상대인력 모델에 기반한 자연적 개체 군집화 알고리즘

(A Natural Clustering Algorithm based on the Relative Gravitation Model)

김 은 주 † 고 재 필 † 변 혜 란 †† 이 일 병 ††
(Eunju Kim) (Jaepil Ko) (Hyeran Byun) (Yillbyung Lee)

요 약 본 논문에서는 상대인력 모델에 기반한 새로운 군집화 알고리즘, G-CLUS를 제안한다. 제한한 방법에서 모든 개체들은 초기에 동일한 질량을 가지고, 개체간의 인력에 의해 인력이 작용하는 방향으로 점진적으로 이동하게 되어, 초기 시작점 선택이나 군집의 개수를 미리 지정하지 않은 상태에서 자연스럽게 군집을 형성한다. 제안한 방법은 인력작용과정에서 군집의 수가 자연스럽게 결정되며, 한 개체가 받는 힘은 개체간의 인력을 합한 합력을 사용하기 때문에 이상치에 대한 민감성을 완화하였다. 본 알고리즘은 계산복잡도를 낮추기 위하여 큐브개념을 적용하여 $O(nk)$의 계산 복잡도를 유지하도록 하였다. 실험에서는, 개체들의 움직임 특성, 군집화 모델에 따른 군집화 과정, 임의의 데이타 집합에 대한 군집화 결과를 보이고, 또한 타 군집화 알고리즘과 제안한 알고리즘의 군집화 결과를 비교한다.

Abstract This paper propose a new clustering algorithm called G-CLUS based on the relative gravitation. In this method, every instance has the same mass at first, the gravitations among instances make each instance move to the attractive direction gradually and eventually natural clusters are formed without the initial seed and the number of clusters. Our proposed method can determine the number of clusters via a process of gravitational agglomeration and it can reduce the sensitivity to outliers by using the resultant of gravitation. We also improved the computational complexity by applying the concept of a cube to the proposed algorithm. In our experiments, we show the behavior of instance movement, clustering process for each model, clustering process and the results for an example data set, and the results of comparison between the other clustering algorithm and our proposed method.

1. 서 론

군집화란 소속 군이 주어져 있지 않은 데이타 집합을 필요에 의해 자연스러운 군집으로 나누는 것을 말한다. 군집화는 데이타 마이닝, 컴퓨터 비젼 등의 여러 분야에서 폭넓게 활용되고 있으며 또한 이러한 목적으로 사용가능한 알고리즘의 종류도 다양하다. 현재도 군집화 알고리즘에 대한 연구는 활발히 진행되고 있으며 새로운 알고리즘에 대한 연구 결과가 발표되고 있다.

군집화 알고리즘은 일반적으로 크게 계층적 군집화 방법과 최적화 기반 방법으로 분류된다[1]. 계층적 방법은 군집화 과정을 통하여 덴드로그램(Dendrogram)이라 불리는 트리 형태[2]의 결과를 형성하게 되는데, 이 트리를 생성하는 혹은 탐색하는 방향(상향식, 하향식)에 따라 다시 세부적으로 병합(agglomerative) 방법과 분할(divisive) 방법으로 나뉘어 진다.

계층적 병합 방법은 초기에 데이타 집합내의 N개의 데이타 각각을 하나의 군집으로 인정하여 N개의 군집을 생성한 후, 주어진 기준(single 혹은 complete)에 따라 차례로 묶어가면서 점진적으로 군집의 개수를 줄여 나가는 방법이다. 분할 방법은 역으로 초기에 전체 데이타 집합을 하나의 군집으로 설정하고, 점차 군집을 주어진 기준

† 비 회 원 : 연세대학교 컴퓨터과학과
outframe@csai.yonsei.ac.kr
nonezero@csai.yonsei.ac.kr
†† 종신회원 : 연세대학교 컴퓨터과학과 교수
hrbyun@aipiri.yonsei.ac.kr
yblee@csai.yonsei.ac.kr
논문접수 : 2000년 11월 14일
심사완료 : 2001년 8월 2일

에 따라 세부 군집으로 나누어가는 알고리즘이다[2-4].

이러한 계층적 군집 방법은 초기 군집의 수를 미리 정해줄 필요 없이 최종 생성된 트리를 사용자가 적정한 수준에서 자름으로써 군집의 수를 결정할 수 있다는 장점을 갖는다. 그러나 매 단계에서 군집을 병합하거나 분할할 때, 이웃에 위치한 근접 군집만을 고려하기 때문에 전체적인 데이타의 패턴을 반영하기 힘들다는 단점이 있다.

