계층 발생 프레임워크를 이용한 군집 계층 시각화

신동화, 이세희, 서진욱
서울대학교 컴퓨터공학부
dhshin@hcil.snu.ac.kr, shlyi@hcil.snu.ac.kr, jseo@snu.ac.kr

Visualizing cluster hierarchy using hierarchy generation framework

DongHwa Shin, Sehi L'Yi, Jinwook Seo
Department of Computer Science and Engineering, Seoul National University

요 약

군집화 알고리즘은 그 종류에 따라 잡아낼 수 있는 군집의 종류와 보여줄 수 있는 정보의 수준이 차이가 난다. 밀도기반 군집화 알고리즘은 데이터 분포 상의 임의의 모양을 가진 군집을 잘 잡아내지만 보여줄 수 있는 계층정보가 매우 적거나 없는 수준이고, 반면 계층적 군집화 알고리즘은 자세한 계층 정보를 보여주지만 구 모양의 군집 외에는 잘 잡아내지 못한다. 이 논문에서는 이러한 두 군집화 알고리즘의 장점만을 취하는 계층 발생 프레임워크를 제시하고 이와 더불어 효과적 데이터 분석을 위한 여러 시각화, 상호작용 기법을 지원하는 시각분석 애플리케이션을 제공하였다.

1. 서 론

가장 보편적으로 쓰이는 데이터 분석 기법 중 하나인 군집화(clustering)는 비슷한 데이터를 하나의 군집으로 묶어서 같은 군집 내의 데이터 간 유사도(similarity)를 최대로 하고, 타 군집과의 유사도는 최소화하는 것을 목적으로 한다. 대표적으로 분할기반(partitioning-based) 방식, 계층적(hierarchical) 방식, 밀도기반(density-based) 방식이 있다. 각 방식 별로 군집에 대한 정의의 차이가 있기 때문에 그 결과로서 도출되는 군집들 또한 고유한 특징을 갖고 있다. 계층적 군집화를 사용하는 데이터 분석가들은 데이터들간의 계층이나 관계는 쉽게 확인 가능하지만, 데이터의 분포가 구 모양이 아닌 임의의 모양을 가진 군집은 쉽게 찾아내지 못한다[1]. 반면 밀도기반 군집화의 경우에는 임의의 모양의 군집들은 잘 찾아내지만, 데이터와 각 군집들이 이루고 있는 관계를 한 눈에 알아보기 힘들다[2]. 이 문제로 인하여 의미 있는 군집들을 찾기 위해 동일한 데이터에 대해서 여러 군집화 알고리즘을 적용해 그 결과를 비교해보아야 하는 어려움이 있다.

기존 연구에는 밀도기반 군집을 시각적으로 표현하는 대표적 방식인 도달가능도(reachability plot)를 각 군집간 계층을 보여주는 덴드로그램(dendrogram)으로 전환시키는 알고리즘이 제시되었다[2]. 그러나 이 경우 정보의 손실 없이 전환을 하기 위한 제약조건이 매우 엄격하여 실용성이 크지 않다는 단점이 있다. 그 외에도 밀도기반으로 군집화를 수행하면서 시각화 표현은 덴드로그램으로 바로 나타내어 주는 HDBSCAN 알고리즘이 제시되었다[3]. 하지만 이 경우 사용자가 분석의 결과를 시각적으로 확인하고 실시간으로 매개변수를 정제하여 다시 반영할 수 있게끔 돕는 시각분석(visual analytics) 체계가 존재하지 않기에 개선의 여지가 있다고 할 수 있다.

이 연구에서는 앞서 소개한 밀도기반, 계층적 군집화 알고리즘의 장점을 모두 취하기 위해서 데이터 분석 시 임의의 모양을 갖는 군집을 찾아내었을 때 그 군집들 간의 관계와 군집 내의 자세한 계층까지 확인 가능하게 하는 계층 발생 프레임워크(hierarchy generation framework)를 제시한다. 또한 사용자의 효과적인 시각분석을 가능케 하는 여러 새로운 시각 표현 및 상호작용 기법들을 제공하는 군집화 시각분석 애플리케이션(application)을 제공한다.

