클러스터 수가 주어지지 않는 클러스터링 문제를 위한 공생 진화알고리즘

신경석* · 김재윤**†

* 전남대학교 산업공학과
** 전남대학교 경영학부

A symbiotic evolutionary algorithm for the clustering problems with an unknown number of clusters

Shin, Kyoung Seok* · Kim, Jae Yun**†

* Dept. of Industrial Engineering, Chonnam National University
** Dept. of Business Administration, Chonnam National University

Key Words : clustering, customer relationship management, symbiotic evolutionary algorithm, data mining

Abstract

Clustering is an useful method to classify objects into subsets that have some meaning in the context of a particular problem and has been applied in variety of fields, customer relationship management, data mining, pattern recognition, and biotechnology etc. This paper addresses the unknown K clustering problems and presents a new approach based on a coevolutionary algorithm to solve it. Coevolutionary algorithms are known as very efficient tools to solve the integrated optimization problems with high degree of complexity compared to classical ones. The problem considered in this paper can be divided into two sub-problems; finding the number of clusters and classifying the data into these clusters. To apply to coevolutionary algorithm, the framework of algorithm and genetic elements suitable for the sub-problems are proposed. Also, a neighborhood-based evolutionary strategy is employed to maintain the population diversity. To analyze the proposed algorithm, the experiments are performed with various test-bed problems which are grouped into several classes. The experimental results confirm the effectiveness of the proposed algorithm.

1. 서 론

클러스터링(Clustering)이란 대용량 데이터로부터 흥미로운 데이터 분포나 패턴을 찾기 위한 방법으로, 주어진 데이터 집합을 유사한 속성을 갖는 몇 개의 그룹으로 분류하는 작업으로 정의된다(Theodoridis and Koutroumbas, 2006). 클러스터링은 데이터 마이닝(Data Mining)의 한 방법으로(Cooley et al., 1997), 특별한 정보나 배경지식 없이 주어진 척도를 사용하여 결과를 도출하는 방법이다. 클러스터링은 고객관리 및 마케팅 전략 수립 등의 경영학 분야와 패턴인식, 영상처리 등의 공학분야, 그리고 유전체 또는 단백체 분석 등의 의료분야 등에 널리 활용되고 있다(오은녕과 이희상, 2002; 황인수, 2002; Xu and Wunsch, 2005; Hruschka et al., 2009). 특히, 클러스터링 분석은 소비자의 쇼핑성향과 구매상황 등을 기초로 동질적인 특성을 지

† 교신저자 jaeyun@jnu.ac.kr
※ 이 논문은 2008년 정부(교육과학기술부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임 (KRF-2008-357-D00288).

닌 집단을 도출, 이를 마케팅에 활용하는 방법으로 매우 유용한 도구이다(김성호와 백승익, 2001). 최근 정보기술의 발달로 데이터를 생성하고 저장할 수 있는 능력이 급격히 증가함에 따라 방대한 양의 데이터로부터 함축적이며 잠재적 가치가 있는 정보를 발견하는데 필요한 핵심적 도구로써 클러스터링 기술은 그 중요성이 더욱 증대되고 있다(Xu and Wunsch, 2005).

클러스터링은 p개의 속성(Attribute, Variable, or Dimension)을 갖는 N개의 데이터 $X=\{X_1,X_2,\cdots,X_N\}$, $X_j=(x_{j1},x_{j2},\cdots,x_{jp})$가 주어졌을 때, 동일한 그룹 내의 유사성은 최대가 되도록, 그리고 서로 다른 그룹간의 유사성은 최소가 되도록 분류하는 것이다. 만약 K개의 클러스터 $C=\{C_1,C_2,\cdots,C_K\}$로 분류되는 경우 식 (1)과 같이 표현될 수 있다.

$$\begin{aligned} &C_i \neq \varnothing \ \text{ for } \ i=1,2,\cdots,K. \\ &C_i \cap C_j = \varnothing \ \text{ for } \ i,j=1,2,\cdots,K, \ i \neq j \\ &\cup_{i=1}^{K} C_i = X \end{aligned} \tag{1}$$

클러스터링을 위한 접근법으로는 크게 계층적(Hierarchical) 방법과 분할적(Partitional) 방법으로 나눌 수 있다(Xu and Wunsch, 2005). 계층적 방법은 각각의 입력 데이터를 하나의 클러스터로 하는 초기해를 가지고 각 계층에서 유사한 클러스터를 합병해 가는 방법(Agglomerative Method)과 반대로 전체 데이터를 하나의 클러스터로 구성한 후, 점차 분할해 가는 방법(Divisive Method)이 있다. 이 방법들은 사전에 클러스터 수를 입력할 필요는 없지만, 알고리즘을 종료해야 할 수준을 결정해야 하는 문제점이 있다.

분할적 방법은 주어진 데이터를 몇 개의 클러스터로 직접 분할하는 방법이다. 이 중 가장 기본이 되고 폭넓게 응용되는 것으로 K-means 알고리즘(Tou and Gonazlez, 1974)이 있다. 이 방법은 간단하면서 구현이 용이하여 여러 문제에 쉽게 적용되어 왔다. 하지만 초기해에 민감하여 쉽게 지역 최적해(Local Optimum)에 수렴하는 단점이 있다. 전역 최적해(Global Optimum)를 찾는 방법으로 Koontz et al.(1975)은 분지한계법(Branch and Bound)에 의한 클러스터링을 제안하였다. 그러나 이 방법은 과도한 계산량으로 인해 대용량 데이터를 포함하여 복잡도가 높은 현실적인 문제에는 적용이 불가능하다.

N개의 데이터가 주어졌을 때 K개 클러스터로 분류하는 클러스터링 문제는 일종의 그룹화(Grouping) 문제로서 NP-hard의 조합최적화에 속하는 문제이다(Liu, 1968). 다수 개의 속성을 갖는 대용량 데이터가 주어졌을 때 주어진 목적함수에 대한 최적 클러스터링 상태(클러스터 수와 각 클러스터의 데이터 집합)를 찾는 것은 현실적으로 불가능하다. 이러한 이유로 클러스터링을 위한 근사 최적해를 찾는 여러 휴리스틱 방법론이 제안되어 왔다. 대표적으로 시뮬레이티드 어닐링(Simulated Annealing) (Selim and Alsultan, 1991; Brown and Huntley, 1992), 타부탐색(Tabu Search) (Al-Sultan, 1995; Sung and Jin, 2000), 그리고 진화알고리즘(Evolutionary Algorithm)에 기반한 다양한 기법들이 제안되었다. 본 연구는 진화알고리즘을 기본 방법론으로 사용하므로, 이에 대한 연구현황은 좀 더 구체적으로 살펴본다.