최적화 기반 방법은 일반적으로 분할(Partitional) 방법이라고도 불리는 데, 현재 널리 사용되는 많은 군집화 알고리즘들이 이 부류에 속한다. 이 방법의 특징은 초기 군집의 수를 미리 정해주면 정해진 목표 함수의 값을 최적화 시키기 위한 방향으로 매 단계마다 군집을 생성하고 변화시켜 나간다는 것이다[5-9]. 이 방법은 군집의 수를 미리 정해주어야 한다는 단점 이외에도 초기 설정(initialization)에 민감하고 국부최적치(local optima)에 빠질 수 있다는 단점을 갖는다. 대표적인 최적화 기반 방법으로는 K-Means, Self-Organizing Maps (SOM), Fuzzy C-Mean(FCM)등을 들 수 있다.

현재까지 발표된 많은 군집화 알고리즘들은 그 특성에 따라 서로 다른 장단점을 가지며 그 성능 또한 문제에 따라 다르다. 군집화라는 문제의 특성상 군집화 알고리즘들의 군집 성능을 비교하기 위한 절대적인 기준을 결정하기란 매우 어려운 일이다. 물론 군집화 이후 각 군집에 대한 분산이 작은 것으로 군집성능을 평가 할 수도 있지만, 시각적으로 확인할 수 있는 2-3차원의 데이타 집합인 경우, 사람이 판단하였을 때 자연스럽다는 느낌을 주는 것이 더 좋은 군집이라고 할 수 있을 것이다. 대부분의 기존 군집화 알고리즘들은 각 개체들을 군집이라는 인위적 목표를 가지고 개체들을 매 단계 조정하면서 군집화를 달성한다. 본 논문에서는 개체가 중심이 되어 군집화라는 목표 없이 인력 모델에 기반한 개체 조정 과정을 거치면서 그 부산물로서 군집이 이루어질 수 있는 방법을 모색한다.

자연계에는 이미 아주 오래 전부터 세상의 모든 개체들을 군집화 과정을 설명하는 중요한 원칙이 존재한다고 알려져 있다. 뉴턴에 의해서 형상화된 만유 인력의 법칙은 뉴턴의 저서인 <프린키피아>가 출판된 1687년부터 오늘날에 이르기까지 수많은 중력 현상들을 기술하는 훌륭한 이론으로 인정받고 있다. 이러한 만유 인력은 오늘날 지구상의 물질 형성 과정뿐만 아니라 우주 전체 및 지구의 형성 과정에 이르기까지 범용적으로 적용되는 자연계의 군집화 알고리즘이라고 할 수 있다. 개체들 간의 인력과 관련된 물리적 현상들을 응용하고자 하는 연구들은 Wright를 비롯한 몇몇 연구 결과들을 통해서도 확인할 수 있다[10-13].

Wright[10]는 공간 내에서 물체들간의 인력이 작용하여 서로를 끌어당김으로써 결과적으로 중앙에서 하나로 병합되는 자연적인 현상에 대하여 언급하였다. Forgy[11]는 잡음 제거를 위하여 간단한 인력 모델을 사용하였으며 Ravi[13]는 인력 모델을 심볼릭 데이타의 군집화에 응용하였다.

기존의 인력 모델에 기반한 군집화 연구들은 인력 모델을 응용에 맞게 정의하여 사용함으로써 일반적인 데이타들의 군집화에는 그대로 적용할 수 없으며 알고리즘의 특성상 데이타의 용량이 커질수록 시간이 오래 소요된다는 약점을 지니고 있어 범용적인 군집화 알고리즘으로 사용되기에는 불충분하였다. 또한 각각의 데이타들 관계에서 발생하는 방향의 정보를 충분히 활용하지 못하고 있었다.

본 논문에서는 만유인력이라는 자연계의 군집화 법칙에 기반한 상대인력 모델을 통하여 개체에 기반한 군집화를 달성함과 동시에 대용량 데이타의 효과적인 처리를 위하여 큐브 개념을 도입한 군집화 방법론을 새로이 제안하고자 한다.

