확률적 관계하의 다차원 개체 CLUSTERING
(Multidimensional Entity Clustering with Probabilistic Relationships)

서울대 산업공학과 이 철
서울대 산업공학과 강 석 호

1. 서 론

개체들이 동일 차원상에 있지 않은 다차원 개체 clustering 문제와 0-1 관계하의 해법으로서의 차이등급 clustering 기법이 저자들에 의해 제시된바 있다.[1] 0-1 관계하의 다차원 clustering 문제를 약술하면 다음과 같다.

개체들의 차원이 N일 때 각 차원의 개체들의 집합을 $E^1, E^2, \ldots, E^N$ 이라 하자. 이 때 E^i는 상호배타적(mutually exclusive)이다. 개체간 관계는 다음 식을 만족한다.

$$r(e_1, e_2, \ldots, e_N) = \begin{cases} 1, & \text{if } e_1, e_2, \ldots, e_N \text{ have a relation} \\ 0, & \text{otherwise} \end{cases} \quad \text{(식-1)}$$
$$\text{where } e_i \in E^i \text{ and } e_i \neq \emptyset.$$

0-1 관계하의 다차원 clustering 문제는 주어진 r 집합에 의거하여 cluster 해 $\mathcal{C} = \{ C_1, C_2, \ldots, C_K \}$ 를 C_i 내의 homogeneity와 C_i 간의 heterogeneity를 향상시키도록 구성하는 것이다. 단 $\mathcal{C}$ 는 다음의 제약식을 만족하여야 한다.

$$C_i = C_i^1 \times C_i^2 \times \ldots \times C_i^N$$
$$\text{where } \forall j,\ C_i^j \subseteq \mathcal{P}(E^j) \text{ and } C_i^j \neq \emptyset$$
$$\forall i,\ \bigcup C_i^j = E^j \quad \text{(식-2)}$$

이러한 0-1 관계하의 다차원 clustering 문제를 위한 차이등급 clustering 기법은 다음의 2 가지 제약을 가지고 있다.

제약 1. 개체간 관계가 0-1 으로 완전한 정보상황을 가정한다.
제약 2. 개체간 관계의 해에 대한 영향도를 동일한 것으로 가정한다.

다차원 개체 clustering 의 응용분야라고 할 수 있는 분산정보시스템의 분석(Analysis of Distributed Information Systems)이나 기계-부품 그룹형성 문제(Machine Part Group Formation) 등의 경우 발생빈도를 가중한 정보량, 또는 부품의 특정 기계에서의 가공시간 등의 가중치를 부여해야 할 필요성이 발생한다. 또한 개체간 관계에 대한 불확실한 정보만이 제공될 경우 차이등급 clustering 기법의 적용은 불가능하다.

본 연구의 목적은 개체간 관계가 확률로서 주어지고 영향도에 따른 가중치가 있는 다차원 개체 clustering 문제의 해법 개발에 있다.

2. 용어 및 부호

본 연구에서 편의 상 사용된 부호 및 용어는 다음과 같다.

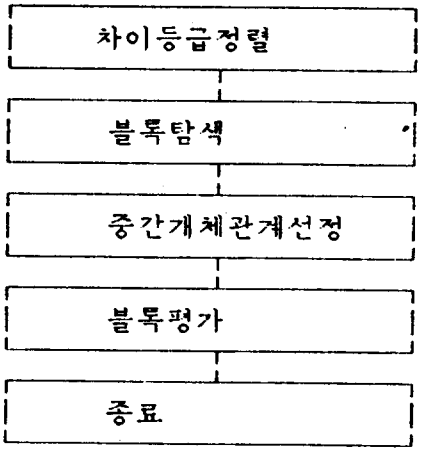
N : 개체들의 차원

E^i : i 차원 개체들의 집합
$|E^i|$: i 차원 개체들의 cardinality
e_i : i 차원 개체
$e_i[k]$: i 차원 상에서의 순서가 k 번째인 개체
개체관계 $r(e_1,e_2,\ldots,e_N)$: $e_1,e_2,\ldots,e_N$ 개체들 간의 관계
$w(r(e_1,e_2,\ldots,e_N))$: 개체간 관계 r의 가중치
관계공간 R : r의 집합
부분관계공간 R_j^k : k 차원이 j 개체인 개체관계들의 집합
$r_j^k\ (e_1,e_2,\ldots,e_{k-1},e_{k+1},\ldots,e_N)$: 부분관계공간 R 의 k 차원을 배제한 개체관계
기준 개체관계 r* : 특정 cluster 의 속성을 대표하는 개체관계
기준 부분관계공간 R_j^{k*} : j 가 r*의 개체와 동일한 부분관계공간
블록 : 예비 cluster (a candidate for the cluster)
예외 개체관계 (Exceptional Elements) : $\{ r \mid \forall C_i,\ r \notin C_i \}$
통합기준 : 복수의 블록을 통합 혹은 분리하는 기준