진화알고리즘은 문제의 잠재해를 유전인자로 구성된 염색체로 표현(Encoding)하여 자연의 진화과정을 모방한 확률적 탐색기법으로써 많은 클러스터링 문제에 적용되어 왔다. Maulik and Bandyopadhyay(2000)은 p개의 속성을 갖는 데이터를 K개의 클러스터로 분류하는 문제에 대해 클러스터 중심을 실수로 표현하여 진화알고리즘을 적용하였다. Bandyopadhyay and Maulik (2002a)는 진화알고리즘에 의해 우수한 클러스터 중심을 찾은 후 Tou and Gonazlez(1974)의 K-means 알고리즘을 이용하여 해를 개선하는 방법을 제안하였다. 이는 K-means 알고리즘이 초기해에 따라 알고리즘 성능에 큰 영향을 미치는 단점을 극복함으로써 성능을 향상시킬 수 있는 방법이다. Garai and Chaudhuri(2004)는 초기 데이터를 유사성에 의해 소그룹들로 묶고, 이들 소그룹들을 진화알고리즘에 의해 보다 큰 그룹으로 클러스터링하는 2단계 과정을 수행하는 방법을 제안하였다. 이 방법은 첫번째 단계에서 문제의 복잡도를 줄여 계산시간을 줄일 수 있다는 장점이 있다. 이상의 연구들은 많은 클러스터링 문제에 좋은 결과를 보여주었지만, 분류되어야 할 클러스터 수가 미리 알려진 경우에 적용 가능하다.

많은 현실적인 문제에서는 클러스터 수가 사용자에게 알려져 있지 않다. 예를 들어, 문숙경과 김우성(2004)이 다룬 문제에서와 같이 특정 제품에 대하여 여러 경쟁 업체들이 각각의 고유 상표를 가지고 생산하는 제품들의 판매량 자료를 바탕으로 비슷한 성장패턴을 가지는 기업들을 그룹화할 때 비슷한 패턴의 수가 사전에 주어지지 않을 수 있다. 이 경우 데이터 자체에서 적절한 클러스터 수를 추정하여야 한다. 이 때 클러스터링

알고리즘의 성능은 추정된 클러스터 수에 따라 많은 영향을 받는다. 이러한 이유로 클러스터의 수와 클러스터 집합을 동시에 결정하는 기법들이 제안되었다. Tseng and Yang (2001)은 nearest-neighbor 알고리즘에 의해 초기 데이터를 m개의 소그룹으로 분류한 다음, m개의 이진(Binary) 인자를 갖는 염색체로 표현하여 진화알고리즘을 적용하였다. 인자값이 1인 소그룹이 클러스터의 중심 역할을 하고, 인자값이 0인 소그룹들은 클러스터간 차별성과 클러스터내의 유사성에 따라 병합된다. 이 방법은 m개로 분류된 소그룹이 클러스터의 중심 역할을 하여 해공간을 축소시키는 단점이 있다. Bandyopadhyay and Maulik(2002b)는 실수와 null 기호를 사용한 유전표현을 사용하였다. 염색체의 크기는 임의의 클러스터 수의 최대값으로 하고, 그 중 일부 인자가 중심위치로 선정되어 실수로, 나머지 인자는 null 기호로써 표현된다. 따라서 선정된 중심위치의 수가 클러스터 수가 되고 데이터들은 이들 중심위치 중 가장 가까운 곳에 할당됨으로써 하나의 해로 해석된다.

본 연구에서는 클러스터 수가 주어지지 않는 클러스터링 문제를 진화알고리즘의 한 변형인 공생 진화알고리즘(Coevolutionary Algorithm)에 의해 해결하는 방법을 제안한다. 전통적인 진화알고리즘은 전체 해를 하나의 염색체로 표현한 모집단을 운영하여 해를 탐색한다. 반면 공생 진화알고리즘은 전체문제를 여러 개의 부분문제로 분해하고, 각 부분문제에 대한 염색체로 이루어진 부분문제 모집단을 운영한다. 즉, 공생 진화알고리즘은 부분문제 염색체로 이루어진 복수개의 모집단을 운영함으로써 해를 탐색하는 방법이다. 모집단이 부분문제를 위한 염색체로 이루어져 있으므로 각 부분문제 모집단 염색체는 상대 모집단의 염색체(다른 부분문제 모집단 염색체)들과 결합하여 적응도가 평가되고, 평가된 적응도에 따라 선택과 유전연산을 수행한다. 이러한 과정은 생물계에서 공생(Symbiosis) 관계를 갖는 서로 다른 종들이 상호작용 및 상호적응하며 공진화하는 과정을 모방한 것이다. 공생 진화알고리즘은 복잡도가 높은 문제나 여러 부분문제가 결합된 통합문제를 해결하는데 전통적인 진화알고리즘보다 좋은 해의 탐색 측면에서 우수한 것으로 알려져 있다(Moriarty and Miikkulainen, 1997; Kim, *et al.,* 2003).

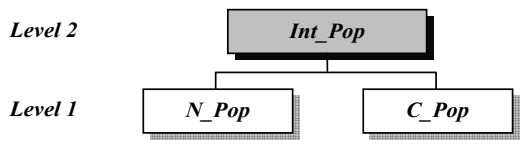
본 연구에서 다루는 클러스터링 문제는 클러스터 수를 결정하는 부분문제와 클러스터 수에 따라 데이터를 분류하는 부분문제로 분해될 수 있다. 이러한 문제에 공생 진화알고리즘을 적용하기 위해 유전표현을 개발하고, 이에 대한 알고리즘 구조와 구성요소, 그리고 진화전략들을 제안한다. 다루는 문제는 복잡도가 매우 높으며, 제안한 방법론을 사용한 기존 연구는 아직까지 존재하지 않는다. 이후 본 논문의 구성은 다음과 같다. 2장에서는 알고리즘의 개념과 절차를 기술하고, 3장에서는 알고리즘 적용을 위한 유전요소를 기술한다. 4장에서는 기존의 알고리즘과 비교분석을 위한 실험 및 결과를 설명한다. 마지막으로 5장은 결론으로 연구내용을 요약 정리한다.