2. 만유 인력과 클러스터링 알고리즘 사이의 관계성 연구

인력이란 공간적으로 떨어진 두 물체가 서로를 당기는 힘을 의미한다. 만유인력이란 우주 공간에 있는 모든 물체 사이에 작용하는 인력을 말하며 특히 지구와 지구상에 있는 모든 물체 사이에 작용하는 만유인력을 중력이라고 한다. 공간상에 위치한 두 물체의 질량이 각각 m_1, m_2라 할 때, 두 물체 사이에 존재하는 힘은 다음과 같이 정의된다[14][15].

$$F = G\frac{m_1 m_2}{r^2} \quad (1)$$

여기서 G는 만유인력상수이고, r 은 두 물체 사이의 거리이다.

현대 우주학에서 만유 인력은 전체 우주의 형태를 형성하는 데 있어 기본적인 원리로 작용하는 힘으로 인지되고 있다. 즉, 먼지와 같은 초기 우주의 형태가 만유인력에 의한 당김 현상으로 오늘날과 같은 군집화 된 행성의 형태로 변화하게 되었다는 것이다[15].

데이타의 군집이란 상대적으로 낮은 밀도를 갖는 구역(Region)에 의해 나뉘어지는 상대적으로 높은 밀도의 구역을 말하며, 이 때 이러한 형태의 군집들을 자연스러운 군집들(natural clusters)이라고 한다[16].

만유 인력은 기본적인 정의에 있어 질량에 비례하고 거리의 제곱에 반비례하는 힘이므로 만유 인력에 의하여 데이타들을 움직이게 되면 데이타들은 전체적인 데이타 공간 내에서 각 데이타에 인접한 밀도가 높은 무게 중심 쪽으로 움직이게 된다. 따라서 자연스럽게 지역적인 무게 중심을 발견할 수 있게 되므로, 만유인력 알고리즘을 활용한 군집화 알고리즘은 자연스러운 군집들을 찾을 수 있다.

3. 상대인력 모델에 기반한 클러스터링 알고리즘

만유인력모델에 의한 군집화에서는, 모든 개체들이 초기에 동일한 질량을 가지고 개체간의 거리의 제곱에 반비례하고 질량에 비례하는 힘으로 서로를 당겨 인력이 작용하는 방향으로 점진적으로 이동하여 자연스런 군집을 형성한다.

그러나, 어느 정도 군집화가 진행되어 커다란 군집들이 형성되면, (그림 1)과 같이 이들은 주위의 작은 군집보다 큰 군집을 더욱 강하게 당기게 된다. 이러한 성질은 데이타 군집화에는 바람직하지 않다. 또한 만유인력은 질량을 가진 개체 사이에서 힘이 소멸되지 않기 때문에 궁극적으로는 모든 개체가 하나의 군집을 형성하게 되는 결과를 낳는다.

본 장에서는 위와 같은 문제를 해결하기 위하여 새로운 인력모델 알고리즘을 위한 개념및 정의를 설명한다. 즉, 본 논문에서 정의한 변형된 만유인력인 상대인력(Relative Gravitational Force), 개체의 이동과 병합, 질량에 대한 감소함수를 갖도록 한 만유인력상수, 그리고 계산 복잡도를 줄이기 위한 방법을 설명한다.

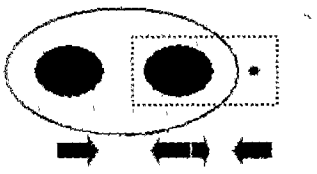


그림 1 개체간의 힘의 크기 비교

3.1 개념 및 정의

(1) 개체

개체 p는 위치와 질량을 가지는 것으로 아래의 수식으로 정의한다.

$$p_i\text{의 질량} : p_{im} \tag{2}$$

$$p_i\text{의 위치벡터} : p_{iv} \tag{3}$$

여기서, 개체의 위치벡터는 개체의 특징값(공간상의 위치)을 의미하며, 질량은 초기에 동일한 값을 갖는다고 가정한다.

(2) 상대인력 (RGF)

상대인력은 커다란 군집끼리 당기는 힘을 줄이고자 제안한 새로운 힘으로 다음과 같이 정의한다.

$$RGF = \frac{F}{m_1^2} = G\frac{m_2}{m_1} \qquad F = G\frac{m_2 m_1}{r^2} \tag{4}$$

여기서 m_1은 현재 인력을 고려중인 개체의 질량, m_2는 상대 개체의 질량, 그리고 G는 인력상수를 나타낸다. 수식(4)로부터, 주어진 개체에 대한 인력은 상대 개체의 질량에 비례하고 자신의 질량 및 두 개체간의 거리의 제곱에 반비례한다. 따라서, 커다란 개체 자신은 더 이상 주위 개체에 큰 인력을 발휘하지 못한다. 그러나, 작은 개체는 질량이 큰 상대 개체에 강한 인력을 가지게 된다.