2. 군집화의 계층 발생 프레임워크

임의의 모양의 군집을 찾기 위해서 먼저 밀도기반 군집화 알고리즘 OPTICS[1]를 수행하여 군집을 구한다. 그 결과로 도출된 군집들 중에서 사용자가 더욱 정밀한 계층을 확인하고 싶은 군집들을 선택하여 그 것 들을 대상으로 응집 계층 군집화(hierarchical agglomerative clustering) 알고리즘[4]을 수행한다(그림 1). 그러한 군집들을 '씨앗 군집' 이라 칭한다. 선택된 씨앗 군집들은 그 것들 중에서 가장 하위의 군집부터 차례로 응집 계층 군집화 알고리즘이 수행된다. 수행 결과로 얻은 하위 군집의 계층 구조상의 뿌리 노드(root node)를 그 상위 군집의 잎 노드(leaf node)로 취급하여 다시 응집 계층 군집화를 수행하면 최종적으로 씨앗 군집들 전체에 대한 하나의 계층, 즉 덴드로그램을 얻게 된다(그림 2). 이러한 방식으로 얻은 덴드로그램은 밀도기반 군집 수행 결과 구조를 그대로 유지하면서 각 점들간의 관계 및 더욱 자세한 계층을 직관적으로 파악할 수 있다는 장점을 가지고 있다. 즉, 밀도기반 군집들의 세세한 계층을 얻음으로써 두

* This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MEST) (No. NRF-2014R1A2A2A03006998)

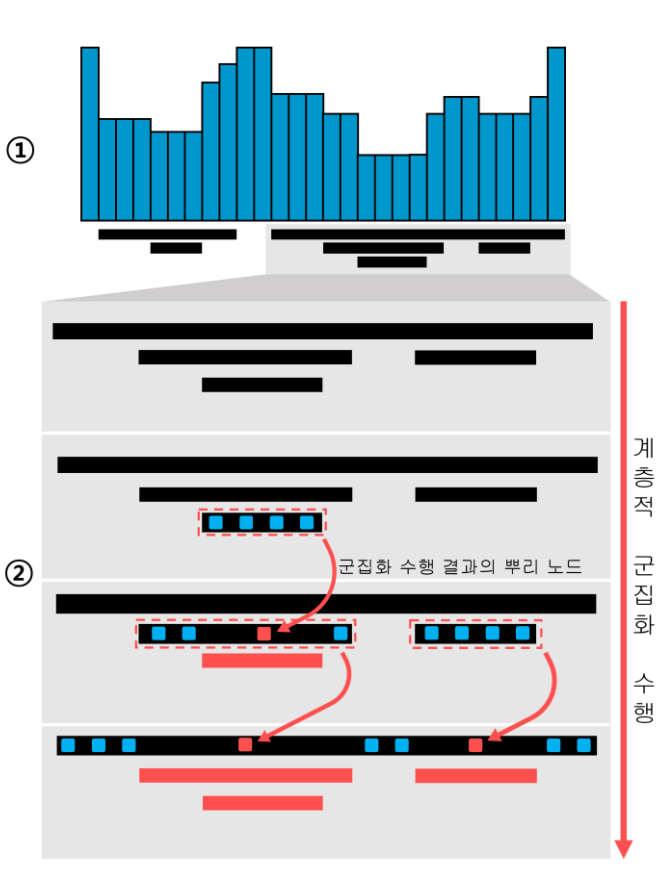

그림 1. 계층 발생 프레임워크
(1) 밀도기반 군집 분석의 결과인 도달가능도 (2) 사용자가 더욱 정밀한 계층 확인을 원하는 씨앗 군집을 선택하여 하위의 군집부터 응집 계층 군집화를 수행

군집화 방식의 장점을 모두 취한 것이다.

