제33회 한국정보처리학회 춘계학술발표대회 논문집 제17권 제1호 (2010. 4)

계층적 클러스터링을 위한 매개변수를 요구하지 않은 초기 데이터 분할 방안

송석순, 윤석호, 김상욱
한양대학교 전자컴퓨터통신공학과
e-mail: bogely@agape.hanyang.ac.kr

Effective Data Partitioning in Hierarchical Clustering: A Parameter-Insensitive Approach

Suk-Soon Song, Seok-Ho Yoon, Sang-Wook Kim
Department of Electronics and Computer Engineering, Hanyang University

요 약

본 논문에서는 계층적 클러스터링을 위한 매개변수에 민감하지 않은 효과적인 데이터 분할 방안을 제안한다. 먼저, 선행 실험을 통하여 기존 방안이 매개 변수에 민감하다는 것을 보인다. 본 논문에 제안하는 방안은 주어진 데이터를 최적의 초기 부분 클러스터의 크기를 결정할 수 있는 측정 함수를 제안하고 제안된 측정 함수를 이용해서 주어진 데이터를 최적의 초기 부분 클러스터들로 분할한다. 또한, 분할된 초기 부분 클러스터들을 병합해서 질이 좋은 최종 클러스터들을 생성한다. 실험을 통하여 제안하는 방안이 기존 방안보다 매개 변수에 민감하지 않는다는 것을 보인다.

1. 서론

계층적 클러스터링 방법은 클러스터링의 대상이 되는 데이터들을 트리 구조로 클러스터링 한다. 클러스터링의 결과로 생성되는 이러한 트리를 덴드로그램(dendrogram)이라 부르는데 덴드로그램은 사용자가 원하는 클러스터의 개수에 맞게 클러스터들을 제공할 수 있다. 따라서 사용자가 요구하는 클러스터의 개수가 바뀌어도 적절한 결과를 제공할 수 있는 장점이 있다[1].

계층적 클러스터링 방법은 덴드로그램을 형성하는 방법이 상향식 또는 하향식인가에 따라 병합(agglomerative) 방법과 분할(divisive) 방법으로 분류된다. 계층적 클러스터링 방법의 대표적인 방법으로는 BIRCH[2], ROCK[3], CURE[4], 그리고 Chameleon[5] 등이 존재한다. 최근 연구에서는 병합 방법과 분할 방법을 통합하는 방법이 주목받고 있다[6]. 병합 방법과 분할 방법을 통합한 대표적인 방법은 Chameleon이다. Chameleon은 두 클러스터 간에 상호연결성과 근접성을 동시에 고려하여 클러스터링 하기 때문에 자연스럽고 균질한 클러스터들을 발견할 수 있다[5].

Chameleon은 크게 세 가지 과정을 통해 주어진 데이터들을 클러스터링 한다. 먼저, 주어진 데이터의 객체를 정점으로 나타내고, 객체들 간의 유사도는 간선의 가중치로 나타낸다. 이로써 주어진 데이터를 그래프로 표현한다. 그런 후에 생성된 그래프를 METIS[7]라는 분할 방법을 이용해서 초기 부분 클러스터들로 분할시킨다. METIS는 두 클러스터 사이의 유사도의 합이 최소가 되도록 주어진 데이터를 비슷한 크기의 두 클러스터들로 분할한다. METIS는 전체 데이터를 사용자가 지정한 부분 클러스터 크기가 될 때까지 재귀적으로 분할한다. 주어진 데이터를 METIS를 이용해서 초기 부분 클러스터들로 분할하기 위해서는 일반적으로 사용자가 초기 부분 클러스터의 크기를 전체 데이터의 m%와 같은 형식으로 분할 전에 미리 제공해야 한다. 마지막으로 분할된 초기 부분 클러스터들 간의 유사도를 계산하여 가장 유사한 두 부분 클러스터들을 병합해 나가는 과정을 반복한다[7].

Chameleon은 초기 부분 클러스터들을 생성하기 위해서 METIS를 사용한다. 따라서 METIS를 이용해서 주어진 데이터를 분할하기 전에 초기 부분 클러스터의 크기를 사용자가 미리 제공해야한다. 그러나 사용자는 데이터 분할 전에 적절한 m값을 예측하기가 어렵다. 따라서 본 논문에서는 이러한 문제를 해결하는 방안에 대해서 논의한다.

