A Genetic Algorithm for Clustering in Data Mining

정지원*, 최인찬
고려대학교 산업공학과

Abstract

본 논문에서는 데이터마이닝 문제에 클러스터링 기법을 적용할 때 발생할 수 있는 문제점 및 속성선택(feature selection)과 그룹 수 산정의 상호연관성을 살펴보고, 데이터 필드의 상대적 중요도와 최적의 그룹 수를 결정하는 수리적 모형을 제시한다. 또한, 이 모형을 풀기 위하여 K-means 알고리즘을 이용한 유전 알고리즘을 제시한다.

1. 서 론

데이터마이닝은 축적된 데이터를 효과적으로 활용하기 위하여 데이터에 내재된 규칙, 패턴 등을 찾아서 적용하는 과정이다. 많은 양의 데이터로부터 효과적으로 지식을 추출하기 위해 가설검증, 지식추론과 관련된 자동화, 반자동화된 기법들이 데이터마이닝에 적용된다. 이러한 기법들 가운데 클러스터링은 서로 유사한 성향을 가진 몇 개의 동질적인 클러스터(그룹)로 데이터를 구분하여 노이즈를 제거하고 그룹간의 관계로부터 또는 각 그룹으로부터 쉽게 필요한 정보를 추출할 수 있도록 하는 기법이다. 클러스터링은 미리 분류된 데이터나 특정 목적 필드를 필요로 하지 않는 지식추론 기법(undirected knowledge discovery)의 하나로서 데이터마이닝의 시작 단계에 적합한 기법이며 그룹을 형성한 후 필요한 규칙이나 패턴을 찾아내기 위해 몇 가지 기법이 함께 적용되기도 한다[1].

클러스터링 문제를 풀기 위한 기존 알고리즘은 크게 K-means를 포함한 비계층적인 접근방법과 최적화 모형에 의한 접근방법, 계층적 접근법 및 인공지능에 기초한 접근방법으로 구분된다. 이들 기존의 알고리즘은 NP-complete 영역에 속하는 클러스터링 문제를 효율적으로 풀어내기 위해 각 속성(feature)에 대한 가중치나 클러스터의 수가 알려진 것으로 가정하거나 주어진 데이터 중 하나를 클러스터의 중심으로 하여 데이터를 할당하고 있다. 즉, 클러스터링 문제를 풀기 위하여 1)생성될 클러스터 수(number of clusters)를 결정하는 문제, 2)클러스터를 구분하는데 불필요한 속성들을 제거하고 기여도에 따라 속성에 대한 가중치를 할당하는 가중치 산정 또는 속성 선택(feature selection) 문제, 3) 데이터를 각각의 클러스터에 할당하는 클러스터링의 세 가지 부문제(sub-problem)에 대한 별도의 절차(procedure)를 필요로 한다.

이들 부문제는 상호 밀접한 연관관계를 가지고 있어서 독립적으로 분리하여 고려하기보다는 통합하여 고려하는 것이 바람직하다. 또한, 데이터에 대한 충분한 사전지식이 없는 일반 사용자는 물론 전문가라 할지라도 반복적으로 시행착오를 거쳐 각 부문제에 대한 해를 구하게 되므로 많은 시간이 소요된다. 특히, 클러스터링이 미지의 데이터에 대하여 마이닝의 초기 단계에 주로 적용되는 점을 감안하면 이들 부문제의 연관관계를 고려한 객관적이고 통합적인 방안이 요구된다. 본 논문에서는 이들 각 부문제를 통합적으로 고려한 수리적 모형을 제시하고 이 모형을 풀기 위한 K-means 기법에 기초한 유전 알고리즘을 제시한다.

2. 기존 연구와 문제점

2.1 기존 연구

클러스터링 문제를 풀기 위해 여러 알고리즘들이 연구되어 있지만 알고리즘의 평가에 대한 일치된 기준은 제시되어 있지 않다. 그러나 Cowgil 등[2]에 의하면 많은 클러스터링 알고리즘이 개념적으로는 그룹내 객체들의 집적도와 그룹간의 객체들의 이산도를 최적화 하는 것을 클러스터링의 우선적인 평가기준으로 다루고 있다. 그럼에도 많은 알고리즘이 그룹간의 거리는 배제하고 그룹내 유사성(similarity) 또는 비유사성(dissimilarity)만의 합 또는 분산을 최대화 또는 최소화하는 것을 목적으로 하고 있다.