2. 클러스터링을 위한 공생 진화알고리즘

2.1 알고리즘의 개념

자연계에서 다양한 종들이 상호작용하고 상호적응하는 메커니즘이 존재하듯이, 이를 모방한 공생 진화알고리즘에서도 부분문제들이 유기적으로 작용하면서 좋은 해를 탐색할 수 있는 진화 메커니즘이 필요하다(Kim et al., 2003). 여기에는 부분문제(모집단)의 수와 구조, 진화전략, 적응도 평가를 위한 공생자(Symbiont) 선택, 그리고 표현과 연산자를 포함한 유전요소 등이 포함된다. 특히, 공생 진화알고리즘은 메타휴리스틱(Meta-heuristic) 기법이므로 다루는 문제에 따른 진화메커니즘을 설계하는 것만으로도 연구의 대상이 된다.

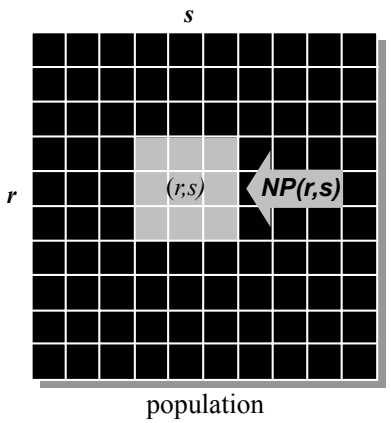
본 연구에서 제안한 공생 진화알고리즘은 <그림 1>과 같은 2계층 구조를 갖는다. 계층 1에서는 두 개의 모집단, 즉 클러스터의 수와 위치를 표현한 염색체로 이루어진 모집단(*N_Pop*)과 클러스터 중심 좌표값을 표현한 모집단(*C_Pop*)을 갖는다. 계층 2에서는 이들 두 모집단 염색체가 결합된 전체문제 염색체로 이루어진 모집단(*Int_Pop*)으로 구성된다. 즉, 제안한 공생 진화알고리즘에서는 3개의 모집단이 상호작용을 하면서 동시에 진화한다. 모든 모집단은 토러스(Torus) 형태의 2차원 격자구조로 구성된다.



<그림 1> 알고리즘의 모집단 구조

각 모집단은 이웃단위 진화와 안정상태 유전알고리

즘(Steady-state Genetic Algorithm)의 진화전략을 사용한다. 이웃진화는 다양한 우수 염색체들이 적소(Niche)를 형성하도록 하여 해의 조기수렴을 방지할 뿐 아니라, 해공간의 효율적 탐색을 가능하게 한다(Kim et al., 2000). 안정상태 유전알고리즘은 높은 적응도를 갖는 염색체가 생산되는 즉시 재생산에 참여하는 방법으로, 우수한 자손 염색체가 갖는 유전정보의 이용을 높일 수 있다. <그림 2>는 한 모집단의 위치 *(r, s)*에서 (3×3) 크기의 이웃 *NP(r, s)*를 보여준다. 그림에서 하나의 격자는 하나의 염색체를 의미한다. 결과적으로 계층 1, 2의 각 모집단 *N_Pop*, *C_Pop* 및 *Int_Pop*은 다른 염색체를 담고 있는 동일한 모양을 갖는 모집단이다.



<그림 2> 이웃정의

2.2 알고리즘의 절차

제안한 알고리즘의 단계 1에서는 각 모집단이 나타내려는 해의 표현방법에 따라 임의로 염색체를 생성하고 초기 적응도를 평가한다. 단계 2는 진화대상이 되는 이웃을 정의하는 단계이며, 단계 3에서는 선택된 이웃이 계층내 및 계층간의 다른 이웃들과 상호작용한다. 계층내 상호작용은 계층 1에 있는 모집단들의 염색체들이 하나의 완전한 해를 표현하지 못하므로 상대 모집단의 염색체들을 공생자로 선택하고, 이들 공생자들과의 결합을 통하여 적응도 평가가 이루어진다. 공생자 선택 전략은 다양할 수 있으나, 본 연구에서는 단순한 전략으로 동일 계층에 있는 상대 모집단의 이웃에서 임의로 한 염색체를 선택하는 전략을 사용한다. 계층 2에 있는 염색체들은 해를 완전하게 표현하고 있으므로, 공생자 선택이 필요없다. 그리고 계층간 상호작용은 계층 1의 적응도 평가 과정에서 적응도가 높은 염색체결합이 발견되면, 적응도를 기준으로 하여 계층 2로 전달하는 상향식 정보전달과정을 통해 이루어진다.

단계 4에서는 모든 계층에 존재하는 모집단의 이웃들이 진화한다. 이때 진화는 앞에서 설명한 바와 같이 안정상태 유전알고리즘의 형태를 따른다. 단계 5에서 종료조건을 판단하여, 알고리즘의 종료 여부를 결정한다. 제안한 알고리즘의 구체적인 절차는 다음과 같다.

단계 1 (초기화와 초기 적응도 평가)

단계 1.1 계층 1의 *N_Pop*, *C_Pop*, 계층 2의 *Int_Pop*의 초기 모집단을 초기해 생성방법을 통해 2차원 격자구조로 생성한다.

단계 1.2 계층 1의 각 부분모집단 염색체들을 상대 모집단의 같은 위치에 있는 염색체와 짝을 지어 적응도를 평가한다. 이 때 가장 높은 적응도를 갖는 결합염색체와 적응도를 각각 Ind^*와 f^*로 둔다.

단계 1.3 계층 2의 모집단내 염색체들의 적응도를 평가하고 가장 좋은 적응도와 해당 염색체를 각각 f_{best}와 Ind_{best}로 둔다. 만약 $f^* > f_{best}$ 이면 $f_{best} = f^*$, $Ind_{best} = Ind^*$로 갱신한다.

단계 2 (이웃 설정)

임의 위치 *(r, s)*를 선택하여 계층 1 모집단 *N_Pop*, *C_Pop*의 이웃 $NP_N(r, s)$, $NP_C(r, s)$, 계층 2 모집단 이웃 $NP_{Int}(r, s)$를 정의한다.

단계 3 (적응도 평가와 계층간 상호작용)

단계 3.1 계층 1의 각 이웃 염색체는 상대 모집단내 임의의 염색체와 짝을 지워 적응도를 평가한다.