3. 시각화 및 상호작용 기법

사용자들로 하여금 앞서 제안한 계층 발생 프레임워크를 직관적으로 사용할 수 있게 하고, 더불어 효과적인 데이터 분석을 위한 여러 시각화 기법을 지원하는 시각분석 프로그램의 사용자 인터페이스(user interface)를 설계 및 구현하였다(그림 3). 인터페이스의 가운데 부분은 이 분석 프로그램의 핵심인 계층 발생 프레임워크 수행 결과를 확인하는 영역이다. 이 영역은 밀도기반과 계층적 군집화의 결과 표시 영역으로 나뉘는데 이를 상단 및 하단으로 병치시켜 표현하였다. 그 이유는 사용자로 하여금 밀도기반 군집화 수행, 씨앗 군집 선택, 그리고 그것들을 대상으로 한 응집 계층 군집화 수행까지의 일련의 과정을 인터페이스의 상단에서 하단으로 자연스럽게 내려가며 직관적으로 수행할 수 있도록 유도하기 위함이다. 이 외의 다른 시각화 기법들은 다음과 같다.

- 도달가능도에서 데이터의 한 점에 해당하는 도달가능막대에 음영처리를 하였다(그림 3-1). 이는 데이터의 밀도와 연결되어 밀도가 높을수록 더욱 음영이 진해지는 효과를 얻는다. 이를 통해 사용자들은 어느 군집이 더욱 높은 밀도의 군집인지를 한눈에 파악 가능하다.
- OPTICS 알고리즘의 매개변수 설정을 하는 시각화 요소에 동적 쿼리(dynamic query) 기법[5]을 적용하였다(그림 3-7). 이는 사용자가 매개변수를 바꾸고 그에 따른 군집화 결과의 변화를 목격하는 과정을 실시간으로 보여준다. 여러 매개변수를 설정하고 변화를 관찰해보아야 하는 군집화 작업의 특성을 고려한 상호작용 기법이다.
- 덴드로그램에서 내부적으로 군집을 품고 있는 노드인 군집 노드와 그렇지 않은 비 군집 노드를 한 눈에 구별하기 위해 군집 노드를 삼각형 모양으로 시각화하였다(그림 3-3, 3-4).
- 덴드로그램에서 사용자의 관심이 집중되는 씨앗 노드의 경우는 변(edge)을 파란 색으로, 비 씨앗 노드의 경우 빨간 색으로 시각화함으로써 분석의 관심 지역에 대한 구별을 더욱 쉽게 하였다(그림 3-3, 3-4).
- 덴드로그램의 확대/축소 기능을 위한 시각화 요소로서 확대 지시자를 추가하였다(그림 3-4). 덴드로그램 우측에 있는 이 지시자를 상하로 조정하면서 덴드로그램의 특정 부분은 더욱 확대하고 특정 부분은 축소를 할 수 있다. 이는 공간상 이유로 인해 유사도가 비슷한 노드들 간의 구분이 어려운 점을 해소하기 위해 고안하였다.