비계층적 클러스터링 기법의 하나인 K-means 알고리즘은 그룹내 가상의 중심(centroid)으로부터 각 데이터에 대한 거리의 제곱의 합을 최소화하도록 K개의 그룹을 형성한다. Mulvey와 Crowder[7]에 의한 p-median 모델에서는 각 그룹에서 주어진 데이터 중의 하나를 그 그룹의 중심으로 하여 이들 중심으로부터 그룹내 각 데이터의 거리의 합을 최소화하고 있으며 Wang[10]의 선형 할당(linear assignment) 모델은 각 데이터의 쌍으로부터 유사성을 계산하고 이로부터 데이터 중에서 그룹의 중심을 먼저 구한 후 주어진 중심 데이터로부터 유사성의 합을 최대화하는 모형을 제시하고 있다. 그룹의 중심을 데이터들 중에서 선택할 경우 그룹이 비교적 분산되어 있고 데이터의 수가 적은 경우에는 실제 중심과 큰 차이가 있을 수 있는 문제점이 있다. Mangasarian[5]은 주어진 K 값에 대하여 2-norm 대신 1-norm 거리에 근거한 bilinear 모델을 제시하고 있다. 1-norm을 사용하는 경우 상대적으로 동떨어진 데이터의 효과를 가중시키는 결과를 보이지만 그룹 수가 미리 정해져 있지 않은 경우에는 이러한 가중효과가 불필요하다.

Cowgill 등[2]은 그룹내 집적도와 그룹간 이산도를 동시에 고려하고 있으나 그룹 수가 주어진 상태에서 이들을 최대화하고 있어 그룹내 집적도와 그룹간 이산도를 함께 고려한 장점을 충분히 살리지 못하고 있다. 이와 같이 대부분의 최적화 모형에 기반을 둔 접근 기법과 비계층적 기법은 클러스

터링 알고리즘에 의하여 생성될 클러스터의 수를 사용자가 미리 결정하도록 하고 있다. 반면에 계층적 접근 기법들은 n개의 데이터에 대하여 1개의 클러스터로부터 n개까지 모든 범위의 클러스터링 해를 제공하고 사용자가 결정하도록 하고 있다.

2.2 문제점

클러스터링에서 속성선택은 최소한의 속성만으로 데이터를 단순화시켜 클러스터링하는 것을 목적으로 한다[5]. 즉, 중요도가 적은 속성들은 고려대상에서 제외시켜 문제의 차원(dimensionality)을 감소시키고 복잡도를 줄일 수 있다. 미지의 데이터를 대상으로 하는 데이터마이닝에서 속성선택은 특히 중요한 의미를 갖는다. 유사한 데이터들로 이루어진 그룹으로부터 유용한 정보를 추출하기 위하여 다른 기법들이 추가적으로 적용되기도 하지만 속성선택 기법을 적용했을 때 선택된 속성들의 가중치로부터 속성별 중요도 및 연관성 또는 보다 의미 있는 정보를 쉽게 추출할 수도 있다.

아래 그림 1은 그룹 수에 따른 속성 가중치의 중요도에서의 변화를 보이고 있다. 그림 1의 데이터를 점선 직사각형 안의 2 개의 클러스터로 구분하는 경우 성별 항목은 구분자(discriminator)로서 의미가 없게되며, 점선 좌표에 따라 4 개의 클러스터로 구분하는 경우에는 구분자로서 중요도가 커지게 된다.

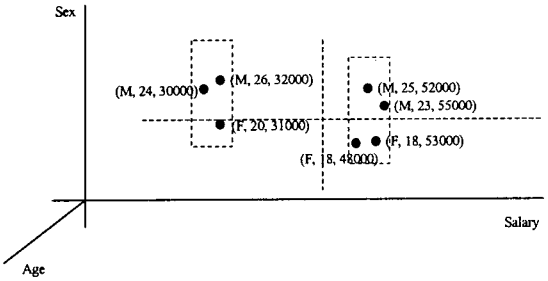

그림 1. 그룹 수에 따른 가중치 변화

클러스터링 문제에 대한 대부분의 기존 연구는 그룹의 수 즉, K 값의 산정과 속성선택을 별개의 문제로 다루고 있다. 즉, 대부분의 속성선택을 위한 기법이 클러스터 수가 미리 결정된 것으로 가정하고 있다. Sarkar 등[9]은 evolutionary programming 기반의 알고리즘에서 그룹내 거리와 그룹간 거리를 동시에 고려하여 그룹 수의 결정과 클러스터링을 최적화하고 있으나 속성선택 문제는 고려하지 않고 있다. Punch 등[8]은 그룹의 수와는 무관한 분류(classification) 문제에서 데이터를 주어진 그룹에 할당하기 위해 각 속성에 적절한 가중치를 부여하는 속성선택 기법을 사용하고 있다.

