Priority Queue 를 이용한 Hierarchical Clustering (Centroid Linkage) 성능 개선

전용권, 윤성로*
고려대학교 대학원 전기전자전파공학과
*e-mail : sryoon@korea.ac.kr (corresponding author)

A Performance Improvement Study On Hierarchical Clustering (Centroid Linkage) Using A Priority Queue

Yongkweon Jeon, Sungroh Yoon
Dept. of Electrical Engineering, Korea University, Seoul 136-713, Republic of Korea

요 약

기존 hierarchical clustering 은 Time complexity 와 space complexity 가 Large data set 을 clustering 하기에는 적당하지 못하며 이것을 일반 PC 의 메모리 내에서 해결하는데 어려움이 있다. 따라서 본 연구에서는 이러한 어려움을 극복하기 위해 기존 Hierarchical clustering 중 Centroid Linkage 에 새로운 Algorithm 을 제안하여 보다 적은 메모리를 사용하고 빠르게 처리하는 방법을 제안하고자 한다.

1. 서론

최근 생명과학분야나 경제학분야 등에서 수많은 데이터가 쏟아져 나오고 있으나 그 많은 데이터를 사람이 일일이 분석하기에는 어려움이 따른다. 그리하여 컴퓨터를 이용한 다양한 방법들을 사용하여 이러한 대용량 데이터를 분석하기 위해 많은 연구가 이루어지고 있다.

그 동안 대용량 데이터를 분석하기 위해 많은 연구가 제안되었으며 그것들 중 Agglomerative Hierarchical clustering (AHC)이 오래되었지만 널리 이용되는 방법 중에 하나이다. AHC 는 정해진 계산법 (metric)에 따라 한 data set 의 모든 객체들간의 비유사도 (dissimilarity)를 계산한다. 그 후 정해진 linkage 방법에 따라 비유사도가 가장 낮은 것부터 병합을 시작하여 최종적으로 하나의 군집을 완성하게 된다 [1].

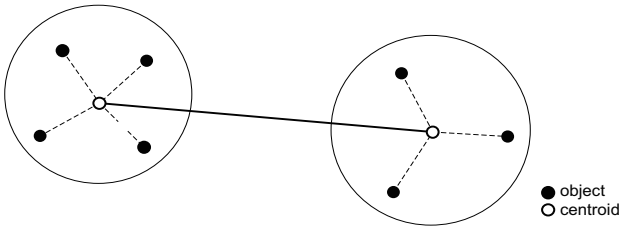
그러나 기존 이 방법은 모든 객체들간의 비유사도를 메모리에 저장해야 하는데 이 경우 공간 복잡도가 $O(N^2)$이기 때문에 대용량 데이터를 처리하기에는 한계가 있다 [2].

따라서 우리는 이러한 문제를 극복하기 위해 AHC 중 Centroid Linkage 방식의 공간 복잡도를 O(N)로 줄이는 새로운 Algorithm 을 제안하여 대용량 데이터를 처리하도록 하였다.

2. 방법

2.1 Centroid Linkage

기존 Hierarchical clustering 에서는 모든 객체간의 dissimilarity 를 계산하고 그것을 matrix 저장하였다. 하지만 Centroid linkage 에서는 어떤 한 cluster 가 병합이 되면 병합된 cluster 내의 object 들의 거리 평균이 이 cluster 의 대표 값이 되어 다른 cluster 와의 비유사도를 측정할 때 기준이 된다. 따라서 객체들은 비유사도가 가장 가까운 객체와 최초 merge 된 후에는 다른 객체와의 비유사도 정보가 필요 없게 된다. 즉, 각 객체들 마다 가장 가까운 객체들과의 정보만 필요할 뿐 그 이후 비유사도가 낮은 객체들과 정보는 불필요하게 된다.



(그림 1) Centroid Linkage

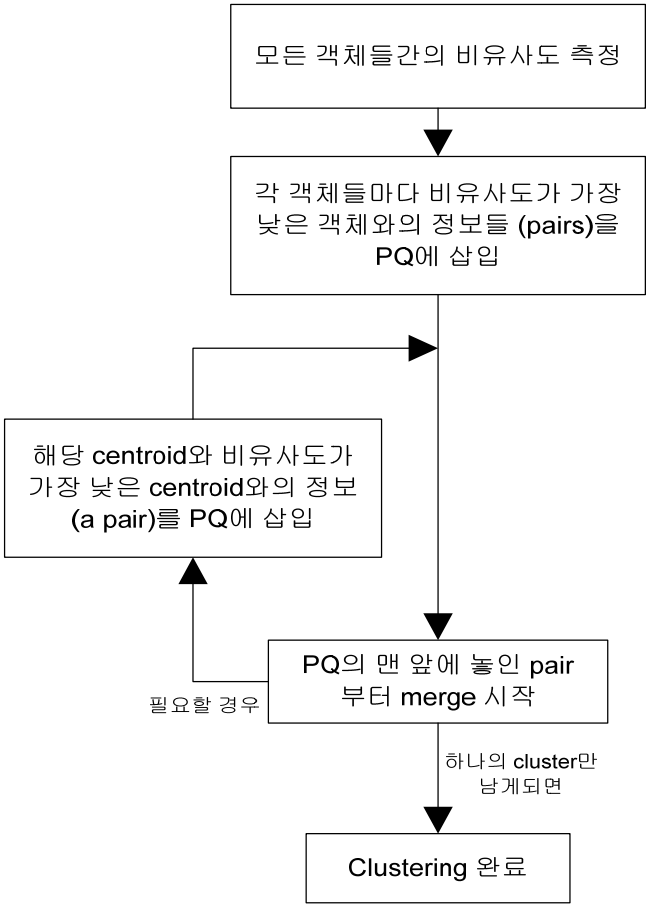
2.2 Algorithm

각 객체마다 비유사도가 가장 낮은 객체를 저장하여 Priority Queue 에 삽입을 하게 된다. 모든 객체에 대해 비유사도가 가장 낮은 객체와의 정보만을 Priority Queue 에 삽입하게 되면 공간 복잡도는 O(N)이 된다. 이 때 시간 복잡도는 모든 객체들간의 비유사도를 측정하기 때문에 $O(N^2)$가 되지만 모든 객체들간의 비유사도를 메모리에 저장할 때보다는 operation cost 를 절약할 수 있기 때문에 이 부분에서 기존보다 속도향상을 얻을 수 있다.

