다계층 밀도기반 군집화 기법

신동문, 정석호, 이경민, 이동규, 손교용, 류근호
충북대학교 데이터베이스/바이오인포매틱스 연구실

e-mail : {mastershin216, sukhojung, min9709, dglee, gysohn, khryu}@dblab.chungbuk.ac.kr

Multi-hierarchical Density-based Clustering Method

Dong Mun Shin, Suk Ho Jung, Gyeong Min Yi, Dong Gyu Lee,
Sohn, GyoYong, Keun Ho Ryu

Database/Bioinformatics Laboratory, Chungbuk National University

요 약

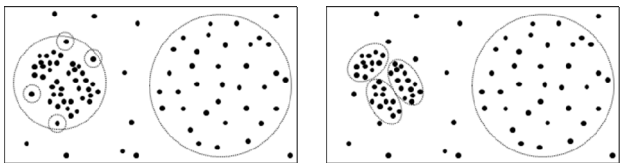
군집화는 대용량의 데이터로부터 유용한 정보를 추출하는 데에 적합한 데이터마이닝 기법들 중 하나이다. 군집화 기법은 주어진 데이터그룹 내에서 사전정보 없이 의미있는 지식을 발견할 수 있으므로 큰 어려움이 없이 실제 응용분야에 적용할 수 있다. 또한, 대용량 데이터를 다룰 때에 개별적인 데이터에 대한 접근 횟수를 줄이고, 알고리즘이 다루어야 할 데이터 구조의 크기를 줄일 수 있다. 본 논문에서는 밀도-기반 군집화 기법을 기반으로 하는 새로운 군집화 기법을 제안한다. 우리가 제안하는 군집화 기법은 반복적인 군집화 과정을 통하여 군집 내 주변 잡음을 제거하고 더 세밀하게 집단을 세분화하는 것이 가능하다. 또한, 군집을 표현하는 데에 계층구조로 나타내어 각 군집의 상관관계를 파악하는 데에 유리하다. 본 논문에서 제안하는 군집화 기법을 통하여 다양한 밀도를 가진 군집들을 효과적으로 분류할 수 있을 거라고 기대된다.

1. 서론

군집화는 대용량의 데이터베이스에서 지식을 발견하는 중요한 데이터마이닝 기법들 중에 하나이다. 군집화 기법은 유사한 특성을 가진 객체들을 각각의 다른 그룹으로 분류하는 것을 말한다.[1]

군집화 기법이 데이터마이닝의 중요한 기술로서 활발히 응용되고 있는 이유는 타 기법들과는 차별되는 몇 가지 강점을 가지고 있기 때문이다. 첫 번째, 대부분의 군집방법은 주어진 데이터 그룹 내에서 사전적인 정보 없이 의미있는 지식을 발견할 수 있다. 두 번째, 대용량 데이터를 다룰 때에 개별적인 데이터에 대한 접근 횟수를 줄이고, 알고리즘이 다루어야 할 데이터 구조의 크기를 줄일 수 있다. 이러한 강점들 때문에 다방면의 분야에 효율적인 응용이 가능하다.

밀도-기반 군집화는 기존의 다른 알고리즘 기법들과는 달리, 다양한 형태의 군집들을 발견할 수 있고, 저밀도 지역을 형성하는 잡음을 제거할 수 있다[1,2]. 하지만 다른 밀도를 가진 군집들이 존재할 때, 만족스러운 군집화 결과를 획득하지 못하는 문제점을 안고 있다.



(a) 밀도-기반 군집화 결과 (b) 새 군집 생성

(그림 1) 서로 밀도를 가지는 군집들

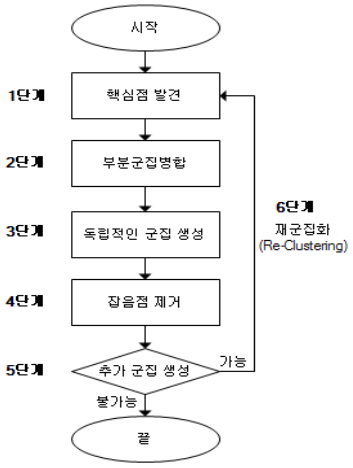
그림 1 은 고밀도 영역과 저밀도 영역으로 구분되어지는 데이터 집합에 대해 군집화한 것이다. 밀도-기반 방법에 의해 군집화를 실행하여 그림 1(a) 에서 보듯이 2 개의 군집이 생성하였다. 상대적으로 고밀도인 군집 지역은 잡음을 포함하여 하나의 군집이 형성되었다. 만약 이 군집에 대해서 또 한번의 군집화를 한다면, 군집 내 주변 잡음을 제거하고 그림 1(b)와 같이 3 개의 군집을 추가 생성할 수 있다. 즉, 반복적인 군집화 과정을 통하여 더 정확하고 세밀하게 집단을 세분화하는 것이 가능하다. 그리하여 대용량의 데이터에 대하여 새로운 패턴을 쉽게 발견할 수 있다. 또한, 군집을 표현하는 데에 계층구조로 나타내어 각 군집의 성격이나 상관관계를 파악하는 데에 유리하게 작용한다.

본 논문에서는 다양한 밀도를 가지는 대용량의 데이터집합에서 효과적으로 군집화할 수 있는 방법을 제안한다.