본 논문에서는 Punch 등[8]의 기법에 근거한 속성선택 기법을 적용함으로써 문제의 차원(dimensionality)을 감소시키고 그룹내의 거리와 그룹간의 거리를 동시에 고려한 최적의 클러스터링을 위하여 속성들에 적절한 가중치를 부여할 수 있는 모형을 제시한다. 본 문에서 적용된 속성선택 기법은 그룹내의 거리를 최소화하고 그룹간의 거리를 최대화하는 것을 목적식으로 하여 각 속성이 클러스터링의 구분자로서 갖는 중요도에 따라 가중치를 부여한다는 점에서 Punch 등[8]의 속성선택 기법과는 차이가 있다.

3. 그룹의 수와 속성 선택을 고려한 모형

아래의 수학적 모형은 그룹내의 거리를 최소화하고 그룹간의 거리를 최대화하는 클러스터 수, 속성 가중치 및 클러스터링을 결정한다.

n개의 데이터를 상호 배타적인 m개의 그룹으로 유클리디안 거리를 이용하여 클러스터링 하는 문제에서 $m \le n$ 이고 각 데이터는 l개의 속성(attributes or variables)을 갖는다고 가정한다. 이때 각 속성의 값은 l 차원의 공간에서 한 축에 대한 거리를 나타내며 각 속성에 대하여 값을 갖는 데이터는 l 차원 공간에서 한 점을 나타내는 벡터로 나타낼 수 있다. 각 속성의 값은 0과 1 사이의 실수로 정규화되어 표현되어 있다고 가정할 때 클러스터링 문제는 다음과 같이 다목적 비선형 정수계획 문제(Multi-objective Non-linear Integer Programming)로 수식화 될 수 있다.

Decision Variables:

x_{ij}: 레코드 j가 클러스터 i에 할당(0,1)

w_k: 필드 k의 가중치($0 \le w_k \le 1$)

y_i: 이진변수, 클러스터 i에 할당된 데이터가 존재할 때 1

v_{ik}: 클러스터 i에 속하는 데이터들의 중심점의 속성 k의 좌표

Coefficients:

d_{jk}: 정규화된 데이터 j의 속성 k의 좌표 값($0 \le d_{jk} \le 1$)

<MOCP: Multi-objective Clustering Problem>

$$\min \quad \left[\sum_{i=1}^{m} \left\{ \sum_{j=1}^{n} \left[\sum_{k=1}^{l} (w_k \cdot d_{jk} \cdot x_{ij} - v_{ik})^2 \right]^{1/2} \right\} / \sum_{j=1}^{n} x_{ij} \right] / \sum_{i=1}^{m} y_i$$

$$\max \quad \sum_{k=1}^{m} \left\{ \sum_{i=1}^{m} \left[\sum_{k=1}^{l} (v_{ik} - v_{hk})^2 \right]^{1/2} \right\} / \left[\left(\sum_{i=1}^{m} y_i - 1 \right) \cdot \sum_{i=1}^{m} y_i \right]$$

$$s.t. \quad \sum_{i=1}^{m} x_{ij} = 1 \qquad \text{for all } j \quad (1)$$

$$x_{ij} \le y_i \qquad \text{for all } i, j \quad (2)$$

$$y_i \le \sum_{j=1}^{n} x_{ij} \qquad \text{for all } i, k \quad (3)$$

$$v_{ik} = \sum_{j=1}^{n} w_k \cdot d_{jk} \cdot x_{ij} \Big/ \sum_{j=1}^{n} x_{ij} \qquad \text{for all } i, k \quad (4)$$

$$v_{ik} \ge 0, \quad 0 \le w_k \le 1,$$

$$x_{ij} \in (0,1), \quad y_i \in (0,1)$$

제약식 (1)은 모든 데이터는 하나의 클러스터에 할당되어야 한다는 것을 의미하고 제약식 (2)와

(3)은 클러스터 i에 할당된 데이터가 존재할 때 y_i가 1의 값을 갖도록 보장한다. 식 (4)는 클러스터 i의 속성 k의 평균값을 계산하는 식으로 클러스터 i의 가상 중심의 속성 k의 좌표를 나타낸다.

