효율적인 계층적 클러스터링 방안

홍지원, 김상욱
한양대학교 전자컴퓨터통신공학과
e-mail : nowiz@dake.hanyang.ac.kr

A Method for Efficient Hierarchical Clustering

Ji-Won Hong, Sang-Wook Kim
Dept. of Electronics Computer Engineering, Hanyang University

요 약

본 논문에서는 계층적 클러스터링의 확장성(scalability)을 향상시키기 위한 방안으로 새로운 초기 클러스터링 기법을 제안한다. 본 논문에서는 k-NN 그래프를 구축하여 초기 클러스터의 중심이 될 객체를 찾고, 이 객체를 중심으로 유사한 다른 객체들을 초기 클러스터에 포함시키는 방법을 제안하고, 실험을 통해 제안하는 방법의 성능 개선 효과를 규명하였다.

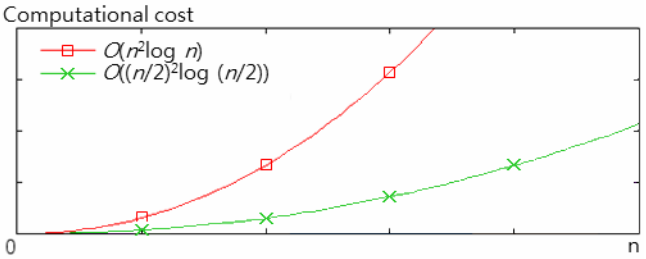
1. 서론

데이터를 유사한 객체들로 이루어진 그룹들로 나누는 것을 클러스터링이라 한다. 클러스터링을 할 때 만들어지는 각 그룹은 클러스터라 불리며, 같은 클러스터로 분류된 객체들은 서로 유사하고, 다른 클러스터로 분류된 객체들은 서로 유사하지 않아야 한다[1]. 기존의 클러스터링 기법들로는 분할(partitioning) 기법, 계층적(hierarchical) 기법, 밀도 기반(density-based) 기법, 격자 기반(grid-based) 기법, 모형 기반(model-based) 기법 등이 존재한다[1].

이 중에서 계층적 클러스터링 기법은 객체들을 클러스터들의 계층적 구조로 그룹화하는 방법이다. 계층적 클러스터링에는 병합식(agglomerative)과 분할식(divisive) 방법이 있다. 병합식 방법에서는 초기에는 각각의 객체가 하나의 클러스터를 나타내며, 이 클러스터들을 병합해 나가는 방법으로 클러스터링을 수행한다. 분할식 방법에서는 반대로 처음에는 모든 객체가 하나의 클러스터에 포함되며, 이 클러스터를 분할해 나가는 방법으로 클러스터링을 수행한다. 본 논문에서는 병합식 방법을 사용하여 계층적 클러스터링을 수행한다. 병합식 방법은 분할식에 비해 계산하기 용이하며, 결과 클러스터의 품질 또한 좋은 것으로 알려져 있다[2].

병합식 계층적 클러스터링은 반복적으로 한 번에 두 클러스터씩 병합하여 클러스터가 하나가 남게 되면 종료된다. 이 때 각 반복에서 모든 클러스터 쌍의 유사도를 계산해야 그 단계에서 병합할 두 클러스터들을 선택할 수 있기 때문에 총 계산 복잡도는 $O(n^2 \log n)$이 된다[2]. $O(n^2 \log n)$은 상당히 큰 복잡도이기 때문에 대상 데이터가 크다면 확장성(scalability) 측면에서 문제가 있다.

확장성 문제를 해결하기 위한 방법으로 초기 클러스터링(initial clustering)이라는 방법을 사용할 수 있다[3]. 초기 클러스터링 방법은 서로 충분히 가까이 있어서 클러스터링을 통해서 같은 클러스터에 포함될 것이 확실한 객체들을 빠른 속도로 초기 클러스터라 불리는 하나의 작은 그룹으로 만드는 방법이다. 초기 클러스터들을 각각 하나의 객체로 보고 클러스터링을 수행한다면, 전체 객체의 수가 아닌 초기 클러스터들의 수가 클러스터링의 대상이 되어 데이터의 규모가 줄어든 것과 같은 효과를 얻을 수 있다. 초기 클러스터링에 사용되는 알고리즘이 충분히 단순하고 전체 객체 수가 크다면 초기 클러스터들을 구성하는 과정에 들어가는 비용은 초기 클러스터링을 수행하여 절감되는 비용에 비해 무시할 수 있는 수준이 된다.