단계 3.2 단계 3.1의 결과에 따라 각 이웃에서 가장 높은 적응도를 갖는 결합염색체와 적응도를 각각 Ind^*와 f^*로 둔다. Ind^*와 f^*를 계층 2의 이웃 $NP_{Int}(r, s)$에서 가장 낮은 적응도를 갖는 염색체와 비교하여 계층 1의 결합염색체 적응도가 높으면 이와 대체한다.

단계 3.3 계층 2의 이웃 $NP_{Int}(r, s)$의 적응도를 구한다. 이웃에서 가장 높은 적응도를 갖는 염색체와 적응도를 Ind_{old}와 f_{old}로 둔다. 만약 $f_{old} > f_{best}$이면 $f_{best} = f_{old}$, $Ind_{best} = Ind_{old}$로 갱신한다.

단계 4 (이웃 진화)

단계 4.1 계층 1, 2의 각 이웃에 대해 적응도를 기준

으로 두 부모 염색체를 확률바퀴 (Roulette Wheel)방법에 의해 선택하고, 교차연산을 통해 두 자손 염색체를 생산한다.

단계 4.2 이웃 내에서 적응도가 가장 낮은 두 염색체를 단계 4.1에서 생산한 자손염색체와 대체한다.

단계 4.3 이웃 내 염색체들을 돌연변이율에 따라 돌연변이 시킨다.

단계 5 (종료조건)

알고리즘 종료조건을 만족하면 끝내고, 그렇지 않으면 단계 2로 간다.

3. 알고리즘의 유전요소

3.1 유전표현과 초기 모집단

클러스터링 문제를 공생 진화알고리즘으로 해결하기 위해 각 모집단은 특성에 맞도록 염색체(해)를 표현해야 한다. 먼저 계층 1에서는 클러스터의 수를 표현한 염색체를 담고 있는 모집단 *N_Pop*과 클러스터 중심의 위치를 표현한 염색체로 이루어진 모집단 *C_Pop*이 있다. *N_Pop*의 염색체는 이진표현으로 그 길이는 분류할 수 있는 클러스터의 최대치(Kmax)이다. Kmax는 파라미터로서 문제의 규모와 성격에 따라 다르게 설정될 수 있다. 염색체의 인자값이 1이면, *C_Pop* 염색체의 동일 위치에 있는 중심값이 해당 클러스터의 중심이 되고, 반대로 인자값이 0이면 무시된다. 따라서 *N_Pop* 염색체에서 인자값이 1인 개수가 분류될 클러스터의 수가 된다. *C_Pop* 염색체의 각 인자는 실수로 표현되며 클러스터의 중심위치를 의미한다. 이 표현은 클러스터의 수와 중심위치가 이진표현과 실수표현으로 사용되어 자연스러운 염색체 해석과 유전연산이 용이하다는 장점이 있다. <그림 3>은 K_{max} = 4일 때, *N_Pop* 염색체와 *C_Pop* 염색체의 표현 예를 보여준다. 계층 2의 *Int_Pop* 염색체는 *N_Pop*과 *C_Pop* 염색체가 결합된 형태로써 염색체의 길이는 두 염색체 길이의 합이 되고, 완전한 전체문제 해를 표현한다.염색체의 해석은 계층 1에서의 경우 먼저 *N_Pop*의 염색체와 *C_Pop* 염색체가 결합됨으로써 완전한 형태의 염색체를 형성한다. 이때 서로 결합되는 염색체의 결정은 알고리즘의 단계 3.1에서 이루어진다. *N_Pop* 염색체에서 인자값이 1인 개수가 클러스터의 수가 되고, 1의 위치와 동일한 위치에 있는 *C_Pop* 염색체의 인자값이 해당 클러스터의 중심 위치이다. 분류할 데이터 각각은 이들 중심위치와 가장 가까운 곳에 할당됨으로써 클러스터 수의 결정과 데이터 클러스터링이 동시에 수행된다. 계층 2는 *N_Pop*과 *C_Pop* 염색체가 결합된 완전한 형태의 염색체이므로, 해석을 위해 별도의 염색체를 선택할 필요가 없으며 동일한 방법으로 해석될 수 있다.

(0 1 1 0)

a) *N_Pop*의 염색체 표현

{(50.11,18.62)(54.91,15.97)(17.81,33.26)(17.29,67.71)}

b) *C_Pop*의 염색체표현

<그림 3> 부분모집단의 염색체표현

초기 모집단 생성은 계층 1의 *N_Pop* 염색체의 경우, [1, K_{max}]사이 임의의 정수를 생성하여 그 수만큼 임의 위치의 인자값으로 1을 할당하고, 남은 인자값은 0으로 채운다. 이러한 과정으로 모집단 크기만큼 염색체를 생성하여 모집단을 초기화한다. *C_Pop*의 염색체는 각 인자값이 중심위치를 표현하므로 데이터의 범위, 즉 $\left[x_d^{\min}, x_d^{\max}\right]$ 내에서 임의의 수를 선택하여 각 인자의 값으로 설정한다. 여기서 $x_d^{\min}$, $x_d^{\max}$는 데이터 집합에서 *d*번째 속성의 최소값과 최대값을 각각 나타낸다. 계층 2는 계층 1의 동일한 표현의 염색체가 결합된 형태이므로 역시 같은 방식으로 생성하여 초기 모집단을 구성한다.

3.2 적응도 평가함수

진화알고리즘에서 적응도 평가는 흔히 다루는 문제의 목적함수를 이용하여 평가된다. 본 연구에서 다루는 클러스터링의 목적은 클러스터 내 데이터들의 유사성 또는 클러스터 간 차이가 최대화 되도록 하는데 있다. 이러한 목적에 클러스터링의 결과가 얼마나 부합하는지에 대한 다양한 척도가 연구되었다(Halkedi et al., 2001). 본 연구에서는 식 (2)의 DB index (Davies and Bouldin, 1979)를 목적함수로 사용하고 이를 이용하여 적응도를 평가한다.

$$DB = \frac{1}{K}\sum R_i$$

$$R_i = \max_{i,i' \neq j}\{R_{ii'}\}$$

$$R_{ii'} = \frac{S_i + S_{i'}}{d_{ii'}}$$

$$S_i = \frac{1}{|C_i|}\sum_{X_j \in C_i} \| X_j - Z_i \|$$

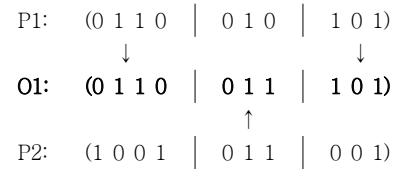
$$\| X_j - Z_i \| = \sqrt{\sum_{d=1}^{p}(x_{id} - z_{id})^2}$$

$$z_{ip} = \frac{1}{|C_i|}\sum_{j} x_{jp},\ d_{ii'} = \| Z_i - Z_{i'} \| \quad (2)$$