4. K-means에 기초한 유전 알고리즘

유전 알고리즘은 자연체계의 유전법칙에 근거한 발전적 탐색기법이다. 유전 알고리즘은 복수개의 점을 동시에 탐색하며 확률개념을 적용함으로서 효과적으로 전역해를 탐색한다. 유전 알고리즘을 적용하기 위해서는 우선, 탐색하고자 하는 파라미터에 대하여 코드화된 스트링을 생성한다. 다수의 스트링으로 구성된 초기해 집단을 만들고 이들에 대하여 자연선택(selection)과 부분적인 스트링의 교환(교체, crossover)이나 단일 유전인자의 변환(돌연변이, mutation)등의 유전 연산자를 적용하여 적합도(fitness value)에서 보다 향상된 다음 세대를 구성하게 된다. 이와 같이 점진적으로 향상된 세대를 반복적으로 생성함으로써 최적해를 탐색하게 된다. 클러스터링 문제에 유전 알고리즘을 적용한 연구들이 이미 이루어져 있으나[2, 4] 본문에서는 MOCP를 위한 K-means 기반의 혼합 유전 알고리즘을 제안한다.

클러스터링 문제에 유전 알고리즘을 적용할 때 가장 일반적인 유전자 구조는 데이터 수만큼의 유전인자를 갖는 스트링 구조에서 각 유전인자의 값(allele)으로 1부터 K까지의 클러스터 수를 갖는 방식이다[4]. 그러나 클러스터의 수가 정해져있지 않은 MOCP에 이러한 유전자 구조를 적용하기는 어렵다. 따라서 본 문에서는 가중치와 그룹의 수로 구성된 유전자 구조를 사용한다. 또한, 이들 가중치와 그룹 수를 이용하여 데이터를 클러스터링하기 위해 K-means 기법을 적용한다. K-means 기법은 지역해에 빠질 수 있는 단점이 있으나 현실적인 문제에서 비교적 우수한 결과를 보이고 있어 가장 널리 사용되는 기법중 하나이다[5].

4.1 유전자 구조

유전자 구조에서 필드의 가중치를 나타내는 유전인자는 0부터 255까지의 0-1 패턴(binary pattern)을 갖게되며 이 값을 10진수로 바꾸어 255로 나눈 값이 각 필드의 가중치가 된다. 그러므로 각 필드의 가중치는 8 비트의 메모리를 차지하게 된다. 속성별 가중치를 나타내는 비트들과 함께 + 또는 - 부호를 나타내는 한 개의 비트와 ±(K × x%) 범위의 정수 값을 나타내는 비트들이 추가된다. 임의의 범위 내의 정수 값으로 표현되는 이 값은 생성될 클러스터 수를 나타낸다. 이러한 범위 값은 사용자에 의해 주어지며 사용자가 임으로 선택한 범위 내에서 K 값을 탐색하게 되므로 모든 클러스터 수를 탐색할 때와 비교할 때 계산 시간을 줄일 수 있다. 이러한 구조를 갖는 유전자의 길이는 '필드 수 × 8 비트 + 부호를 표시하는 1 개의 비트 + K값의 ±x%에 해당하는 크기의 비트 수'이다.

4.2 초기해의 생성 및 유전 연산자의 적용

초기 세대는 임의로 생성된 50개의 유전자로 구성된다. 세대(Population)의 각 유전자에 자연선택(selection)을 적용하여 선택된 유전자간에 확률

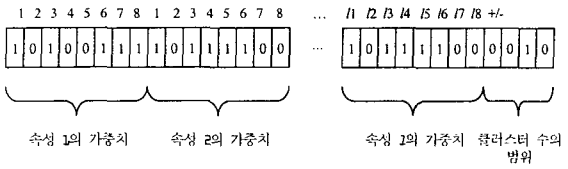

그림 2. 유전자 구조

에 따른 교체와 돌연변이를 적용시킨다. 교체는 임의의 위치에서 일점교체(one-point crossover) 방식을 채택하였다. 돌연변이는 선택된 비트의 값을 현재 값 이외의 값으로 변환시키는 것으로 발생 확률을 0.03으로 하였다.