2. 제안하는 방안

2.1. 선행 연구

본 절에서는 선행 실험을 통하여 기존 Chameleon이 초기 부분 클러스터의 크기가 변함에 따라 최종 클러스터링의 결과가 어떻게 달라지는지 알아보고자 한다. 선행 실험에서 사용할 데이터는 Chameleon 방법을 제안한 [5]에서 사용한 4개의 데이터 중 1개를 사용한다. 그림 1은 실험에 사용한 2차원 데이터를 나타낸다. 최종 클러스터의 수는 6으로 설정한다.

그림 2는 초기 부분 클러스터의 크기를 2~4%까지 변화시키면서 Chameleon으로 그림 1의 데이터를 클러스터링 한 결과이다. 그림 2에서 같은 색상으로 표시된 점들은 동일한 클러스터에 포함되어 있는 것을 나타낸다. 그림 2를 보면 초기 부분 클러스터의 크기가 변함에 따라서 각각의 클러스터링 결과가 크게 달라지는 것을 알 수 있다. 이는 METIS로 데이터를 분할하기 전에 사용자가 매개변수 m값을 적절하게 설정하지 않으면 Chameleon으로 도출한 최종 클러스터의 질이 낮아질 수 있다는 것을 의미한다.

2.2. 제안하는 방안

본 논문에서 제안하는 초기 부분 클러스터 분할 방법은 주어진 데이터에 포함된 객체들의 수가 N개 일 때 1~N으로 클러스터의 개수로 정하고 기존 클러스터링 방법을 통해서 주어진 데이터를 클러스터링 한다. 본 논문에서는 기존 클러스터링 방법으로서 spectral clustering[8]을 사용한다. 그런 후에 클러스터링의 결과로 도출된 초기 부분 클러스터들의 질을 제안하는 측정 함수를 이용하여 평가하고 평가한 점수가 가장 높을 때의 초기 부분 클러스터들을 최적의 초기 부분 클러스터들로 정한다. 도출된 최적의 초기 부분 클러스터들을 병합하여 최종 클러스터들을 도출한다.

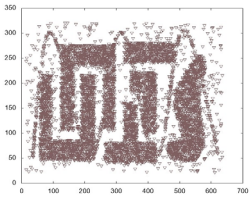

그림 1. 선행 실험을 위한 2차원 데이터.

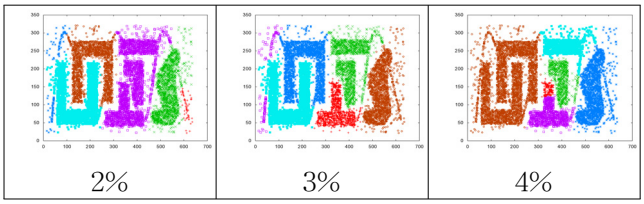

그림 2. METIS를 이용한 초기 부분 클러스터 크기 변화에 따른 클러스터링 결과.

본 논문에서는 클러스터들의 질을 평가하기 위해서 [9]에서 두 개의 클러스터들의 질을 평가하는 측정 함수를 N개의 클러스터들의 질을 평가할 수 있는 측정 함수 (1)로 확장한다. 수식 (1)에서 A_k는 k번째 클러스터를 의미하며, V는 모든 클러스터들의 합집합을 의미한다. $Nassoc(A_1,...,A_n)$은 클러스터들의 질을 평가하는 함수이며 값이 높을수록 클러스터들의 질이 좋은 것을 의미한다. $assoc(A_k,V)$는 A_k k번째 클러스터에 속한 객체들과 V에 속하는 객체들의 유사도의 합을 의미한다. 수식 (1)이 의미하는 바는 N의 클러스터가 있을 때, 모든 객체들이 가지고 있는 유사도의 전체 합 분에 동일 클러스터에 있는 객체들 간의 유사도이다. 다시 말해, 같은 클러스터에 속한 객체들끼리는 유사하고 다른 클러스터에 속한 객체들끼리는 유사하지 않으면 평가 점수가 높다. 이는 유사한 객체들은 같은 클러스터에 포함되고 유사하지 않은 객체들은 다른 클러스터에 포함되는 클러스터링의 정의와 부합한다[1].

$$Nassoc(A_1,...,A_n) = \sum_{k=1}^{n} \frac{assoc(A_k,A_k)}{assoc(A_k,V)} \quad (1)$$

제안하는 방안은 모든 가능한 클러스터들의 개수가 크면 클수록 수행 속도가 급격하게 느려진다는 문제점이 있다. 따라서 수행 속도가 빠른 METIS로 주어진 데이터를 일정한 크기로 분할해서 임시 부분 클러스터들을 생성한 후에 제안하는 방안을 적용하여 최적의 초기 부분 클러스터들을 생성한다. 제안된 방안으로 최적의 초기 부분 클러스터들을 생성한 후에 기존 Chameleon으로 초기 부분 클러스터들을 병합하여 최종 클러스터들을 생성한다.

