만유인력에 기반한 자연적 개체 군집화

김은주° , 고재필, 변혜란, 이일병
연세대학교 컴퓨터 과학과

A Natural Clustering of Instances Based On Universial Gravity

Eunju Kim, Jaephil Ko, Hyeran Byun, Yillbyung Lee
Dept. of Computer Science, Yonsei University

요 약

현존하는 다양한 군집화 알고리즘들이 개체들을 군집화하기 위하여 사용하는 기준들은 일반적으로 인위적으로 설정된 것들이다. 이러한 기준들은 개체들 자체로부터 나오는 자연스러운 기준이라기 보다는 군집을 위하여 임의로 선정된 것이므로 군집화의 기본 목적인 개체들을 자연스러운 그룹들로 분할하고자 하는데 있어 한계를 갖게 된다. 본 논문에서는 이러한 점에 주목하여 현존하는 자연계의 군집 법칙으로 대표되는 만유인력의 법칙을 사용한 개체 군집화 알고리즘을 제안함으로써 기본적인 목적에 충실한 군집화를 실현하고자 한다. 이 방법은 기존의 방법론들에서 찾아 볼 수 없었던 자연 법칙에 근거한 새로운 군집화 시도일 뿐만 아니라, 초기 조건에 관계없이 안정적인 성능을 보이고 또한 군집의 수가 자연 법칙에 따라 자동으로 결정되는 특성을 지니고 있어 다양한 실질적인 응용 분야에서 효과적으로 사용될 수 있는 새로운 군집화 도구가 될 수 있을 것으로 보인다.

1. 서 론

군집화 기법들은 각 개체들에게 미리 예측된 클래스가 없을 때 그 개체들을 자연스러운 그룹들로 분할하고자 할 때 사용된다[1]. 이러한 목적으로 사용되는 군집화 알고리즘들의 종류는 매우 다양하며 그들 각각은 기본적인 아이디어에 따라 서로 다른 장단점을 띄게 된다.

군집화의 목적으로 가장 일반적으로 많이 사용되는 방법들로는 K-Means 알고리즘, SOM(Self-Organizing Map), 계층적 군집화 방법(Hierarchical Clustering:최장연결, 최단 연결, 중심연결, 평균연결법)등을 들 수 있다[2][3][4].

이들 방법들은 군집화를 목표로 각기 서로 다른 기준들을 적용하여 개체들을 분할해 나간다 그러나 이러한 기준들은 설정 당시부터 군집 생성이라는 목표에 기반하여 만들어진 인위적인 것으로 개체들을 자연스러운 방법으로 묶어 나가기 보다는 기준에 근거한 강압적인 방식을 취하는 부자연스러운 특성을 띤다. 즉 개체들의 특성이나 전체적인 분포, 개체 상호간의 연관성으로부터 나오는 자연스러운 군집화 과정이라기 보다는 주어진 기준에 맞추거나 최적화 하고자 하는 목표값에 맞추어 데이터들을 군집화하게 된다. SOM과 같은 군집화 기법은 인간의 뇌구조를 모방한 신경회로망 모형의 일종이므로 어느 정도 자연스러운 군집화 방법을 표방한다고도 할 수 있다. 그러나 이 방법 또한 개체들의 자연스러운 분포로부터 군집화가 진행되어가기 보다는 고정된 지도의 구조를 데이터의 위상 구조를 반영하도록 변화시켜 나가고자 하는 인위성을 지니고 있다.

자연계에는 이미 아주 오래전부터 세상의 모든 개체들을 군집화하는 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 과정 뿐만 아니라 우주 전체 및 지구의 형성 과정에 이르기까지에 범용적으로 적용되는 자연계의 군집화 알고리즘이라고 할 수 있다.

따라서 본 논문에서는 만유인력이라는 자연계의 군집화 법칙에 기반하여 개체들이 서로에게 상호 영향을 미치면서 운동함으로써 자연스럽게 군집화되는 매우 새로운 방법론을 제안하고자 한다.

2. 만류인력

인력이란 공간적으로 떨어져 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[5].

공간상에 위치한 두 물체의 질량이 각각 M, n라 할 때, 두 물체 사이에 존재하는 힘은 다음과 같이 정의된다.

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

여기서 G는 만유인력상수이고, r은 두 물체 사이의 거리이다.

이러한 만유 인력은 지상에서의 물체의 운동이나 태양계 내의 행성의 운동 등을 결정할 뿐만 아니라, 별, 은하, 더 나아가 우주 전체와 같은 물체들의 구조를 결정하는 데 가장 중요한 역할을 하게 된다.

3. 만유인력모델을 통한 군집화 알고리즘

수식(1)에 의하면 인력은 물체의 질량이 증가함에 따라 커지며, 거리가 가까워짐에 따라 급속도로 커지게 된다. 이때 힘과 가속도와의 관계식(2)에서 가속도 a는 질량 m과 반비례하며, 가속도는 속도의 변화량으로 움직인 거리에 비례한다.

$$F = ma \quad (2)$$

본 장에서는 이러한 자연법칙을 모델링하여 밀도가 높은 지역으로 개체들을 군집화하는 새로운 알고리즘을 설명한다.