4.3 적합도 값의 산정

유전 연산자의 적용을 통하여 생성된 새로운 유전자로 임시 세대를 구성하고 원래의 세대와 혼합하여 각 유전자의 적합도 값(Fitness Value)을 구한다. 이때 적합도 값은 유전자 구조에 나타난 가중치를 데이터에 적용하여 의미 있는 속성(가중치가 0 또는 특정 한도값(threshhold value) 보다 큰 값을 갖는 속성)에 대해서 K-means 기법을 적용하여 클러스터링 한 후 적합도 함수를 이용하여 구한다. 적합도 함수는 MOCP의 최소화 목적식을 최대화 문제로 전환하기 위해 임의의 큰 수에서 각 그룹의 중심(centroid)으로부터 데이터의 평균거리를 뺀 값과 그룹간의 평균거리를 합한 값을 사용한다. 그룹간 평균거리는 각 그룹 쌍의 중심까지의 유클리디안 거리의 평균이다. 유전자 s의 적합도 함수 F_s를 수식으로 나타내면 다음과 같다.

$$F_s = c_1 \cdot \{(c_2 - \alpha \cdot W) + (1-\alpha) \cdot B\} + (w_{\max} - wt_s)$$

단, c_1, c_2 = 임의의 큰 값을 갖는 상수,
W = 그룹내 유클리디안 거리의 평균,
B = 그룹간 유클리디안 거리의 평균,
$w_{\max}$ = 최대 가중치의 합,
wt_s = 유전자 s의 가중치의 합,
α = 0과 1 사이의 상수

4.4 다음 세대의 구성

임시 세대와 이전 세대의 유전자들을 적합도 값에 따라 정렬하고 적합도 값의 순위에 따라 상중하위 클래스로 구분한 후 각각의 클래스에서 일정 비율의 유전자를 추출하여 다음 세대를 형성하는 엘리트 그룹 기법을 적용하였다[11]. 본 문에서는 상중하위 3개의 클래스를 각각 20, 50, 30 %로 구분하고 다음 세대의 20 %를 상위 클래스에서, 40 %의 유전자를 중간 클래스, 그리고 나머지 40 %는 하위 클래스에서 추출하도록 하였으며 최상위 하나의 유전자는 다음세대에 그대로 보존되도록 하였다. 이는 전에 탐색된 우수한 형질의 유전자를 보존(elitism)할 뿐 아니라 전 세대의 열악한 유전자를 일정 비율 선택함으로서 나쁜 형질의 유전자가 가지고 있는 정보가 교체와 돌연변이에 의하여 최대한 이용될 수 있도록 하였다.

5. 실험 결과 및 결론

혼합 유전 알고리즘을 비교·분석하기 위하여 2 차원, 3 차원의 데이터 셋 각 3개씩을 임의로 생성하였다. 각 차원의 데이터 셋은 클러스터간의 구분이 명확한 데이터 셋과 클러스터간의 구분이 명확하지 않은 데이터 셋 및 한 속성의 값이 모두 같은 데이터 셋의 세 가지로 생성하였다. 표 1은 혼합 유전 알고리즘의 실행 후 얻어진 속성 가중치이고, 그림 3과 4는 경계가 명확한 2 차원 데이터와 경계가 명확하지 않은 2 차원 데이터의 클러스터링 결과이다. 그림 4의 실선 클러스터는 적합도 함수의 α 값으로 0.55를 사용한 결과이며, 점선 클러스터는 α 값으로 0.5을 사용한 결과이다. 그림 5는 한 속성의 값이 모두 같은 2 차원 데이터 셋에서 첫 번째 속성 가중치의 수렴 과정을 보이고 있다.