(3) 개체 이동

인력에 의해 개체들은 힘을 받는 방향으로 자연스럽게 이동하게 된다. 이때 개체가 받는 힘은 주위의 모든 개체로부터 받는 합력으로 이동 방향 및 크기를 결정한다. 두 개체 p_i, p_j 에 대한 만유인력은 수식(5)와 같다.

$$F = G\frac{p_{im}p_{jm}}{|p_{iv} - p_{jv}|^2} \tag{5}$$

개체 p_j에 의해 개체 p_i가 받는 상대인력은 수식(6)과 같고, 주위의 N개 개체로부터 받은 힘들의 합력에 의한 개체 p_i의 새로운 위치벡터는 수식(7)과 같다.

$$RGF_j^i = G\frac{p_{jm}}{|p_{iv} - p_{jv}|^2 p_{im}} \tag{6}$$

$$p_{iv} = p_{iv} + \sum_{k=(N)} u_k^i RGF_k^i \tag{7}$$

수식(7)에서 u_k^i는 p_i 에서 p_k 방향으로의 방향 단위벡터를 나타낸다.

(그림 2)는 어떤 주어진 개체가 받는 힘의 크기 및 방향을 보여 준다. k번째의 p_i가 p_n 및 p_m에 의해 받는 힘 RGF_n^i , RGF_m^i에 의해 $k+1$번째의 p_i방향으로 힘을 받게 된다.

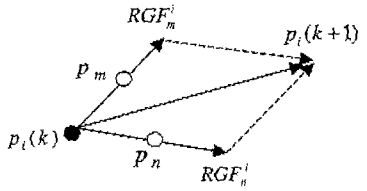


그림 2 개체가 받는 힘의 크기 및 방향

(4) 개체 병합

개체들이 상대인력에 의해 움직이고 난 후, 경우에 따라서 개체간의 거리가 단위거리 즉 1보다 작아 질 수 있다. 이때 우리는 개체간의 거리가 1이내인 모든 개체들을 하나의 개체로 병합하여 새로운 개체를 생성하고 다음과 같이 그 특성을 할당한다.

(a) 생성되는 새로운 개체의 위치 = 병합되는 모든 개체의 평균 위치

(b) 생성되는 새로운 개체의 질량 = 병합되는 모든 개체의 총 질량

병합과정을 통해 첫째, 시간이 지남에 따라 개체가 줄어들어 계산량이 줄어드는 잇점을 얻을 수 있고 둘째, 개체들의 움직임과 이들의 병합과정에서 최종적으로 살아남은 개체가 결국 하나의 군집을 나타내게 되어 자연스럽게 군집화 결과를 얻을 수 있다.

(5) 인력상수

인력모델에 의한 군집화에서는 결국 인력이 소멸되지 않는 한 궁극적으로 모든 개체가 하나의 군집을 형성하게 된다. 이 문제를 해결하기 위해 우리는 인력상수를 더 이상 상수가 아닌 변수로 취급한다. 즉, 인력상수 G를 (그림 3)과 같은 질량에 따른 감소함수로 정의한다.

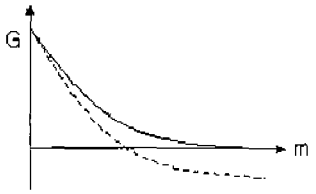


그림 3 개체의 질량에 따른 인력상수

그림에서 실선은 본 논문에서 실제로 사용한 감소함수이며, 점선은 대체하여 사용할 수 있는 함수를 나타내고 있다. 음수로까지 값이 떨어지는 것은 어느 정도 개체가 성장하였을 때 더 이상 인력이 아니라 척력이 작용하는 것을 의미한다.

인력상수를 감소함수로 정의하여 사용함으로서 개체가 커짐에 따라 상대인력이 감소하게 되며 궁극적으로는 힘이 소멸하게 되는 효과를 얻을 수 있다. 따라서, 일정 수준 이상의 무거운 개체는 더 이상 힘을 받지 않게 되어 어느 방향으로도 이끌리지 않게 된다. 그렇다고 일정수준 이상의 개체가 더 이상 성장하지 않는다는 것은 아니다. 왜냐하면, 주위의 작은 개체는 큰 개체에 계속해서 이끌림을 받기 때문이다.