3. 차이등급 CLUSTERING 기법의 확장

3.1 차이등급 clustering

차이등급 clustering은 차이등급정렬, 블록탐색, 중간개체관계선정, 블록평가의 4 부분으로 이루어져 있다.(그림-1)



(그림-1) 차이등급 CLUSTERING 기법

3.1.1. 차이등급정렬 (Difference Ordering)

두 부분관계공간의 차이 부분관계공간은 다음과 같이 정의된다.

$$D(R_i^k,R_j^k) = \{ d(e_1,e_2,\ldots,e_{k-1},e_{k+1},\ldots,e_N) \}$$
$$\text{where } d = 1 \text{ if } r(e_1,e_2,\ldots,e_{k-1},e_{k+1},\ldots,e_N) = r(e_1,e_2,\ldots,e_{k-1},e_{k+1},\ldots,e_N)$$
$$0 \text{ , otherwise} \qquad \text{(식-3)}$$

기준 개체관계 r*가 확정되었을 때 D의 차이등급 vector $V_j = (v_1,v_2,\ldots,v_m)^t$는 다음과 같다.

$$v_n = \sum_{\sum i_k = N+n-1} d(e_1[i_1],e_2[i_2],\ldots,e_N[i_N]) \qquad \text{(식-4)}$$

차이등급정렬은 어느 개체차원에서도 $r_{[1]} \prec r_{[2]} \prec \ldots \prec r_{[N]}$이 만족되도록 개체들의 순서를 결정하는 것이다. 이때 $\prec$는 'lexicographically not greater than'을 의미한다.

3.1.2. 블록탐색

블록탐색은 차이등급정렬이 수행된 후 기준 개체관계로부터 시작하여 통합기준에 근거하여 해당 기준 개체간 관계에 적정한 블록을 결정하는 절차이다.

step 1. $B_0 \Leftarrow r(e_1[1], e_2[1], \ldots, e_N[1])$
$E_0 \Leftarrow e_1[1], e_2[1], \ldots, e_N[1]$
step 2. while $r(e_1[k], e_2[k], \ldots, e_N[k]) = 1$
$E_0 \Leftarrow e_1[k], e_2[k], \ldots, e_N[k]$
$B_0 \Leftarrow r$, where r's entities belong to E_0
$k = k + 1$
step 3. $B_r \Leftarrow r$, where $\forall e_i \in E_0$
step 4. while there exists e_i which does not increase r's that has value 1 and does not belongs to $B_0 \cup B_r$
$E_0 \Leftarrow e_i$
$B_0 \Leftarrow r$ i.e., $e_i \in E_0$
$B_r \Leftarrow r$ i.e., $e_i \notin E_0$

3.1.3. 중간개체관계선정

임의의 개체간 관계를 기준개체관계로 선정한 경우 병목개체(Bottle-neck Entity)가 선정될 위험이 있다. 이를 회피하기 위하여 다음과 같은 중간개체관계선정 절차를 수행한다.

step 1. find new r* i.e., $e_i[\lceil |e/2| \rceil]$'s
step 2. if r* = 0 , find the closest r
step 3. difference ordering
step 4. compare the blocks with following criteria
first criteria : high density
second criteria : larger number of entities
third criteria : less exceptional elements

3.1.4. 블록 평가

새로운 블록이 중간 개체관계선정 절차에서 결정되면 기존의 블록들과의 관계를 분석하여 병목개체를 파악한다. 새로운 블록을 B , 기존의 블록들의 집합을 $\mathcal{C} = \{ C_1, C_2, \ldots, C_K \}$ 라 하자.

case 1. $B \cap C_i = \emptyset$ for all $C_i \in \mathcal{C}$,
$\mathcal{C} \Leftarrow B$
case 2. $B \supseteq C_i$ for some C_i's ,
$\mathcal{C} \Leftarrow B \backslash \cup C_i$ and $B \backslash \cup C_i$ is bottle-neck entities
if number of C_i's > 1 or $\forall i$, $|C_i| > |B|$
$\mathcal{C} = (\mathcal{C} \backslash C_i \Leftarrow B)$, otherwise
case 3. $B \cap C_i \neq \emptyset$ for some C_i's ,
$C_i \Leftarrow C_i \backslash B$ for all C_i and $\mathcal{C}$
$\mathcal{C} \Leftarrow B \backslash \cup C_i$
$B \backslash \cup C_i$ is bottle-neck entities

3.1.5. Algorithm 과 Complexity

algorithm : 총괄적인 algorithm은 다음과 같다.

step 1. 임의의 기준 개체관계 선정
step 2. 차이등급 정렬
step 3. 블록탐색
step 4. 중간개체관계선정
setp 5. 블록평가.
step 6. 모든 개체가 cluster에 포함되면 terminate, 아니면 포함되지 않은 개체간 관계 중에서 가장 순서가 먼 1의 값의 개체간 관계를 기준 개체관계로 선정

complexity : R 전체를 포괄하는 차이등급 vector V를 상정하여보자. 한차원에서 차이등급정렬을 할 때마다 최소한 lexicographically $(0,0,\ldots,0,1)^t$ 만큼 감소되므로 최소치에 다다를 때 정렬은 완성된다. 따라서 complexity는 $O(\sum|E^i|\cdot|R|\cdot$ sorting $)$로서 polynomial이다.

