그래프 클러스터링 방법들 비교 분석

이계준

성균관대학교 시스템경영공학과

성균관대학교 컴퓨터공학과

지도교수 황지영 교수님

요약

그래프(Graph)란 노드(Node)의 집합과 간선(Edge)의 집합으로 이루어진 구조를 말한다. 여기서 노드와 노드 사이에 연결 관계를 간선을 통해 표현하며 만약 간선으로 이루어져 있다면 둘은 이웃 관계에 있다고 명시한다. 정형화되지 않은 소셜 네트워크 상의 데이터(Stanford, Snap, Facebook data)를 각각의 특성을 통해 그래프화 하고 이를 기반으로 Neo-K-means, Spectral clustering, Girvan-Newman method 총 3가지의 군집화 방법을 통해 각각의 성능을 비교 분석하고 군집화 방법들의 특성에 따라 효과적으로 사용될 수 있는 데이터 종류를 찾아내고자 한다.

1. 서론

* 1. 제안배경 및 필요성

인터넷의 발전으로 인해 Facebook, Instagram과 같은 소셜 네트워크의 급격한 발전이 이루어졌다. 그 결과 소셜 네트워크 상에 무수히 다양한 종류의 정보가 존재한다. 하지만 이러한 정보는 정형화되지 않고 방대하기 때문에 정리하기 어려운 상황에 놓여있다. 이러한 데이터들을 각각의 특성들에 따라 정확히 군집화(Clustering)하고 선별화(Classification) 할 수 있다면 시각화 할 수 있고 후에 기계학습(Machine learning)을 통해 예측, 추천 등을 진행할 수 있다. 현재 소셜 네트워크 상의 데이터를 그래프화 하고 이를 여러 군집화 방법들에 적용하여 군집화 방법들의 특성들을 파악하고자 이 작품을 제안한다.

* 1. 졸업 작품의 목표

정형화되지 않은 소셜 네트워크 상의 데이터(Stanford, Snap, Facebook data)를 각각의 특성을 통해 그래프화 하고 이를 기반으로 Neo-K-means, Spectral clustering, Girvan-Newman method 총 3가지의 군집화 방법을 통해 각각의 성능을 비교 분석하고 군집화 방법들의 특성에 따라 효과적으로 사용될 수 있는 데이터 종류를 찾아내고자 한다.

* 1. 작품의 Overview

본 작품에서는 Neo-K-means, Spectral clustering, Girvan-Newman method 알고리즘 기법들을 실제 데이터에 적용한 후 그 결과에 따른 한계점과 장점 그리고 대해서 기술한다. 순서는 다음과 같다. 섹션 2에서는 그래프의 특성을 소개한 후 K-means, Neo-K-means, Spectral clustering, Girvan-Newman method 알고리즘에 대한 논문에 대한 간략한 소개와 알고리즘을 설명한다. 섹션 3에서는 실제 사용되어질 Facebook 데이터를 그래프화 하는 방법과 실제 코딩에 사용될 알고리즘의 수식을 소개한다. 섹션 4에서는 사용되어진 데이터의 분석과 군집화의 결과에 대한 분석한다. 섹션 5에서는 결론을 맺고 향후 발전 방향에 대해 설명한다.

2. 관련 연구

2-1) 그래프의 구조

그래프(Graph)란 노드(Node)의 집합과 간선(Edge)의 집합으로 이루어진 구조를 말한다. 여기서 노드와 노드 사이에 연결 관계를 간선을 통해 표현하며 만약 간선으로 이루어져 있다면 둘은 이웃 관계에 있다고 명시한다.

2-2) K-means clustering

[1]K-means clustering 알고리즘은 클러스터링 방법 중 Classification에 속한다. 이는 주어진 데이터를 여러 그룹으로 나누는 방법이다. 다시 말해, 데이터를 한 개 이상의 데이터 오브젝트로 구성된 k개의 그룹으로 나누는 것이다. 이 때 그룹을 나누는 과정은 거리 기반의 그룹간 비유사도 (dissimilarity) 와 같은 비용 함수 (cost function)을 최소화하는 방식으로 이루어지며, 이 과정에서 같은 그룹 내 데이터 오브젝트 사이의 유사도는 증가하고, 다른 그룹에 있는 데이터 오브젝트와의 유사도는 감소하게 된다. 각 그룹의 중심 (centroid)과 그룹 내의 데이터 오브젝트와의 거리의 제곱합을 비용 함수로 정하고, 이 함수 값을 최소화하는 방향으로 각 데이터 오브젝트의 소속 그룹을 업데이트 해 줌으로써 클러스터링을 수행하게 된다.

2-3) NEO-K-Means

[2]NEO-K-Means clustering은 그래프 군집화 알고리즘 중 중첩된 군집을 탐지하고 이상 탐지를 동시에 진행 가능한 알고리즘이다. 이 알고리즘에서는 군집의 중첩 정도를 나타내는 매개변수 알파(와 이상점의 비율을 나타내는 베타(를 사용한다. 이 알고리즘은 그래프의 노드 특성만을 이용하여 군집화 작업을 할 수 있고, 그래프의 구조만으로도 그래프 군집화 작업을 할 수 있다. 이를 위해 데이터를 그래프화 한 후 노드 간의 거리를 통한 K-Means 군집화 방법을 이용하여 군집화를 지속적으로 시행한다.