[1]본 연구는 건설교통부 첨단도시기술개발사업 - 지능형국토정보기술혁신사업과제의 연구비지원(과제번호:07 국토정보 C05)에 의해 수행되었습니다.

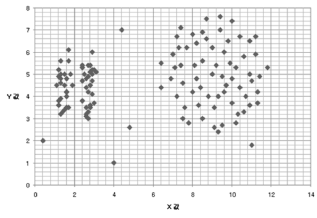
2. 알고리즘

이번 장에서는 우리가 제안하는 새로운 알고리즘을 소개한다.

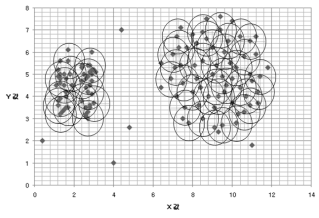


(그림 2) 알고리즘 순서도

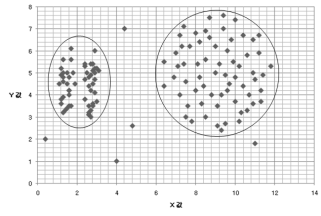
(그림 2)는 우리가 제안하는 군집화 기법에 대한 처리 단계를 보여준다. 우선, 1 단계에서 대용량의 데이터베이스의 모든 객체들을 핵심객체(core object), 경계객체(border object), 그리고 잡음객체(noise object)로 표시한다. 그 다음, 2 단계에서는 한 핵심객체로부터 밀도-도달가능인 객체들을 반복적으로 수집하여 작은 밀도-도달가능인 군집들을 병합해 나간다. 밀도-도달 가능인 객체들은 서로 밀도-연결되어 있다. 3 단계에서 최대로 밀도-연결 되어있는 객체들의 집합으로 독립적인 군집을 형성한다. 독립적으로 형성된 각각의 군집들 중 하나에 경계객체를 포함시킨다. 그리고 4 단계에서 어떤 군집에도 속하지 않는 잡음객체들은 제거한다. 5, 6 단계에서 독립적으로 생성된 군집들에 대하여 더 이상 새로운 군집이 발견되지 않을 때까지 재군집화(re-clustering)를 실시한다. 이때 최소 이웃 수는 그대로 유지하고, 반경 값만 변경시킨다.



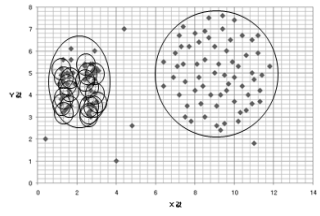
(a) 다른 밀도를 포함하는 데이터집합



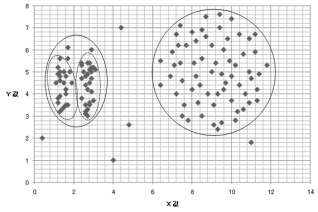
(b) 밀집영역 확장



(c) 2 개의 군집 생성



(d) 재군집화 실시



(e) 군집의 추가 생성

(그림 3) 새로운 알고리즘을 적용한 예

그림 3 은 우리가 제안하는 군집화 알고리즘의 처리 과정을 보여준다. 그림 3(a) 와 같이 예제를 활용하기 위하여 임의적으로 선택한 속성값들을 설정한 다음, 다른 밀도를 포함하는 데이터 집합을 구성하였다. 그림 3(b) 에서 보듯이 한 핵심객체로부터 밀도-도달가능인 객체들의 밀도 영역을 결합하여, 그림 3(c) 와 같이 2 개의 독립적인 군집을 생성하였다. 그런데 왼쪽 지역에 형성된 군집을 살펴보면 추가적으로 군집 생성이 가능하다는 것을 알 수 있다. 그러므로 그림 3(d) 에서와 같이 왼쪽 지역의 군집에 대해서 재군집화를 실시한다. 결국에는 그림 3(e) 에서 왼쪽지역에 형성하고 있던 하나의 군집에서 2 개의 군집이 추가 생성되었음을 확인할 수 있다. 만약 군집화 과정이 계속 반복된다면 군집의 개수는 더 증가할 수 있다.

3. 결론 및 향후 연구

본 논문에서는 다양한 밀도를 가지는 데이터집합에서 효율적으로 군집을 분류할 수 있는 새로운 군집화 기법을 제안하였다. 본 논문에서 제안하고 있는 반복적인 군집화 과정을 통하여 세밀한 군집이 형성된다. 그리고 대용량의 데이터로부터 새로운 패턴을 쉽게 찾아낼 수 있다. 또한, 계층구조를 사용하여 군집을 표현함으로써 각 군집의 성격이나 상관관계를 파악하기에 유리하다. 우리는 임의의 표본 데이터집합을 이용한 예제를 통하여 다양한 밀도를 가진 군집들을 효율적으로 분류할 수 있음을 확인하였다. 우리가 제안한 군집화 기법이 경제학, 사회학, 지리학, 의학, 유전학 등의 여러 응용 분야에서 다양하게 활용될 수 있을 거라고 기대된다.

참고문헌

[1] M. Ester, H.-P., Kriegel, J. Sander, X. Xu, A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, Proc. Int. Conf. on Knowledge Discovery and Data Mining, pp. 226-231, 1996.

[2] Zhang T., Ramakrishnan R., Linvy M, BIRCH: An Efficient Data Clustering Method for very Large Databases, Proc. ACM SIGMOD. Int. Conf. on Management of Data, ACM Press, pp. 103-114, 1996.