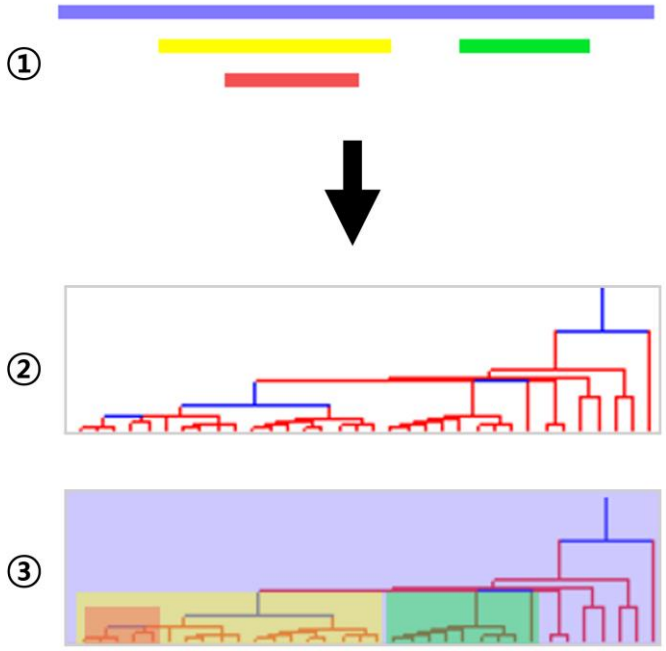

그림 2. 밀도기반 군집을 통해 얻은 덴드로그램
(1) 밀도기반 군집 (2) 덴드로그램 (3) 씨앗 군집의 구조가 덴드로그램 내에 보존되어 있는 모습