[정의]

1 개체정의

개체는 질량과 공간상의 위치를 갖는 것으로 정의한다

개체 p_i의 질량 : p_m^i, 개체 p_i의 위치벡터 : p_v^i

2 개체 p_i, p_j 간에 작용하는 힘 정의

$$F_j^i = G\frac{p_m^i p_m^j}{|p_v^j - p_v^i|^2} \quad (3)$$

3 개체 p_j에 의한 개체 p_i의 움직임 단위를 결정하기 위한 힘 정의

$$Fd_j^i = G\frac{p_m^j}{|p_v^j - p_v^i|^2 p_m^i} \quad (4)$$

4 영향을 주는 다른 개체들에 의한 개체 p_i 의 이동방향벡터 정의

$$p_d^i = \sum_k u_k^i Fd_k^i \quad (5)$$

여기서 u_k^i는 개체 p_i, p_k사이의 단위벡터이다

5 영향을 주는 다른 개체들에 의한 개체 p_i 의 이동거리 정의

$$p_s^i = \frac{1}{1+e^{-\frac{8(|p_d^i|-1)}{G}+4}} \quad (6)$$

위에서 정의한 수식에 의한 이동 방향은 그림 1에 도식화하였다. 같은 거리에 위치하는 질량이 다른 개체는 상대방의 질량에 따라 서로 다른 크기를 가지며, 이들의 합벡터로 최종 이동방향을 결정할 수 있다.

개체가 힘을 받아 움직인 이동거리는 이동방향벡터의 크기(힘)를

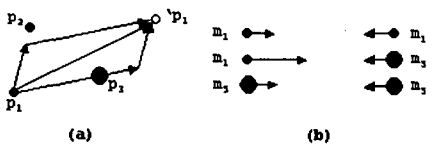

그림 1. (a) 개체 p1의 이동방향벡터, (b) 질량에 따라 받는 힘의 크기비교

시그모이드 함수를 이용해 0-1사이의 값으로 변환하여 최대 단위벡터 크기만큼 이동할 수 있도록 수식(6)을 정의하였다[그림2].

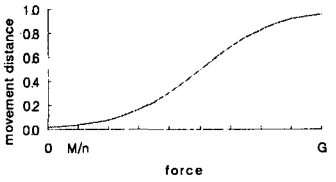

그림 2. 시그모이드함수를 이용한 이동거리 변환

[군집화 알고리즘]

군집화 알고리즘은 크게 병합과 이동으로 다음 과정을 따른다

1 모든 개체의 질량을 $p_m^i = 1$로 초기화한다

2 개체 p_i에 대해 거리가 1 이하인 모든 $p_{j\neq i}$에 대하여 다음 수식, $p_m^i = p_m^i + p_m^j$,으로 질량을 갱신하고, 개체 $p_{j\neq i}$는 소멸시킨다. 이를 모든 개체에 대해 반복한다(병합)

3 개체 p_i에 대해 모든 $p_{j\neq i}$를 선택하여 모든 F_j^i를 계산한다

4 $F_j^i > \theta$인 모든 p_j에 대해서 Fd_j^i를 계산한 후 p_d^i를 계산한다

5 p_v^i를 p_d^i방향으로 길이 p_s^i만큼 이동한다(이동)

6 데이터 집합에 포함된 모든 개체에 대해 단계2를 반복한다

7 이때 모든 개체에 대해 $F_j^i < \theta$면 군집화를 종료한다

4. 실 험

실험은 100x100 2차원 공간에 존재하는 개체들을 대상으로 하여 본 논문에서 제안한 방법의 특징 및 군집화 과정을 보였다

4.1 질량 및 거리에 따른 움직임단위 비교

그림 3은 질량이 1로 동일하고 거리가 5만큼 떨어져 있는 두 개의 개체에 대한 매회 이동단위를 보여준다. 45회에 병합이 이루어 졌으며, 두 개체가 근접할수록 급격하게 움직임 단위가 커지는 것을 확인할 수 있다.

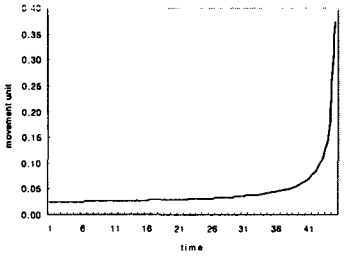

그림 3. 거리가 5이고 질량이 1인 두 개체간의 시간별 이동 단위

그림4는 질량이 각각 1과 5이고 거리가 5인 개체의 매회 이동단위를 보여준다. 14회만에 병합이 이루어 졌으며, 근접할수록 움직임 단

위가 급격히 커지며, 질량이 5인 경우 질량이 1인 개체에 비해 움직임 단위가 크다.