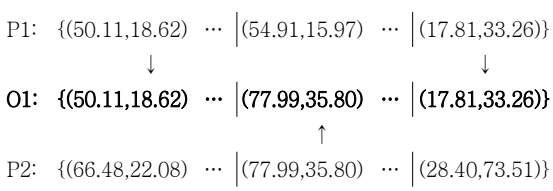
S_i값은 클러스터 C_i에 속하는 데이터들의 유사정도를 나타내는 값이고, $d_{ii'}$는 클러스터간 거리를 의미한다. S_i와 $d_{ii'}$를 이용하여 각 클러스터에 대한 $R_{ii'}$ 값을 구하고, 이중 가장 큰 값을 클러스터C_i의 R_i값으로 한다. DB index는 각 클러스터가 갖는 R_i의 합을 클러스터 수 K로 나눈 평균이므로 클러스터간 거리가 멀수록 클러스터 내 데이터의 유사성이 높을수록 낮은 값을 갖게 된다. 즉, 우수한 클러스터가 형성될수록 더 적은 값을 갖는다. 우수한 염색체에 대해 높은 적응도 값을 갖도록 하기 위해 DB index의 역수를 취하여 적응도 값으로 한다.

3.3 유전연산

진화알고리즘에서 유전연산은 흔히 교차(Crossover)와 돌연변이(Mutation)가 있다. 교차는 부모가 갖는 정보를 자손에 전달하는 역할을, 돌연변이는 모집단의 다양성을 유지하는 역할을 한다. 특히 교차연산은 다루는 문제가 갖는 의미있는 정보가 교차에 의해 파괴되지 않도록 하면서 그 과정이 단순한 방법이 좋다.



a) *N_Pop* 모집단의 교차연산



b) *C_Pop* 모집단의 교차연산

<그림 4> 교차연산

계층 1의 *N_Pop*과 *C_Pop* 염색체는 각각 이진표현과 실수표현으로서 이 표현에 맞는 전통적인 교차연산자를 그대로 적용할 수 있다. 본 연구에서는 적용이 용이한 이점교차(Two-point Crossover)를 사용한다. <그림 4>는 *N_Pop*과 *C_Pop* 염색체의 이점교차를 이용하여 자손(O1)을 생산하는 방법이다. 또 다른 자손(O2)은 부모의 역할을 바꾸어 생산한다.

돌연변이 연산은 *N_Pop* 염색체의 경우 돌연변이율에 의해 인자를 선택하고, 해당 인자값이 0이면 1로, 1이면 0으로 변경한다. 이를 통해 클러스터의 수에 변화를 줄 수 있다. 중심위치가 실수값으로 표현된 *C_Pop* 염색체는 돌연변이율에 의해 인자를 선택하고, 선택된 인자값은 다음 식 (3)에 의해 돌연변이를 수행한다.

$$z_{kj}^{t+1} = z_{kj}^{t} + N(0,1) \quad (3)$$

여기서 z_{kj}^{t}는 세대 t에서 k번째 클러스터 중심위치의 j번째 속성이고, N(0, 1)은 평균이 0, 표준편차가 1인 정규분포를 갖는 난수이다. *C_Pop* 염색체의 돌연변이 수행과정은 예를 들어 설명하고자 한다. 예로, <그림 4> (b)의 O1 염색체에서 위치 z_{12}, 즉 첫번째 클러스터의 두번째 속성값(18.62)이 돌연변이 대상 인자로 선택되고, 이에 대한 정규난수가 -1.36이면, 돌연변이 후 인자값은 17.26 (= 18.62 - 1.36)이 된다. 따라서 O1 염색체는 돌연변이를 통해 첫번째 클러스터 중심위치가 (50.11, 18.62)에서 (50.11, 17.26)으로 변경된다. 계층 2의 *Int_Pop* 염색체의 교차와 돌연변이는 *N_Pop*와 *C_Pop* 염색체가 결합된 형태이므로 계층 1에서와 동일한 방식으로 유전연산이 수행된다.

4. 실험과 분석

4.1 실험설계

클러스터의 수와 클러스터의 중심위치를 동시에 결정하는 문제는 클러스터링 수행 후 구해진 클러스터의 유효성(Validity) 측정이 쉽지 않기 때문에 수행하는 알고리즘의 비교가 용이하지 않다. 서론에서 언급한 바와 같이 클러스터링의 목적은 클러스터 내 데이터들의 유사성은 최대가 되도록, 또한 클러스터 간의 유사성은 최소가 되도록 하는 것이다. 그러나 이 두 목적은 상충적인 면이 존재한다. 이러한 상충적인 관계를 하나의 지표로 측정하는 방법, 즉 클러스터 유효성을 평가하는 여러 기법들이 있지만 본 연구에서는 적응도 평가를 위

해 사용된 식 (2)의 DB index를 그 척도로 삼았다.

제안한 알고리즘은 Bandyopadhyay and Maulik (2002b)의 GCUK(Genetic Clustering Unknown K) 및 Kim et al.(2003)의 단일 계층 공생 진화알고리즘과 해의 탐색 성능 측면에서 비교 분석하였다. GCUK에서 표현은 본 연구에서 제안한 알고리즘과 유사하지만, 하나의 염색체에 클러스터 수와 중심 위치를 함께 표현하여 전통적인 진화알고리즘의 절차를 따른다. 단일 계층 공생 진화알고리즘은 계층적 구조의 효과를 보이기 위해 비교한 것으로, 제안한 알고리즘에서 계층 2는 제외되고 계층 1의 구조로만 수행되는 알고리즘이다. 설명을 위해 본 연구에서 제안한 알고리즘을 TSEA(Two-leveled Symbiotic Evolutionary Algorithm), 단일 계층 공생 진화알고리즘을 SSEA(Single-leveled Symbiotic Evolutionary Algorithm)라고 부르기로 한다.