(그림 1) 데이터 크기에 따른 계산 비용

(그림 1)은 계산 복잡도 별로 데이터 크기에 따른 계산 비용의 변화를 그래프로 나타낸 것이다. 그래프에 따르면, 클러스터링 대상 데이터가 포함하는 객체의 수 n 이 $n/2$ 이 된다면 확장성이 크게 향상된다. 따라서 본 논문에서는 초기 클러스터들을 효율적으로 구성하여 계층적 클러스터링의 확장성을 향상시키는 새로운 방안을 제안한다.

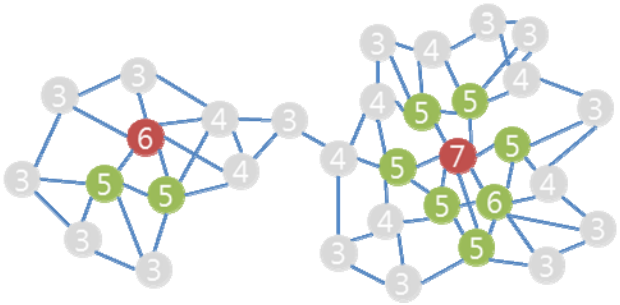
2. 초기 클러스터링

초기 클러스터링은 일반적으로 초기 클러스터 구성의 기준이 되는 중심점들을 찾고, 다른 객체들을 가장 유사한 중심점에 병합하는 방법으로 수행한다[3]. 따라서 초기 클러스터링의 중요한 문제는 객체들의 중심에 위치하는 중심점들을 효과적으로 찾는 것이다. 본 논문에서는 적절한 개수의 중심점들을 자동으로 찾아내기 위해 다음과 같이 정의되는 중심 객체를 찾아내고, 이를 중심점으로 사용한다. 이를 위해 데이터를 k-NN 방법(k-Nearest Neighbor)을 사용하여 그래프로 모델링한다.

정의 1. (중심 객체, centric object) 주어진 k-NN 그래프[1]에서 각 정점의 차수(degree)를 구한 뒤, 그 차수가 간선으로 직접 연결된 다른 모든 정점보다 높은 객체를 중심 객체라 한다.

k-NN 그래프 상에서 각 정점의 차수는 그 정점이 나타내는 객체 주변에 존재하는 다른 객체들의 밀도와 밀접한 연관이 있다. 그러나 단순히 주변 객체들의 밀도가 높은 객

체를 찾는다면 밀도가 높은 지역에서만 많은 수의 중심점들이 나타나고, 밀도가 낮은 지역에는 중심점이 나타나지 않게 된다. 따라서 중심점은 절대적인 밀도와는 관계없이 다른 객체들의 중심에 있는 객체를 선택해야 한다. 이러한 조건을 만족하기 위해 정의 1 의 중심 객체를 중심점으로 사용한다.



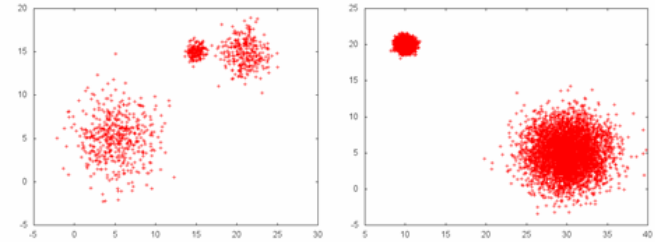
(그림 3) 합성 데이터의 3-NN 그래프

(그림 3)은 합성(synthetic) 데이터의 3-NN 그래프를 찾은 것이다. 그림의 원은 그래프의 정점을 의미하며, 선은 간선을 의미한다. 원 안의 숫자는 해당 정점의 차수를 나타낸다. 만약 절대적인 밀도를 본다면 차수가 6 인 두 정점과 7 인 정점이 모두 중심점으로 선택되거나, 7 인 정점 하나만 중심점으로 선택되어 그래프의 왼쪽 부분에는 중심점이 없는 결과가 나온다. 중심 객체의 정의를 사용하면, 가장 짙은 색으로 표시된 정점들이 중심 객체로 선택되어 적절한 중심점을 찾아내는 것을 볼 수 있다.

중심 객체들이 정해지고 나면 이제 중심 객체가 아닌 다른 모든 객체들이 어느 초기 클러스터에 속할지를 정해야 한다. 본 연구에서는 각 객체에 대해 중심 객체들 중 그 객체와 가장 유사한 중심 객체를 선택하여 같은 클러스터에 속하도록 한다. 가장 유사한 중심 객체를 선택하는 방법에는 코사인 유사도(cosine similarity), 유클리드 거리(Euclidean distance) 등의 여러 방법이 사용될 수 있다.