Priority Queue 에 들어있는 pair 들 중에 비유사도가 낮은 순으로 dequeue 가 되면서 cluster 들의 merge 가 진행 된다. 각 merge 발생 마다 생기는 cluster 의

centroid 에 대해 다른 cluster 들의 centroid 와의 비유사도를 측정하고 그 것들 중 비유사도가 가장 낮은 pair 하나를 다시 enqueue 를 하게 된다. 따라서 모든 객체들의 하나의 cluster 로 군집화 될 때까지 이러한 방식으로 clustering 이 진행 되어 공간 복잡도는 O(N)으로 계속 유지 할 수 있다.

Priority Queue 에서 dequeue 되는 pair 들이 모두 merge 가 되지는 않는다. dequeue 되어 나온 pair 의 한 centroid 가 이미 이전에 다른 centroid 와 merge 가 되어 다른 cluster 를 이루고 있다면 이 pair 의 정보는 잘못된 것으로 merge 를 진행해서는 안 된다. 이 경우 해당 pair 는 버려지게 되며 해당 pair 를 이루는 centroid 중 다른 cluster 와 merge 되지 않은 cluster 의 centroid 가 있다면 이것과 비유사도가 가장 낮은 centroid 를 다시 찾아 Priority Queue 에 삽입을 하게 된다.

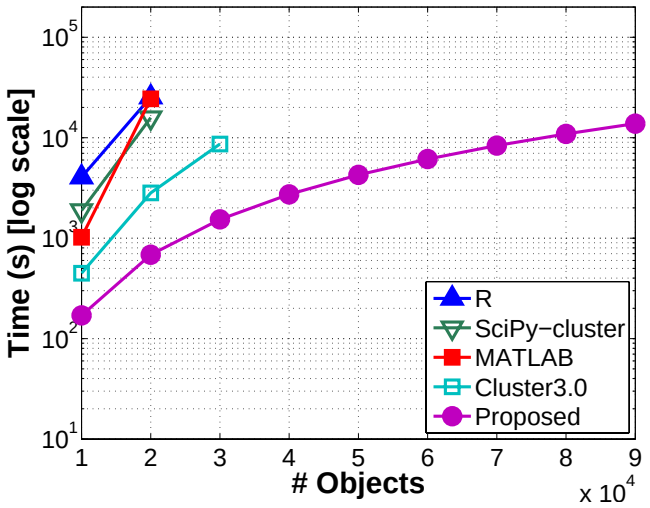


(그림 2) 알고리즘

3. 결과

새로운 알고리즘을 적용하여 기존의 방법을 사용하는 다양한 tools 과 속도 비교를 하였다. (그림 3)은 1000 차원을 갖는 데이터들을 총 객체 수를 변화시켜 MATLAB[i], SciPy [3], R [4], Cluster3.0 [5] 등 다양한 소프트웨어를 사용하여 Centroid Linkage 의 처리시간을 측정한 결과이다.

새로운 알고리즘에서는 Centroid Linkage 의 특성을 파악하여 공간 복잡도를 O(N)으로 줄여 기존 방법보다 더 큰 사이즈의 데이터를 처리할 수 있었다. 그 뿐만 아니라 기존 방법에서는 매 merge 발생마다 matrix 를 update 해줘야하기 때문에 계산에 필요한 시간 복잡도 O(N)과 memory operation 에 필요한 시간 복잡도 O(N) 이 필요한 반면에 새로운 알고리즘에서는 계산에 필요한 시간 복잡도는 O(N)으로 같지만 memory operation 에 필요한 시간 복잡도를 O(1) 으로 줄일 수 있기 때문에 시간 복잡도는 동일하지만 실제적인 처리 시간은 줄일 수 있었다.



(그림 3) < 1000 dimensions >

4. 결론

Centroid Linkage 의 Linkage 특성을 파악하여 기존보다 빠르고 대용량을 처리할 수 있는 방법을 제안하였다.

5. 사사

이 논문은 2010 년 정부(교육과학기술부)의 재원으로 한국 연구재단 (NRF)의 지원을 받아 수행된 기초연구사업임. (NRF 2010-0000631, NRF 2010-0000407)

참고문헌

[1] P. Tan, M. Steinbach, V. Kumar, *Introduction to Data Mining*, Addison Wesley, 2005.

[2] A. Jain, M. Murty, and P. Flynn, "Data clustering: a review," *ACM computing survey*, vol. 31, no. 3, 1999.

[3] D. Eads, 2008, hcluster: Hierarchical Clustering for SciPy. [Online]. Available: http://scipy-cluster.googlecode.com

[4] R Development Core Term, R: A language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2009.

[5] M. de Hoon, S. Imoto, J. Nolan, and S.Miyano, "Open source clustering software," *Bioinformatics*, vol. 20, no. 9, pp. 1453-1454, 2004.

[i] http://www.mathworks.com