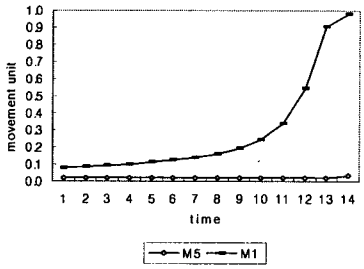

그림 4. 거리가 5이고 질량이 1과 5인 두 개체(M1, M5)의 시간별 이동단위 비교

그림 3과 4에서 확인할 수 있는 두 개체 간의 거리가 가까워질수록 움직이는 속도가 빨라지는 현상은 만유인력의 크기가 거리의 제곱에 반비례하는 속성에 의하여 발생하게 된다.

본 실험 결과를 통해 본 논문에서 제안한 군집화 알고리즘이 만유인력법칙에 의한 개체들의 움직임을 특성을 잘 반영하고 있음을 확인할 수 있다.

4.2 군집화 과정

그림 5는 군집화 단계를 시각적으로 보여준다. 시간이 지남에 따라 개체 집합의 분포에 의해 4개의 중심점으로 개체들이 자연스럽게 모이는 것을 볼 수 있다.

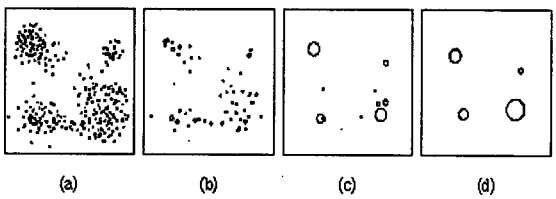

그림 5. 시간별 군집화 과정,
(a)초기 개체집합, (b)(c)군집화 진행 과정, (d)최종 군집결과

그림6은 그림5의 군집결과를 원 개체집합에 대하여 출력한 것이다. 각 개체들이 4개의 군집으로 적절히 군집화 되어 있는 것을 알 수 있다.

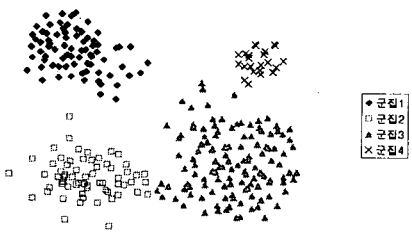

그림 6. 군집된 초기 개체집합 표현

5. 결 론

우리는 본 논문을 통하여 자연계의 군집화 법칙을 데이터의 군집화 과정에 적용하는 새로운 방법론을 제시하였다. 또한 실험 결과를 통하여 본 방법을 적용한 군집화 과정이 데이터들의 분포를 반영하여 자연스럽게 진행되어어 가는 과정을 확인 할 수 있었다.

이 방법의 장점은 가장 자연스러운 군집화 과정이라는 것 이외에도 최종 군집의 수가 초기에 인위적으로 주어지는 것이 아니라 자연적인 힘의 평형현상에 의해 자동으로 결정된다는 것을 들 수 있다. 이것은 대용량의 데이터나 혹은 다차원의 데이터를 다루는 많은 응용분야에서 매우 중요한 장점으로 부각될 수 있다. 데이터의 양이나 차원이 높을수록 인위적으로 적절한 군집의 수를 미리 결정한 다는 것은 매우 어려운 문제이기 때문이다.

본 연구 결과를 통하여 우리는 제안한 방법의 가능성과 효용성을 검증할 수 있었으며, 이러한 결과에 힘입어 제안하는 방법의 실용성을 높이기 위한 다음과 같은 향후 연구 계획을 세우게 되었다.

첫 번째는 계산량 축소에 관한 연구이다. 제안한 알고리즘은 데이터 각각의 개체에 기반한 방법이기 때문에 그 계산량에 있어 계층적 군집화 방법과 같은 $O(n^2)$의 계산량을 갖게 된다. 따라서 이와 같은 문제를 해결하기 위한 방법론 튜닝에 관한 연구를 수행할 것이다.

두 번째로는 방법의 응용 분야 확대를 들 수 있다. 클러스터링을 원하는 많은 분야에 현존하는 데이터들은 일반적으로 다차원 데이터의 성격을 지닌다. 따라서 본 방법을 여러 분야에 존재하는 다차원 데이터에 확대 적용함으로써 그 응용 분야를 실질적인 분야로 확대해 나갈 계획이다.

참고문헌

[1] Ian H.Witten, Eibe Frank, Data Mining, Morgan Kaugmann, 2000.

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

[3] R. Rojas, Neural Networks - A Systematic Introduction, Springer Berlin, 1996.

[4] H.H. Bock, Antmatic Classification, Vandenhoeck and Ruprecht, Gottingen, 1974.

[5] Paul A. Tipler, Physics, Worth Publishers, 1976.