3. 실험 및 성능 평가

초기 클러스터링의 성능 평가 척도로는 초기 클러스터링을 적용하지 않은 계층적 클러스터링에 비해 초기 클러스터링을 적용했을 때 어느 정도의 속도 향상이 있는지, 초기 클러스터링을 적용함으로써 클러스터링의 질이 저하되지는 않는지 등을 살펴볼 필요가 있다.



(그림 3) 실험 데이터 (좌: Data1, 우: Data2)

본 논문에서는 쉽게 결과를 확인하기 위해 가우시안(Gaussian) 분산을 이용하여 생성된 두 가지의 2D 합성 데이터를 사용하여 클러스터링을 수행하였다. (그림 2)는 두 데이터의 분포를 2 차원 상에 표시한 것이다. Data 1, 2 는 객체 수가 10^3 개, 10^4 개이며, 각각 3 개와 2 개의 뚜렷한 클러스터들로 나뉘어진다. 실험에서는 유사도 측정 방안으로 유클리드 거리를 사용하였으며, 10-NN 그래프를 구성하였다.

<표 1>은 제안하는 방법을 사용하였을 때 전체 객체 수가 어떻게 줄어드는지를 나타낸다. <표 1>을 보면, 초기 클러스터링을 통해 대상 객체의 수를 전체 객체 수의 약 3% 정도로 줄였음을 알 수 있다. 이렇게 줄어든 객체를 이용하여 계층적 클러스터링을 수행했을 때의 시간은 <표 2>와 같다. <표 2>에 따르면, 데이터의 개수가 10^3 개일 때에는 20 배, 10^4 개일 때에는 약 300 배에 달하는 속도 향상을 보였다.

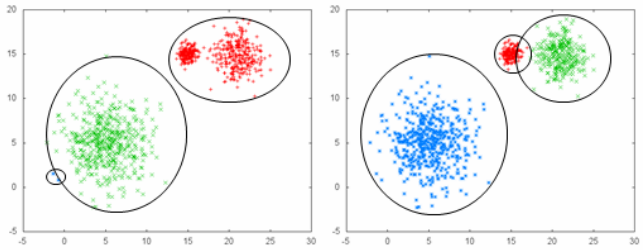
클러스터링 결과에 대해서는 초기 클러스터링 방법을 적용하여도 품질이 손상되지 않았다. Data 2 의 경우에는 제안하는 방법을 적용한 것과 적용하지 않은 것이 동일한 클러스터링 결과를 보였다. 그러나 Data 1 의 경우에는 초기 클러스터링을 적용한 경우에 더 높은 품질의 클러스터링 결과가 나왔다. (그림 4)는 Data 1 에 대해 초기 클러스터링을 적용하지 않은 경우와 적용한 경우의 결과이다.

<표 1> 제안하는 방법으로 줄어든 객체 수

	Data 1	*Data 2*
전체 객체 수	10^3	10^4
초기 클러스터 수	32	375
비율 (%)	3.2	3.75

<표 2> 소요 시간 비교

	Data 1	*Data 2*
기존 방법 (s)	40	39222
제안 방법 (s)	2	118
향상률 (배)	20	332



(그림 4) Data 1 의 클러스터링 결과
(좌: 초기 클러스터링 미적용, 우: 초기 클러스터링 적용)

4. 결론

본 논문에서는 새로운 초기 클러스터링 기법을 이용하여 계층적 클러스터링의 효율을 향상시킬 수 있는 방안을 제안하였다. 이 방법은 데이터를 k-NN 그래프로 모델링하고, 이를 통해 찾을 수 있는 중심 객체를 이용하여 적절한 초기 클러스터들을 생성한다. 또한 본 논문에서는 실험을 통해 제안하는 방안이 계층적 클러스터링의 속도를 데이터의 크기가 10^3 일 때에는 약 20 배, 10^4 일 때에는 약 300 배 향상시킴을 확인하였다.

감사의 글

"본 연구는 지식경제부 및 정보통신산업진흥원의 'IT 융합 고급인력과정 지원사업'의 연구결과로 수행되었음" (NIPA-2010-C6150-1001-0005)

참고문헌

[1] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, 2006.

[2] Y. Zhao, G. Karypis and U. Fayyad, "Hierarchical Clustering Algorithms for Document Datasets," In *DMKD*, Vol. 10, No.2, pp. 141-168, 2005.

[3] D. R. Cutting, D. R. Karger, J. O. Pedersen and J. W. Tukey, "Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections," In *SIGIR*, pp. 318-329 1992.