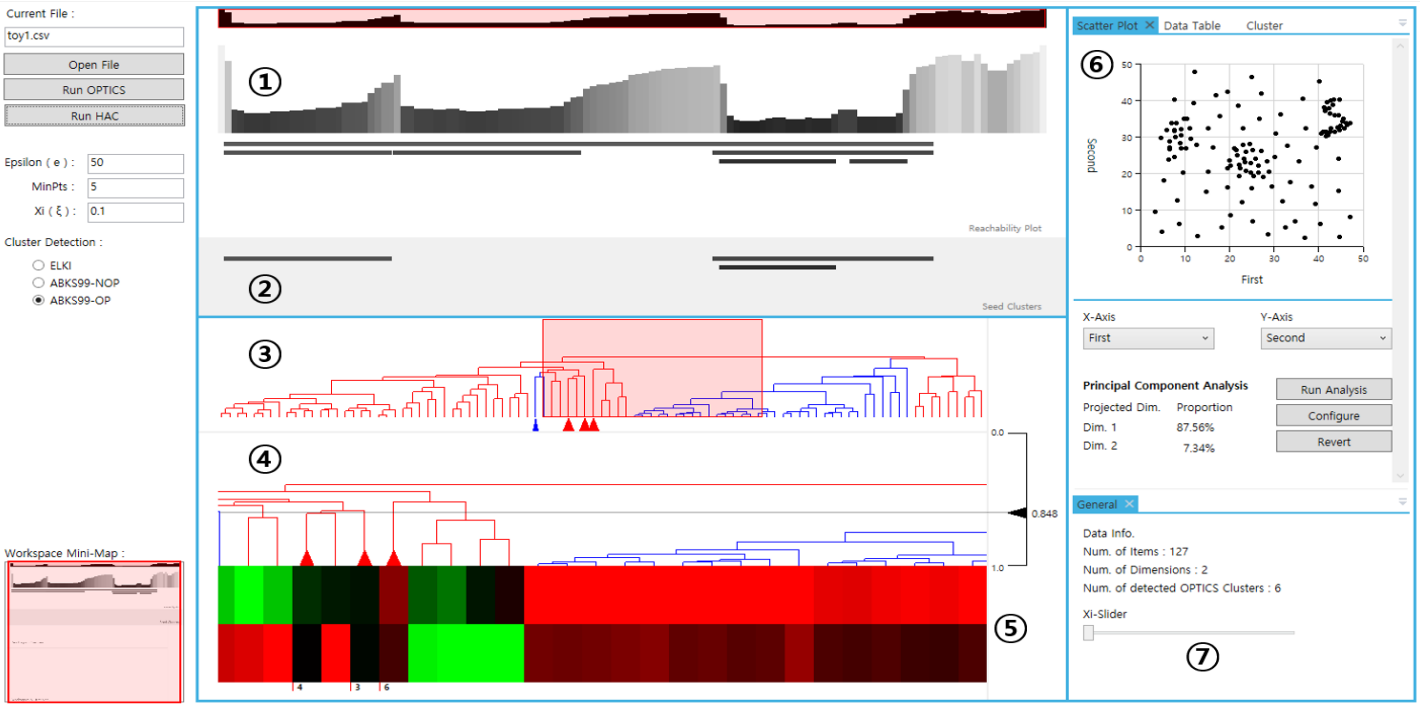

그림 3. 계층 발생 프레임워크을 지원하는 군집화 시각분석 프로그램의 사용자 인터페이스

(1) 도달가능도 및 밀도기반 군집화 결과 표시 영역 (2) 끌어다 놓은 씨앗 군집을 표시하는 영역 (3) 덴드로그램 개요(Overview) 영역 (4) 덴드로그램 확대 표시 영역 (5) 열지도 영역 (6) 산점도(scatter plot) 영역 (7) OPTICS 알고리즘의 매개변수 동적 쿼리를 지원하는 슬라이더(slider)

- 데이터를 표시하는 각 영역들 간에 공통되는 데이터들을 더욱 잘 파악하기 위해 시각화 기법인 브러시 앤 링크(brush & link)[6]를 지원한다.

4. 데이터 분석 및 논의

데이터 공간 상에서 구 모양이 아닌 여러 임의의 모양을 띤 군집을 포함하는 데이터를 이 연구에서 제시한 시각 분석 프로그램을 통해 분석했다. 그 결과, 동적 쿼리 기법을 이용하여 매우 다양한 밀도기반 군집을 짧은 시간 안에 다수 찾아 낼 수 있었다. 또한 씨앗 군집은 동적 쿼리 수행에 의해 계속 군집화 결과가 달라져도 씨앗 군집 영역에 그대로 남아있기 때문에 다양한 매개변수 하에서 얻을 수 있는 밀도기반 군집에 대한 덴드로그램을 한 번에 얻을 수 있었다. 그렇게 해서 얻은 덴드로그램은 밀도기반 군집 간의 관계, 각각의 군집 내부에 또 다른 군집들, 그리고 점들간의 관계 등 기존 밀도기반 군집 분석 혹은 계층적 군집 분석 하나만 따로 수행해서는 얻을 수 없는 정보들을 사용자에게 보여주는 것을 확인 할 수 있었다.

5. 결론 및 향후 연구

이번 연구를 통해 서로 다른 특징을 지닌 밀도기반, 계층적 군집화의 장점들을 모두 취하면서 동시에 효과적인 시각화 기법을 제공하는 계층 발생 프레임워크 및 시각 분석 프로그램을 제시하였다.

향후 연구에서는 이 논문에서는 다루지 않은 분할기반 군집화 및 기타 다른 방식의 군집화 기법들로의 분석 확장을 진행할 것이다. 또한 더욱 정량적인 데이터 분석과 그 결과에 대한 논의를 통해 완성도 높은 시각분석 프레임워크로 나아갈 것이다.

참 고 문 헌

[1] Ankerst, M., Breunig, M. M., Kriegel, H., Sander, J. OPTICS: Ordering Points To Identify the Clustering Structure. In *Proc. SIGMOD*, 49-60, 1999.

[2] Sander, J., Qin, X., Lu, Z., Niu, N., Kovarsky, A. Automatic Extraction of Clusters from Hierarchical Clustering Representations. In *Proc. PAKDD*, 75-87, 2003.

[3] Campello, R. J. G. B., Moulavi, D., Sander, J. Density-Based Clustering Based on Hierarchical Density Estimates. In *Proc. PAKDD*, 160-172, 2013.

[4] Balcan, M. F., Gupta, P. Robust Hierarchical Clustering. In *Proc. COLT*, 282-294, 2010.

[5] Shneiderman, B. Dynamic Queries for Visual Information Seeking. *Software, IEEE*, 11(6), 70-77, 1994.

[6] Buja, A., McDonald, J. A., Michalak, J., Stuetzle, W. Interactive Data Visualization using Focusing and Linking. In *Proc. Visualization*, 156-163, 1991.