4. 실험

본 실험은 선행 실험과 동일한 방식으로 매개변수에 변화에 따라 제안하는 방안으로 클러스터링한 결과가 어떻게 변화하는지 확인한다.

그림 3은 초기 부분 클러스터의 크기를 2~4%까지 변화시키면서 제안하는 방안으로 그림 1의 데이터를 클러스터링 한 결과이다. 그림 3에서 같은 색상으로 표시된 점들은 동일한 클러스터에 포함되어 있는 것을 나타낸다. 초기 부분 클러스터의 크기를 2~4%로 변화시켰기 때문에 그림 3의 각 그림들에서 나타나는 클러스터들의 색상은 그림마다 다를 수 있다. 그림 3을 보면 초기 부분 클러스터의 크기를 변화시켰음에도 불구하고 최종 클러스터링의 결과는 거의 변화가 없으며 6개의 클러스터들을 정확하게 찾은 것을 알 수 있다. 따라서 본 논문에서 제안하는 방안이 기존 방안보다 매개변수에 민감하지 않으면서 클러스터링의 정확도가 높다는 것을 알 수 있다.

5. 결론

본 논문에서는 계층적 클러스터링을 위한 매개변수에 민감하지 않은 효과적인 데이터 분할 방안을 제안하였다. 먼저, 선행 실험을 통하여 기존 방안이 매개 변수에 민감하다는 것을 보였다. 제안하는 방안은 주어진 데이터를 최적의 초기 부분 클러스터의 크기를 결정할 수 있는 측정 함수를 제안하고, 제안된 측정 함수를 이용해서 주어진 데이터를 최적의 초기 부분 클러스터들로 분할한다. 또한, 분할된 초기 부분 클러스터들을 병합해서 질이 좋은 최종 클러스터들을 생성한다. 실험을 통하여 제안하는 방안이 기존 방안보다 매개 변수에 민감하지 않는다는 것을 보였다.

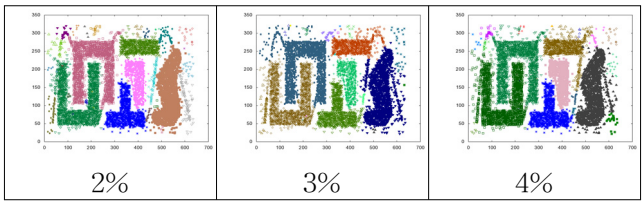

그림 3. 제안하는 방안을 이용한 초기 부분 클러스터 크기 변화에 따른 클러스터링 결과.

감사의 글

이 논문은 2009년도 정부(교육과학기술부)의 재원으로 한국연구재단(No. 2008-0061006)의 지원을 받아 수행되었으며, 또한 지식경제부 및 정보통신산업진흥원의 대학 IT연구센터 지원사업(NIPA-2010-(C1090-1011-0009))의 부분적인 지원을 받았습니다.

참고문헌

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

[2] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An Efficient Data Clustering Method for Very Large Databases," *ACM SIGMOD Record archive*, Vol. 25, No. 2, pp. 103-114, 1996.

[3] S. Guha, R. Rastogi, and K. Shim, "ROCK: A Robust Clustering Algorithm for Categorical Attributes," In *Proc. Int'l Conf. Data Engineering*, pp. 512-521, 1999.

[4] S. Guha, R. Rastogi, and K. Shim, "CURE: An Efficient Clustering Algorithm for Large Databases," In *Proc. ACM SIGMOD Int'l Conf. Management of Data*, pp. 73-84, 1998.

[5] G. Karypis, E. Han, and V. Kumar, "Chameleon: Hierarchical Clustering Using Dynamic Modeling," *IEEE Computer*, Vol. 32, pp. 68-74, 1999.

[6] T. Li and S. Anand, "DIVA: A Variance-Based Clustering Approach for Multi-type Relational Data," In *Proc. Int'l. Conf. on Information and Knowledge Management*, pp. 147-156, 2007.

[7] G. Karypis, and V. Kumar, "A Fast and High Quality Multilevel Scheme for Partitioning Irregular Graphs," *Journal of Society of Industrial and Applied Mathematics*, Vol. 20, No. 1, pp. 359-392, 2008.

[8] A. Ng, M. Jordan, and Y. Weiss, "On Spectral Clustering: Analysis and an Algorithm," *In Advances in Neural Information Processing Systems 14*, 2001.

[9] J. Shi and J. Malik, "Normalized Cuts and Image Segmentation," In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1997.