3.2. 확률적 관계의 도입

3.2.1. 문제 및 해법의 성격

차이등급 clustering 기법은 개체간 관계가 0-1 인 경우만을 상정하고 있으므로 확률적 관계 및 가중치 부여의 경우 적용이 불가능하다. 그러나 기본 idea인 차이 부분관계공간 및 차이등급 vector의 확장이 가능하다. 따라서 본 연구에서는 차이등급 clustering 기법을 모체로 하여 각 절차에서의 확장된 방법을 제시하도록 하겠다.

3.2.2. 기대차이등급정렬

확율로서 개체간 관계가 표현된 경우 개체간 관계 r 은 다음과 같이 표현된다.

$$r(e_1,e_2,\ldots,e_N) \in [0,1] \qquad \text{(식-5)}$$

두 부분관계공간 R_i^k, R_j^k의 차이를 표현하는 차이부분관계공간 D 는 다음 식과 같이 나타난다.

$$D = \{\, d(e_1,e_2,\ldots,e_{k-1},e_{k+1},\ldots,e_N) \,\}$$
$$\text{where } d = r_i^k\,(1-r_j^k) + (1-r_i^k)\,r_j^k \qquad \text{(식-6)}$$

이때 d 값은 두 부분관계공간이 차이를 보일 확률이다. 가중치가 부여된 경우 차이 부분관계공간은 기대 가중치가 된다.

$$d = w(r_i^k)\,r_i^k\,(1-r_j^k) + w(r_j^k)\,(1-r_i^k)\,r_j^k \qquad \text{(식-7)}$$

가중치가 부여되지 않은 경우는 $w = 1$ 의 특별한 경우로 간주된다. 이 경우의 차이 부분관계공간을 기대차이 부분관계공간이라 부르자. 기대차이 부분관계공간에 따른 기대차이 등급 vector $V_j = (v_1,v_2,\ldots,v_m)^t$는 다음과 같다.

$$v_i = \sum_{\sum k_n = N+i-1} d(e_1[k_1],\ldots,e_N[k_N]) \qquad \text{(식-8)}$$

이에 따라 각 차원에서 lexicographically decreasing order로 개체들의 순서를 결정한다.

3.2.3. 블록탐색

블록탐색의 근거로 사용된 통합기준은 예외개체의 최소화였다. 여기서는 예외개체들의 기대 가중치를 사용한다. 블록 B와 B에 포함되지않는 개체들로만 이루어진 개체간 관계들의 집합 R이 주어졌을 때의 예외개체들의 기대가중치의 합을 $\mathcal{N}(B,R)$이라 하자. 이때 블록탐색 절차는 다음과 같다

step 1. $B \Leftarrow e_1[1], e_2[1], \ldots, e_N[1]$
$R \Leftarrow E^1 \backslash e_1[1], E^2 \backslash e_2[1], \ldots, E^N \backslash e_N[1]$
step 2. if $\mathcal{N}(B,R)$ does not increase for next e_i in any i
$B \Leftarrow e_i$
$R = R \backslash e_i$
otherwise terminate

3.2.4. 중간개체관계선정

criteria 중 density 외에는 그대로 적용된다. density의 경우 기대치를 사용한 수 있다. 즉, $\sum w(r)/|B|$이며 $r \in B$ 이다.

3.2.5. 블록평가

블록평가의 경우 병목개체의 집합에 확률이 부여된다는 점 외에는 차이등급 clustering 기법과 동일하다. 이 확률은 당해 개체가 병목개체로서의 역할을 수행하게 될 정도를 나타낸다.

3.2.6. Algorithm 과 Complexity

차이등급 clustering 기법과 동일하다.

4. 결론

다차원 개체 clustering 문제에 있어서 개체간 관계가 확률적이고 가중치가 부여된 경우를 위한 기대차이등급 clustering 기법을 제시하였다. 기대차이등급 clustering 기법은 해법의 필요성에 비해 상대적으로 해법개발이 미진한 분산정보시스템을 대상으로 한 전산화 master plan의 수립이나 기계-부품 그룹형성, FMS에서의 주문선정(Part Type Selection) 등에 기여할 수 있을 것으로 기대된다. 반면 해법의 타당성 검토를 위한 이론적 연구가 제시되지 않아 추후 이의 보완을 위한 연구가 요망된다.

REFERENCES

1. 이 철, 강 석호, "다차원 개체를 위한 차이등급 CLUSTERING", 한국경영과학학회지, 제 14권, 제 1호, 1989.