각 알고리즘은 모두 JAVA 언어로 구현되었으며, Intel Core2 Duo 3.0GHz CPU를 장착한 IBM 계열 PC에서 수행되었다. TSEA, SSEA, 그리고 GCUK에 필요한 파라미터로, 먼저 모집단 크기는 TSEA의 계층 1과 계층 2 모두 100, SSEA의 각 부분 모집단과 GCUK의 모집단 크기를 100으로 두었다. GCUK의 교차율은 0.7로 설정하였으며, TSEA, SSEA는 이웃내 9개의 염색체들 중 2개를 선택하여 교차가 진행되므로 교차율은 약 0.28이라고 볼 수 있다. 세 알고리즘 모두 돌연변이율은 0.1을 적용하였다. 종료조건은 자손의 생산 개수로 두고 5,000개의 자손이 생산되면 종료하도록 하였고, 종류 후 가장 우수한 해를 비교대상으로 삼았다.

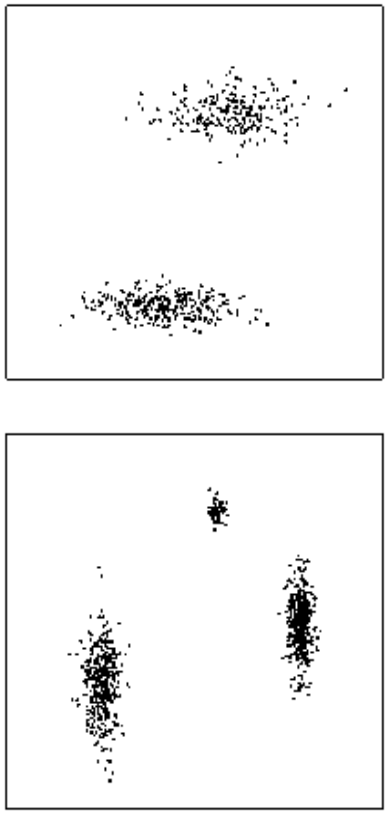
4.2 분석결과

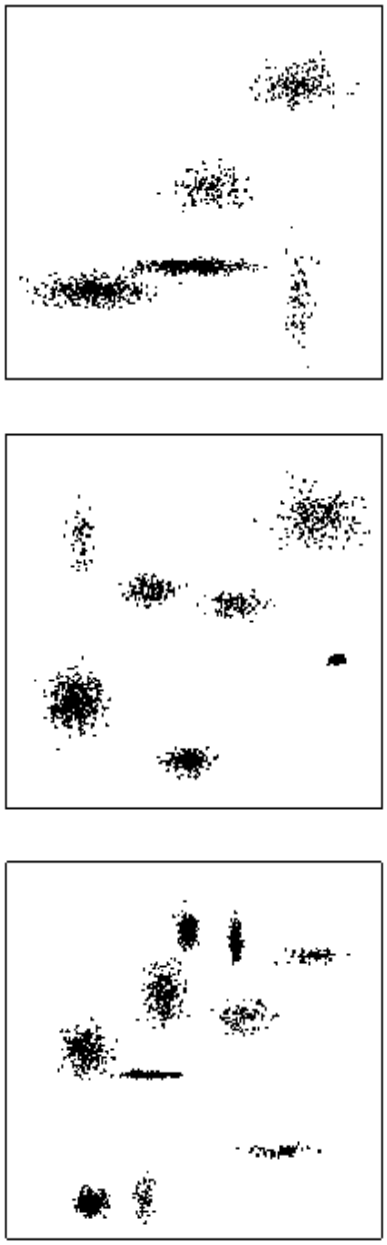
1) 2개의 속성을 갖는 실험문제

시각적인 분석을 위해 2개의 속성을 갖는 실험문제를 생성하여 실험하였다. 2개의 속성을 갖는 실험문제는 고객관계관리에서 N명의 고객과 2개의 제품으로 이루어진 제품-고객 관계테이블에서, 미리 결정되지 않은 K개의 그룹(클러스터)으로 고객을 분류하는 문제와 같다. 예를 들어, 2개의 제품 A와 B가 있고 각 고객은 두 제품에 대하여 숫자로 표시된 선호도를 가질 때, 주어진 N명의 고객을 선호도의 유사 정도가 가장 낮도록 분류하는 것이 된다. 이는 고객관리 및 홍보에 필요한 전략을 수립하는데 있어 각 고객 그룹의 특성에 맞는 전략을 개발하는데 유용하게 사용될 수 있다.

실험문제는 클러스터링 결과를 비교하기 위해 몇 개의 그룹을 형성하도록 생성되었다. 먼저 그룹에 포함될 데이터 수를 [50, 450]에서 임의로 취한 후 그룹 내 각 데이터는 $N(\mu,\sigma)$의 정규난수를 생성하여 구성하였다. 이 때 평균 μ와 표준편차 σ는 각각 [0, 100], [0, 5]의 범위에서 임의로 선택되었다. 그 결과 각 문제에 따라 그룹은 <그림 5>와 같이 2, 3, 5, 7, 10개를 가지며, 데이터 크기는 각각 510, 837, 1406, 1630, 2054개로 이루어져 있다. 각 문제는 (D, C, N)의 형태로 표기하였고, D는 속성 수, C는 그룹 수, N은 총 데이터 크기이다.

<표 1>은 실험문제에 대한 결과를 보여준 것으로, 각 실험문제에 대하여 비교 알고리즘들이 구한 클러스터 수와 DB index 값을 표시하였다. 모든 값들은 20회 반복실험의 평균값이고, 개선율(Improved Rate) = {(비교알고리즘의 평균 TSEA의 평균)/(비교알고리즘의 평균)}×100(%) 이다. 표에서 알 수 있듯이 TSEA와 비교알고리즘은 모두 실험문제가 갖는 그룹의 수를 찾았다. 이는 문제에서 그룹의 경계가 명확하여 쉽게 클러스터링이 이루어진 결과이다.





<그림 5> 2개의 속성을 갖는 실험문제

그러나 DB index 값의 비교에서는 제안한 TSEA가 가장 우수하였고, 다음으로 SSEA, 그리고 GCUK가 가장 낮은 성능을 보였다. 이 결과는 구해진 클러스터의 중심위치가 실제 문제의 클러스터 중심위치에 얼마나 근사한가를 보여준다고 할 수 있다. 2계층으로 이루어진 TSEA가 단일 계층구조의 SSEA보다 성능이 좋은 이유는 내공생진화(Endosymbiotic Evolution) 과정(Margulis, 1981)을 모방하고 있기 때문이다. 즉, 계층 1에서 분리된 공생진화는 병렬탐색을 강화하여 넓은 해공간을 탐색하고, 계층 2에서 결합된 형태의 내공생진화는 유망한 좁은 해공간의 집중탐색에 기여한 것으로 판단된다. 또한 진화가 진행되는 동안 하위계층에서 발견된 좋은 해들이 결합된 내공생자로 만들어져 상위계층으로 이동하는 진화전략과 조화되어 모집단의 다양성(Diversity) 및 해의 수준(Quality) 향상에 기여함으로써 우수한 해를 탐색한다고 볼 수 있다.