결국, 인력상수를 감소함수로 정의함으로서 일정 수준 이상의 무거운 개체들은 성장이 둔화되고 궁극적으로는 성장이 멈추게 된다. 이를 통해 우리는 군집의 개수를 알고리즘 수행이전에 미리 정하지 않더라도, 성장이 멈춘 개체들의 수로서 자연스럽게 군집의 개수가 정해지는 효과를 얻을 수 있다.

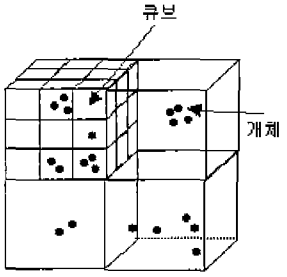


그림 4 3차원 큐브 예

(6) 계산복잡도를 줄이기 위한 큐브(Cube)방식 제안

어떤 개체가 받는 힘은 주위의 모든 개체에 의해 받는 힘들의 합력으로 계산된다. 이때 개체의 수가 n개인 데이타 집합이라면, 계산복잡도는 $O(n^2)$이 된다. 이는 대용량 데이타 집합에 대한 군집화 알고리즘으로서 치명적인 단점이 될 수 있다.

이를 보완하고자 제안한 알고리즘에 큐브개념을 적용하면 계산복잡도를 줄일 수 있다. 데이타 공간을 같은 크기의 공간, 즉 (그림4)와 같이 큐브로 분할한다. 이렇게 공간을 세분화함으로써 주어진 개체에 영향을 주는 주변 개체의 범위를 이웃하는 큐브에 속하는 개체들로 한정할 수 있다. 이웃하는 큐브에 들어갈 수 있는 개체의 최대 수가 k ($k < n$)개라면, 계산복잡도는 $O(kn)$으로 줄게 된다.

3.2 새로운 군집화 알고리즘 (G-CLUS)

제안하는 군집화 알고리즘의 주요 부분은 모든 개체들의 상대인력을 구하는 것이다. 상대인력에 따라 각 개체들은 새로운 위치로 이동하게 되고 이동의 결과로 서로간의 거리가 단위거리 이내로 들어오는 개체들이 생겨나게 된다. 그런 다음, 그들을 하나의 개체로 병합한다. 이 과정, 즉 상대인력에 의한 이동과 병합을 반복하면, 더 이상 개체들이 이동하지 않게 되는 평형상태에 도달할 수 있다. 이때, 마지막까지 살아남은 개체들 각각이 하나의 군집을 형성하게 된다.

제안하는 알고리즘은 다음과 같은 단계로 구성된다.

G-CLUS 알고리즘

1. 데이타 집합 내의 모든 개체의 질량을 1로 초기화 한다
2. 거리가 1이내인 개체들을 하나의 개체로 병합한다
3. 각 개체가 데이타 집합 내의 다른 개체에 의해 받는 상대 인력, RGF를 계산한다
4. RGF의 합력을 계산한다
5. 모든 개체에 대해 단계3-4를 반복한다
6. 각 개체의 합력에 따라 개체의 위치벡터를 갱신한다
7. 평형상태에 도달할 때까지 단계 2-6을 반복한다

4. 실험결과

이 장에서는 시간별 개체의 움직임크기(속력) 변화를 살펴보고, 2차원 데이타에 대하여 제안한 군집화 알고리즘의 군집화 과정 및 결과를 보여준다. 그리고, 대표적 군집화 알고리즘인 K-Means 군집화 알고리즘과의 비교결과를 제시한다.

4.1 개체이동 행위

(그림 5)는 질량이 1인 두 개체가 서로의 인력으로 인해 움직여 갈 때, 시간별 움직임 단위 변화를 보여준다. 여기서 움직임 단위는 1단위 시간동안 개체가 움직인 거리 즉, 속력을 말한다. 그림을 보면 시간이 지남에 따라 움직임 단위가 급격하게 커지는 것을 볼 수 있다. 이는 질량이 같은 두 개체가 가까워짐에 따라 둘 사이의 인력이 급격하게 커지기 때문이다. 다시 말하면, 근접한 개체들끼리의 병합이 빠르게 이루어진다는 것을 말한다.