2-4) The Girvan-Newman Method

[3] Girvan-Newman 알고리즘은 betweenness centrality를 이용한다. betweenness centrality는 네트워크 내의 한 지점 i에서 j를 연결하는 최단 거리의 선상에 특정 지점 k가 얼마나 많은 빈도로 등장하는지를 나타낸다. 최단거리 알고리즘을 통해 노드 내의 모든 지점 i와 j간의 최단거리 루트를 구하고, 그 안에 등장하는 노드들 k의 빈도를 조사하여, 수치화 한다. 즉, betweenness centrality를 통해 우리는 Community간의 bridge역할을 하는 노드들과 연결들을 조사할 수 있다. betweenness centrality가 높을 수록 커뮤니티간 네트워크에서 중요한 위치를 차지할 것이다.

2-5) Spectral clustering

[4]Spectral clustering은 Unsupervised learning 군집분석 방법 중 한 가지로, 라플라시안 (Laplacian) 행렬을 이용한 그래프 분할(graph partitioning) 기반의 군집분석 기법 이다. 이 방법은 k-means와 같은 전통 적인 군집분석 기법과는 대조적으로, 볼록 한 모양 뿐 아니라 다양한 군집의 형태에 대해서도 군집화가 가능하다.

3. 제안 작품 소개

3-1) 실험 환경

실험을 위해 CPU intel core u3-6100U 2.30GHz, 메모리 8GB를 사용하였다. 실험을 위한 프로그램으로 Window10 환경에서 Python 3.6을 사용하였다.

3-2) 소셜 그래프

앞서 관련 연구에서 언급한 Graph 구조에서 사람을 노드로, 사람사이의 관계를 간선으로 나타낸 그래프 모델을 소셜 그래프라고 한다. 그래프 상의 간선이 많은 노드로부터 군집화를 실행한다면 보다 효율적이고 쉬운 군집화가 실행될 것으로 보인다.

3-3) 데이터의 소개

Stanford Network Analysis Platform(SNAP)에서 제공하는 Facebook network의 일부인 Ego414 network를 사용한다. Ego414 network는 150개의 노드, 1873개의 간선으로 이루어져 있다. 각 노드에 대해 분류된 105개의 세부특성이 존재한다. 또 세부특성이 있으면 1, 없으면 0으로 표기한다.

3-4) 결과의 측정 방법

각각의 군집화를 실행한 후 각 군집화 별로 Ground Truth와 비교하여 각 군집화 별로 F1 score를 비교하여 성능을 측정한다.

3-5) K-means 구현

우선, 사용할 클래스/그룹을 선택하고 각각의 중심점을 무작위로 초기화한다. 중심점은 각 데이터 요소 벡터와 동일한 길이의 벡터이다. 각 데이터 포인트는 해당 포인트와 각 그룹 센터 사이의 거리를 계산 후 다음 그 포인트가 가장 가까운 그룹에 속하도록 그룹화하여 분류한다. 이렇게 분류된 점을 기반으로 그룹의 모든 벡터의 평균을 취하여 그룹 중심을 재 계산한다. 설정된 반복 횟수 동안 또는 그룹 중심이 반복하는 사이에 많이 변하지 않을 때까지 이 단계를 반복한다.

K = 군집 수, C = 중심점, X = 모든 노드

3-6) Neo-K-means 구현

Matlab을 사용하여 mantis graph를 구현 한 후 mantis graph 를 사용하여 NEO-K-means를 실행한다. 아래의 수식에서 의 값을 조절하여 겹침의 비율과 이상점의 수치를 조절하여 최적의 군집화 상태를 찾으면 된다.

s.t

U = , 만약 가 군집에 포함될 경우, 포함되지 않을 경우 0.

= 클러스터의 개수

= 클러스터의 크기

3-7) The Girvan-Newman Method

Girvan-Newman 알고리즘은 betweenness centrality가 높은 연결부터 한 개씩 한 개씩 끊어 나간다. Community들간의 연결을 한 개씩 끊어 나가는 것이다. 그리고는 그 때마다 나누어지는 그룹에 대해 Modularity를 계산한다. 그리고 마지막 연결까지 끊었을 때, 지금까지 나왔던 Modularity 중 가장 큰 값을 가지는 지점이 최상의 Community 검출이 이루어진 상태라고 가정한다.