2) 다수 개의 속성을 갖는 실험문제

다음으로, 다수 개의 속성을 갖는 문제에 대해 TSEA의 성능을 분석하였다. 이는 문제의 복잡도 증가에 따른 TSEA의 탐색 성능 변화를 살펴보기 위한 것이다. 각 문제는 2개의 속성을 갖는 문제와 동일하게 몇 개의 그룹을 형성하도록 하였고 데이터 생성방법도 앞에서 설명한 것과 모두 동일하다. <표 2>는 속성 수(D), 그룹 수(C), 데이터 크기(N)로 표현된 (D, C, N) 형태의 실험문제와 각 비교알고리즘에 의한 DB index 값을 보인 것이다. 개선율은 2개의 속성을 갖는 문제의 실험에서와 동일한 방식으로 제시하였다.

표에서 알 수 있듯이 3개의 문제를 제외한 대부분 문제에서 제안한 TSEA가 좋은 결과를 보였다. 실험 결과는 하나의 전체문제 염색체를 통한 단일모집단을 운영하는 경우보다 여러 부분문제의 모집단을 운영하는 병렬적 탐색이 복잡도가 높은 문제에 보다 효율적임을 의미한다. 이는 기존 공생 진화 알고리즘의 많은 연구 결과(Kim et al., 2000; 2003)와 부합된다.

다수 개의 속성을 갖는 실험문제의 경우 각 문제에 따라 찾아진 클러스터 수가 일정하지 않았다. 본 연구에서 다루는 클러스터링 문제는 분류되어야 할 클러스터 수가 미리 주어져 있지 않은 문제로 제안한 알고리즘은 주어진 척도(DB index)를 최소화하는 클러스터의 수와 클러스터 집합을 동시에 결정한다. 따라서 실험데이터가 몇 개의 그룹을 형성하도록 생성되었다 할지라도, 그룹들을 병합(클러스터 수 감소)하거나 분할(클러스터 수 증가)하는 것이 DB index값을 보다 최소화할 수 있다.

알고리즘 수행을 위한 계산소요시간은 많은 계산시간을 요구할 것으로 예상되는 다수 개의 속성을 갖는 문제에 대하여 제시하고자 한다. 16개의 실험문제에 대하여, 3개 비교 알고리즘의 (최소, 최대, 평균: 단위 sec.) 계산소요시간은 GCUK (9.4, 136.9, 45.8), SSEA (27.4,

360.5, 123.2), TSEA (20.0, 273.1, 91.4) 이다. 동일한 규모의 문제에 대해 GCUK가 가장 적은 시간이 소요됐으며, 다음으로 TSEA, SSEA가 가장 많은 계산시간을 요구하였다. 이는 단일 모집단의 염색체가 해석되는 전통적인 진화알고리즘과는 달리 공생 진화알고리즘은 각 부분모집단의 염색체가 상대모집단 염색체와 결합하여 해석되어야 하기 때문에 그만큼 적응도 평가시간을 더 요구하기 때문이다. 본 연구에서 다루는 대용량 데이터의 클러스터링은 실시간을 요구하기 보다는 사전계획(pre-planned) 문제의 특성을 갖기 때문에 좋은 해를 탐색하는 것이 적은 계산 소요시간보다 우선한다고 판단된다.

5. 요약 및 결론

본 연구에서는 클러스터 수가 주어지지 않은 클러스터링 문제를 해결하기 위해 공생 진화알고리즘 기반의 방법론을 제시하였다. 공생 진화알고리즘을 적용하기 위해 다루는 문제를 클러스터 수와 중심위치를 결정하는 부분문제로 각각 분해하여 각 부분문제에 맞는 유전표현과 유전연산을 제안하고, 해 탐색의 효율성을 위해 공생 진화알고리즘을 수행하는 부분문제 계층과 전통적인 진화과정을 수행하는 전체문제 계층을 결합한 2계층 구조를 제안하였다.

<표 1> 2개의 속성을 갖는 실험문제 결과

Problems	GCUK		SSEA		TSEA		Improved Rate(%)	
	# of cluster	DB index	# of cluster	DB index	# of cluster	DB index	vs GCUK	vs SSEA
(2,2,510)	2	0.187	2	0.182	2	0.182	3.0	0.0
(2,3,837)	3	0.184	3	0.163	3	0.160	13.2	2.0
(2,5,1406)	5	0.317	5	0.318	5	0.312	1.9	1.9
(2,7,1630)	7	0.384	7	0.362	7	0.359	6.6	0.9

<표 2> 다수개의 속성을 갖는 실험문제 결과

Problems	GCUK	SSEA	TSEA	Improved Rate(%)	
				vs GCUK	vs SSEA
(3,3,759)	0.130	0.130	0.122	6.3	6.3
(3,5,1242)	0.337	0.346	0.271	19.6	21.7
(3,7,1734)	0.271	0.324	0.306	-12.9	5.5
(3,10,2725)	0.446	0.439	0.431	3.3	1.7
(5,3,990)	0.142	0.142	0.137	3.6	3.6
(5,5,1369)	0.233	0.254	0.209	10.0	17.6
(5,7,2248)	0.253	0.235	0.275	-8.7	-17.0
(5,10,2297)	0.491	0.533	0.475	3.1	10.8
(7,3,592)	0.148	0.165	0.144	2.7	12.3
(7,5,1377)	0.289	0.284	0.246	14.9	13.3
(7,7,1495)	0.342	0.386	0.321	6.4	17.0
(7,10,1991)	0.409	0.400	0.442	-8.2	-10.4
(10,3,1016)	0.194	0.194	0.183	5.3	5.3
(10,5,1047)	0.290	0.353	0.248	14.5	29.6
(10,7,1581)	0.422	0.460	0.357	15.5	22.4
(10,10,2894)	0.514	0.446	0.486	5.5	-8.9

제안한 알고리즘의 성능을 분석하기 위해 여러 형태로 이루어진 실험문제를 생성하여 비교 분석하였다. 실험결과, 기존의 단일 모집단을 이용한 전통적인 진화알고리즘에 비해 부분문제로 이루어진 부분모집단을 운영하는 공생 진화알고리즘의 탐색성능이 우수하였고, 부분모집단만을 운영하는 단일계층 공생 진화알고리즘보다 본 연구에서 제안한 2계층 공생 진화알고리즘이 전반적으로 우수함을 보였다. 이러한 결과는 클러스터링의 유효성을 평가하는 적절한 척도와 함께 적용되어 클러스터링을 위한 유망한 도구로 사용될 수 있음을 보여준다. 또한, 공생 진화알고리즘은 복잡도가 높은 부분문제들을 통합적으로 해결할 수 있다는 특징을 갖기 때문에 알고리즘 적용의 유연성을 보인 결과라고 생각된다.