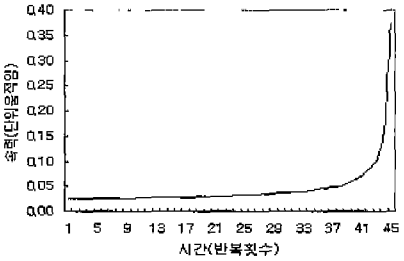


그림 5 질량1인 개체의 속력

(그림 6)은 질량이 각각 1과 5인 개체의 시간별 움직임 단위의 변화를 보여준다. 그림에서 알수 있듯이, 질량이 1인 개체는 질량이 5인 개체에 비해 월등히 움직임 단위가 커진다. 이 결과로 볼 때, 상대개체의 질량에 비해 자신의 질량이 큰 개체는 움직임이 둔화되고 상대적으로 주위의 작은 개체들을 강하게 끌어당긴다는 것을 알 수 있다. 이는 상대인력을 통해 제안한 알고리즘에서 얻고자 하는 개체의 이동특성이다.

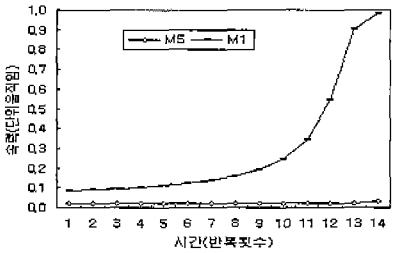


그림 6 질량1과 5인 개체의 속력

4.2 인력모델별 군집화 과정 비교

(그림 7)에 제시한 임의의 2차원 데이타 집합에 대하여 만유인력 모델, 상대인력 모델, 인력상수변화 모델에 대한 군집화 과정을 보이고 최종 제안 모델인 인력상수 변화모델의 우위성을 보여준다.

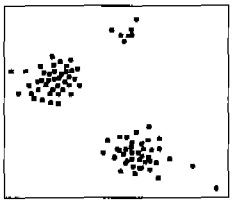


그림 7 실험 데이타 집합

(1) 만유인력 모델에 기반한 군집화 과정

만유인력 모델에 의한 시간별 군집화 과정을 그림8에 보인다. 시간이 지남에 따라 두 커다란 개체가 주위의 작은 개체를 흡수하기에 앞서 그들이 먼저 병합되는 것을 볼 수 있다. 이것은 자연계의 자연스런 인력의 결과이지만, 데이타 군집화에 있어서는 바람직하지 못한 결과를 초래한다. 또한 궁극적으로는 모든 개체가 하나의 개체로 통합되어서야 비로소 평형상태를 유지한다.

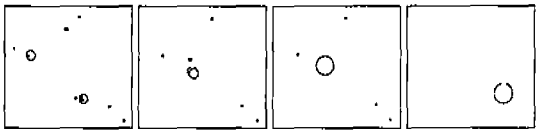


그림 8 만유인력 모델에 기반한 군집화 과정

(2) 상대인력(RGF) 모델에 기반한 군집화 과정

(그림 9)는 상대인력 모델에 의한 군집화 과정을 보여준다. 상대인력 모델의 경우, 만유인력모델에서와 같이 커다란 개체끼리의 병합이 주위의 작은 개체와의 병

합보다 선행되지 않는다. 그러나 이 모델 역시 궁극적으로는 인력이 소멸되지 않는 관계로 모든 개체가 하나의 개체로 통합되는 결과를 보인다.

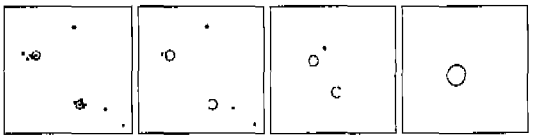


그림 9 상대인력 모델에 기반한 군집화 과정

(3) 가변인력상수를 적용한 상대인력 모델에 기반한 군집화 과정 (제안모델)

상대인력 모델에 인력상수를 변수로 변화시킨다. 이때 인력상수는 질량에 대한 감소함수로 가변적인 값을 적용한 모델이다. 군집화 과정은 (그림 10)과 같다. 그림에서 볼 수 있듯이, 본 모델은 위 두 모델에서 보이는 단점을 극복하여, 큰 개체끼리의 병합을 완화함은 물론, 최종 개체의 수도 군집화 과정을 통해 자연스럽게 결정될 수 있도록 하고 있다. 이는 군집수행 이전에 군집의 개수를 미리 정해주어야 하는 K-Means나 SOM과 같은 다른 군집화 알고리즘에 비해 장점으로 부각될 수 있다.