구분		속성별 가중치		
데이터 구분	데이터 셋	x	y	z
경계가 명확한 데이터 셋	2 차원(D)	1	1	n/a
	3 차원(D)	1	1	1
경계가 명확하지 않은 데이터 셋	2 차원(B)	0.858824	1	n/a
	3 차원(B)	1	1	1
한 속성의 값이 같은 데이터 셋	2 차원(S)	0	1	n/a
	3 차원(S)	1	1	0

표 1 속성 가중치 결과

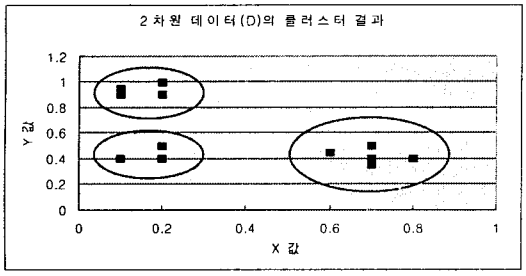

그림 3. 2 차원 데이터(D)의 클러스터링 결과

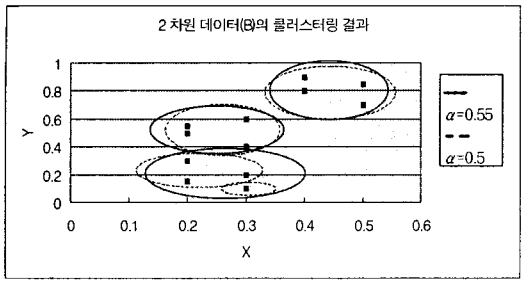

그림 4. 2 차원 데이터(B)의 클러스터링 결과

본 문에서 제시하고 있는 혼합 유전 알고리즘은 클러스터링에 있어서 그룹 수의 결정과 속성선택 문제를 함께 고려하여 이들 상호 관련성으로부터 초래될 수 있는 문제점들을 최소화할 뿐 아니라 사용자가 알고리즘을 실제 문제에 적용할 때의 현실적인 문제점들을 해결하고 있다. 그러나 이러한 적용상의 현실성은 알고리즘의 실행시의 컴퓨팅 시간이라는 비용을 초래한다. 현재, 실제 데이터에 의한 적용 실험과 함께 다양한 클러스터링 측정치의 성능평가, 알고리즘의 시간 효율성 및 클러스터링 해의 품질에 대한 추가적인 연구가 수행 중에 있다.

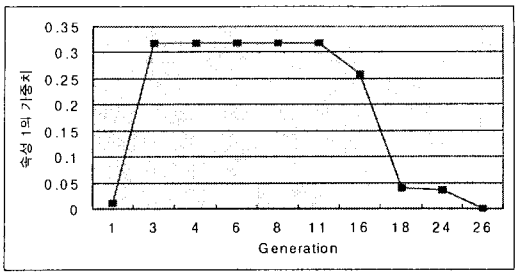

그림 5. 속성 가중치의 수렴

참고 문헌

[1] M. J. A. Berry, Gordon Linoff, "Data Mining Techniques", John Wiley & Sons, Inc., 1997

[2] M. C. Cowgil, R. J. Harvey, and L. T. Watson, "A genetic algorithm approach to cluster analysis", Computers & Mathematics with Applications, Vol. 37, No. 7, pp99-108, 1999

[3] D. E. Goldberg, "Genetic Algorithms", Addison-Wesley Publishing Co., 1989

[4] K. Krishma and M. N. Murty, "Genetic K-Means Algorithm", IEEE Trans. Syst. Man and Cybern., Vol. 29, No. 3, pp433-439, 1999

[5] O. L. Mangasarian, "Mathematical Programming in data mining", Data Mining and Knowledge Discovery, Vol. 1, No. 2, pp183-201, 1997

[6] G. W. Milligan and M. C. Cooper, "An examination of procedures for determining the number of clusters in a data set", Psychometrika, Vol. 50, pp159-175, 1985

[7] J. Mulvey and H. Crowder, "Cluster analysis: An application of Lagrangian relaxation", Management Science, Vol. 25, pp329-340, 1979

[8] W. F. Punch, E. D. Goodman, Min Pei, Lai Chia-Shun, P. Hovland and R. Enbody, "Further Research on Feature Selection and Classification Using Genetic Algorithms", ICGA93, pp557 - 564, Champaign Ill

[9] M. Sarkar, B. Yegnanarayana, and D. Khemani, "A clustering algorithm using an evolutionary programming-based approach", Pattern Recognition Letters, Vol. 18, No. 10, pp975-986, 1997

[10] J. Wang, "A linear assignment clustering algorithm based on the least similar cluster representatives", IEEE Trans. Syst. Man and Cybern., Vol. 29, No. 1, pp100-104, 1999

[11] 최인찬, 김성인, 황대호, "경쟁적 입지 선정 문제의 안정집합을 찾기 위한 수리적 모형과 유전 알고리즘", 대한산업공학회지, 제23권, 제1호, pp223-234, 1997