최근 연구들은 공생 진화알고리즘이 기존의 진화알고리즘에 비해 해 탐색성능이 우수함을 보여주고 있다. 하지만 공생 진화알고리즘은 각 부분모집단에서의 평가로 인해 계산시간의 증가를 초래한다. 따라서 클러스터링을 위한 공생 진화알고리즘을 적용하는 데 있어 계산시간 단점을 보완하기 위한 각 부분문제의 표현방법과 평가방법 등은 향후 연구주제가 될 수 있다.

참고문헌

[1] 김성호, 백승익(2001), "인위적 데이터를 이용한 군집분석 프로그램간의 비교에 대한 연구,"「지능정보연구」, 7권, 2호, pp. 35-49.

[2] 문숙경, 김우성(2004), "마케팅자료에서 특성점들을 이용한 군집방법,"「품질경영학회지」, 32권, 4호, pp. 265-273.

[3] 오은녕, 이희상(2002), "클러스터링 기법을 이용한 이동통신의 고객 세분화 연구,"「한국경영과학회 추계논문집」, pp. 421-424.

[4] 황인수(2002), "데이터 마이닝에서 그룹 세분화를 위한 2단계 계층적 클러스터링 알고리듬,"「경영과학」, 19권 1호, pp. 189-196.

[5] Al-Sultan, K.(1995), "A Tabu search approach to the clustering problem," *Pattern Recognition*, Vol. 28, No. 9, pp. 1443-1451.

[6] Bandyopadhyay, S. and Maulik, U.(2002a), "An evolutionary technique based on K-Means algorithm for optimal clustering in R^N," *Information Sciences*, Vol. 146, pp. 221-237.

[7] Bandyopadhyay, S. and Maulik, U.(2002b), "Genetic clustering for automatic evolution of clusters and application to image classification," *Pattern Recognition*, Vol. 35, pp. 1197-1208.

[8] Brown, D. and Huntley, C.(1992), "A practical application of simulated annealing to clustering," *Pattern Recognition*, Vol. 25, No. 4, pp. 401-412.

[9] Cooley, R., Mobasher, B. and Srivastava J. (1997), "Web Mining: Information Pattern Discovery on the World Wide Web," Proc. of the 9th IEEE International Conference, pp. 558-567.

[10] Davies, D.L. and Bouldin, D.W.(1979), "A cluster separation measure," *IEEE Transactions on Pattern Recognition and Machine Intelligence*, Vol. 1, No. 2, pp. 224-227.

[11] Garai, G. and Chaudhuri, B.B.(2004), "A novel genetic algorithm for automatic clustering," *Pattern Recognition Letters*, Vol. 25, pp. 173-187.

[12] Halkidi, M., Batistakis, Y. and Vazirgiannis, M. (2001), "On clustering validation techniques," *Journal of Intelligent Information Systems*, Vol. 17, pp. 107-145.

[13] Hruschka E.R., Campello, R.G.B., Freitas, A.A. and Carvalho, A.P.L.(2009). "A survey of evolutionary algorithms for clustering," *IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews*, Vol. 39, No. 2, pp. 133-155.

[14] Kim, Y.K., Kim, J.Y. and Kim, Y.(2000), "A coevolutionary algorithm for balancing and sequencing in mixed model assembly lines," *Applied Intelligence*, Vol. 13, pp. 247-258.

[15] Kim, Y.K., Park, K.T. and Ko, J.S.(2003), "A symbiotic evolutionary algorithm for the integration of process planning and job shop scheduling," *Computers & Operations Research*, Vol. 30, pp. 1151-1171.

[16] Koontz, W.L.G., Narendra, P.M. and Fukunaga, K.(1975), "A branch and bound clustering algorithm," *IEEE Transactions on Computers,* Vol. 24, No. 9, pp. 908-915.

[17] Liu, G.(1968), Intoduction to combinatorial mathematics, NewYork: McGraw-Hill.

[18] Margulis, L.(1981), Symbiosis in cell evolution, W.H.Freeman, SanFrancisco.

[19] Maulik, U. and Bandyopadhyay, S.(2000), "Genetic algorithm-based clustering technique," *Pattern Recognition*, Vol. 33, pp. 1455-1465.

[20] Moriarty, D.E. and Miikkulainen, R. (1997), "Forming neural networks through efficient and adaptive coevolution," *Evolutionary Computation*, Vol. 5, pp. 373-399.

[21] Selim, S. and Alsultan, K.(1991), "A simulated annealing algorithm for the clustering problems," *Pattern Recognition*, Vol. 24, No. 10, pp. 1003-1008.
[22] Sung, C. and Jin, H.(2000), "A Tabu-search-based heuristic for clustering," *Pattern Recognition*, Vol. 33, pp. 849-858.
[23] Tou, J.T. and Gonazlez, R.C.(1974), Pattern Recog nition Principles, Addison-Wesley, Reading, MA.
[24] Theodoridis, S. and Koutroumbas, K.(2006), Pattern Recognition, 3rd Edition, Academic Press.
[25] Tseng, L.Y. and Yang, S.B.(2001), "A genetic approach to the automatic clustering problem," *Pattern Recognition*, Vol. 34, pp. 415-424.
[26] Xu, R. and Wunsch, D., II(2005), "Survey of clustering algorithms," *IEEE Transactions on Neural Networks*, Vol. 16, No. 3, pp. 645-678.

2011년 1월 7일 접수, 2011년 1월 27일 1차 수정, 2011년 1월 28일 채택