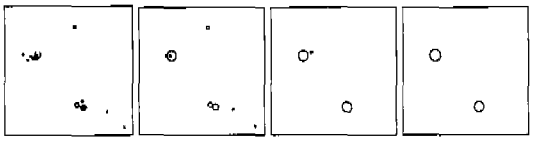
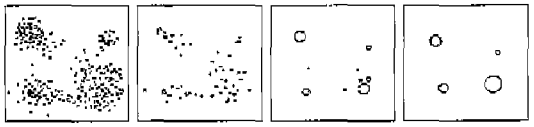


그림 10 가변인력상수를 적용한 상대인력 모델의 군집화 과정

4.2 제안모델의 군집화 과정 및 군집결과

(그림 11)은 4개의 큰 덩어리를 가지고 있는 개체들을 제안한 알고리즘에 의해 군집화를 수행하는 과정을 보여준다. 시간이 지남에 따라 4개의 커다란 개체들이 형성되는 것을 볼 수 있다. 이때 각 개체들은 각각의 군집을 의미한다. 최종 군집결과를 (그림 12)에 보인다.



(a) (b) (c) (d)

그림 11 군집화 과정 : (a) 초기개체, (b)(c) 개체이동 및 병합과정, (d) 최종생존개체

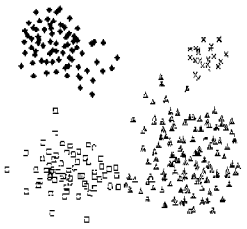


그림 12 최종 군집화 결과

4.4 제안한 알고리즘(G-CLUS) 과 타 군집화 알고리즘의 군집 결과

제안하는 알고리즘과 타 군집화 알고리즘과의 비교를 위해 임의로 생성한 데이타 집합에 대하여 KMeans 군집화 알고리즘, FCM 군집화 알고리즘, 완전결합 및 평균결합 계층적 군집화 알고리즘과의 비교 실험을 수행하였다. 비교를 위해 각 알고리즘의 최종 군집의 개수를 2개로 지정하였다. (그림 13) 및 [표 1]에서 제안하는 알고리즘이 보다 자연스러운 군집을 형성한 것을 확인할 수 있다. 표 1에서 C1 및 C2는 각각의 군집을 나타낸다. [표 2]는 (그림 12)의 데이타 집합에 대한 표준편차 비교 결과로서 제안하는 알고리즘이 가장 낮은 표준편차를 보인다.

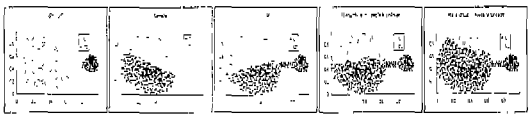


그림 13 군집화 알고리즘별 군집 결과

표 1 그림 13 데이타 집합의 군집 표준편차 비교

구분	C1	C2	Total(C1+C2)
G-CLUS	0.4051	0.1197	**0.5249**
KMeans	0.4068	0.3055	**0.7124**
FCM	0.3859	0.3049	**0.6867**
Hierarchical-Complete	0.3905	0.2796	**0.6702**
Hierarchical Average	0.4272	0.1860	**0.6133**

표 2 그림 12 데이타 집합의 군집 표준편차 비교

구분	C1	C2	C3	C4	Total (C1+C2+C3+C4)
G CLUS	0.1858	0.2579	0.0967	0.1878	**0.7284**
KMeans	0.1567	0.2033	0.3205	0.1279	**0.8086**
FCM	0.1858	0.1970	0.2482	0.2137	**0.8449**
Hierarchical-Complete	0.1858	0.2462	0.2673	0.1778	**0.8772**
Hierarchical-Average	0.1858	0.2654	0.2527	0.1722	**0.8762**

5. 결론 및 향후 연구 계획

만유인력은 우주의 형성에 기여하는 자연계의 기본적인 힘으로 간주된다. 본 논문을 통하여 자연계의 군집화 법칙을 데이타의 군집화 과정에 적용하는 새로운 방법론으로서 상대만유인력 모델에 기반한 새로운 군집화 알고리즘을 제안하였다. 또한 실험결과를 통하여 본 방법을 적용한 군집화가 데이타들의 분포를 반영하여 자연스럽게 진행되어 가는 과정을 확인 할 수 있었다.

