의 동질성과 분할한 후의 동질성을 살펴보면 (그림-11)과 같다. 그림의 선은 y = x 를 추가한 것이다. 이 그림에서 분할된 후의 동질성은 분할 전보다 높은 값을 보이는데 이는 분할이 동질성에는 기여하고 이질성에는 부정적 역할을 수행할 것이라는 추측을 가능케 한다.

6. 결 론

본 연구에서는 0-1 다차원 개체 clustering 문제에 있어서의 척도를 제시하였고 2 차원의 경우에 있어서의 분석을 수행하였다. 본 연구에서 제안된 척도는 clustering 기법의 개발 및 평가에 기여할 수 있을 것으로 기대된다. 추후 연구과제로서 보다 정밀한 분석결과를 얻기 위해서는 simulation 수를 증가시킬 필요가 있으며 특히 임의의 특정 행렬하에서의 임의의 cluster 해들을 발생시켜 분석한다면 clustering 문제에 대한 보다 많은 정보를 얻을 수 있을 것이다.

REFERENCES

[1] 이 철, 강 석호,"다차원 개체를 위한 차이등급 CLUSTERING",한국경영과학회지,제14권,제1호,1989.

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

[3] A. W. F. Edward and L. L. Cavalli-Sforza,"A method for cluster analysis",BIOMETRICS,Vol.21,pp.362-375,1965.

[4] W. D. Fisher,"On grouping for maximum homogeneity",ASAJ,December, 1958.
[5] R. E. Jensen,"A dynamic programming algorithm for cluster analysis",Operations Research,Vol.17,No.6,pp.1034-1057,1969.
[6] B. King,"Step-wise clustering procedures",ASAJ,pp.86-101,March,1967.
[7] J.R. King,"Machine-component grouping in production flow analysis : an approach using a rank order clustering algorithm",International Journal of Production Research,Vol.18,No.2,pp.213-232,1980.
[8] W. T. McCormick, Jr., P. J. Schweitzer and T. W. White,"Problem decomposition and data reorganization by a clustering technique",Operations Research,Vol.20,pp.993-1009,1972.
[9] R. Rajagopalan and J. L. Batra,"Design of cellular production systems : A graph-theoretic approach", International Journal of Production Research,Vol.13,No.6,pp.567-579,1975.
[10] M. R. Rao,"Cluster analysis and mathematical programming",JASA,Vol.66,No.335,pp.622-626,September,1971.
[11] H. D. Vinod,"Integer programming and the theory of grouping",ASAJ,pp.506-519,June,1969.