이 방법의 장점은 군집화가 개체관점으로 개체들의 자연스런 운동결과로 주어지는 부산물로서 군집화가 이루어지는 다소 자연스런 방법이라는 것 이외에도 최종 군집의 수가 초기에 인위적으로 주어지는 것이 아니라 자연적인 힘의 평형 현상에 의해 자동으로 결정된다는 것을 들 수 있다. 이것은 대용량의 데이타나 혹은 다차원의 데이타를 다루는 많은 응용분야에서 매우 중요한 장점으로 부각될 수 있다. 데이타의 양이나 차원이 높을수록 인위적으로 적절한 군집의 수를 미리 결정하는 것은 매우 어려운 문제이기 때문이다. 또한 제안한 알고리즘은 계산 복잡도에 있어서도 큐브개념을 적용하여 $O(kn)$를 달성할 수 있었다.

실험에서는 정의한 상대인력에 의한 개체들의 움직임, 인력 모델별 군집화 과정 및 결과를 통해 본 알고리즘이 추구하고 있는 자연스런 군집화 특성을 보여주었으며, 또한 타 군집화 알고리즘과의 비교 결과를 제시하여 군집내의 표준편차 측면에서도 제안한 알고리즘이 우수하다는 것을 보여주었다.

향후 연구에서는 실제 데이타 집합에 제안한 방법을 적용한 실험과, 이 과정을 통해 제안한 인력모델의 문제점을 발견하고 수정해 나갈 계획이다.

참 고 문 헌

[1] A.K. Jain and R.C. Dubes, Algorithms for Clustering Data. Englewood Cliffs, N.J.: Prentice Hall, 1988.

[2] B.D. Ripley, Pattern Recognition and Neural Networks, Cambridge, 1996.

[3] R. O. Duda and P. E. Hart, Pattern Classification and Scene Analysis, New York: Wiley, 1973.

[4] H. H. Bock, Automatic Classification, Vandenhoeck and Ruprecht, Göttingen, 1974.

[5] K. Fukunga, Introduction to Statistical Pattern Recognition, San Diego, CA, Academic Press, 1990.

[6] S.L. Lauritzen, "The EM algorithm for graphical association models with missing data, Computational Statistics and Data Analysis," pp.191-201, 1995.

[7] R.T. Ng, J. Han, "Efficient and Effective Clustering Methods for Spatial Data Mining," Proc. 20th Int. Conf. on Very Large Data Bases, pp. 144-155, 1994.

[8] T. Kohonen, K. Makisara, O.Simula and J. Kangas, Artificial Networks, Amsterdam, 1991.

[9] S. Haykin, Ne-ural Networks - A Comprehensive Foundation, Prentice Hall, 1999.

[10] W. E. Wright, "Gravitational Clustering," Pattern Recognition vol. 9, no. 3, pp. 151-166, 1977.

[11] P.H.A. Sneath, "A method for curve seeking from scattered points," Computer. J., Vol. 8, pp. 383-391, 1966.

[12] E. W. Forgy, "Evaluation of several methods of detection sample mixtures from different N-dimensional populations," American. Psych. Assoc., Los Angeles, CA, 1965.

[13] T.V. Ravi and K. Chidananda Gowda, "Clustering of Symbolic Objects Using Gravitational Approach," IEEE Transactions on Systems, Man, And Cybernetics-Part B: Cybernetics, Vol. 29, No. 6, December 1999.

[14] A. P. Tipler, Physics, Worth Publishers, 1976.

[15] F. J. Hawley & K.A. Holcomb, Foundations of Modern Cosmology, Oxford University Press, 1998.

[16] B. Everitt, Cluster Analysis, 2nd ed., Halsted Press, 1981.

김 은 주
1994년 연세대학교 전산과학과 학사 졸. 1996년 연세대학교 전산과학과 석사 졸. 1996년 ~ 현재 연세대 컴퓨터과학과 박사과정. 관심분야는 데이타마이닝, CRM, 패턴인식, 신경망

고 재 필
1996년 연세대 전산학과 학사 졸. 1998년 연세대 컴퓨터과학과 석사 졸. 1999년 ~ 현재 연세대 컴퓨터과학과 박사과정. 관심분야는 영상이해, 패턴인식, 영상처리

변 혜 란
정보과학회논문지 : 소프트웨어 및 응용 제 28 권 제 3 호 참조

이 일 병
정보과학회논문지 : 소프트웨어 및 응용 제 28 권 제 6 호 참조