3-8) spectral clustering

[4]Spectral clustering은 그래프 분할 기반의 군집분석 기법이다. Spectral clustering을 위해서는 우선 각 데이터 개체를 노드(node)로 하고 개체 간의 유사도를 간선(edge)의 가중치로 하는 그래프 G(V,E)를 구성한 다. 이렇게 구성한 그래프에서 일부 에지를 제거하여 k개의 서로 고립된 성분 그래프로 분할하는데, 이 때 동일한 성분 그래프 내 노드 간의 유사성의 합을 최대화하는 동시에, 서로 다른 성분 그래프에 속한 노드 간의 유사성의 합을 최소화하는 형태로 그래프를 분할한다

V = 전체 노드의 집합

k = 분할된 그래프의 수,

번째 분할된 그래프의 노드의 집합( =1…k)

번째 노드와 번째 노드 간의 유사도

두 집합 A, B의 유사도

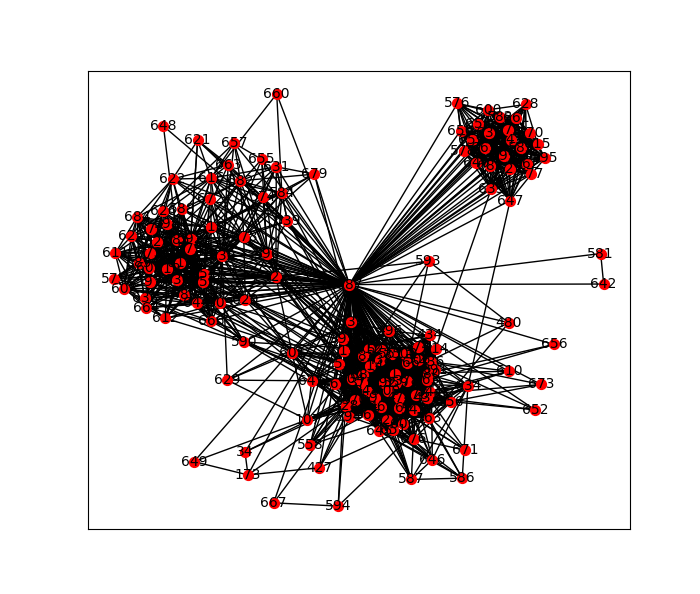
4. 구현 및 결과분석

4-1) Ground Truth

주어진 데이터 상에서 총 7개의 군집이 존재하며 주어진 데이터는 소셜 그래프로 군집 간에 중복되는 노드가 존재한다. 제일 큰 군집에는 총 57개의 노드가 존재하며 모든 노드가 중복을 이루는 1개의 군집이 존재한다.

4-1) Graph 구현

먼저 Connected component graph를 구현하기 위해 모든 노드들과 연결된 ego node를 설정 한 후 Python library중 하나인 network를 통해 그래프를 형성 한다. 아래의 그림에서 보이기에 ego node를 중심으로 크게 3개의 군집으로 나누어 보이는 것을 확인 할 수 있다.



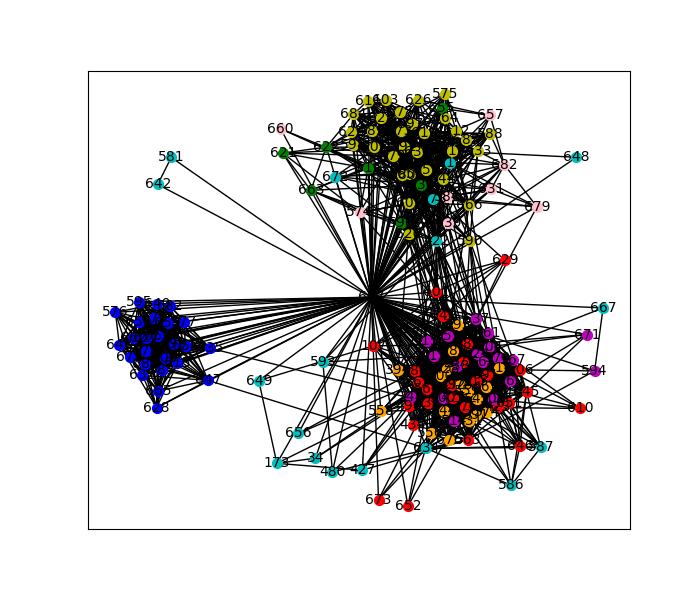
<그림1>

4-2)Neo-K-Means 결과

먼저 Linux 환경에서 metis graph를 형성한 후 , cluster의 개수는 Ground Truth와 같은 7개로 설정하였다. 이 경우에 가장 높은 F1-scroe를 기록하였기에 이를 토대로 clustering을 진행하였고 그 결과는 아래 <그림2>와 같은 결과를 보였다. <그림2>상의 옅은 핑크색이 중복을 허용하여 6번과 7번 군집을 중복되게 포함하는 노드들이다. 하지만 그림상 빨강색, 보라색, 주황색 노드들이 혼재되어 나타나는 현상을 보이는데 이는 중복되는 노드로 표현되어야 하는 노드들이 그러지 못하고 각각의 군집으로 나타나 일어난 현상이다. Ground truth 상에는 하나의 군집으로 표현이 되어진다. 중복을 조금 더 확실히 표현할 수 있다면 F1-score가 매우 높아질 것으로 기대가 되어진다.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 | Cluster 6 | Cluster 7 | Avg  F1-Score |
| Neo-k-means | 0.1765 | 0.4595 | 0.2581 | 0.3158 | 0.9804 | 0.400 | 0.3856 | 0.4251 |

<표1>



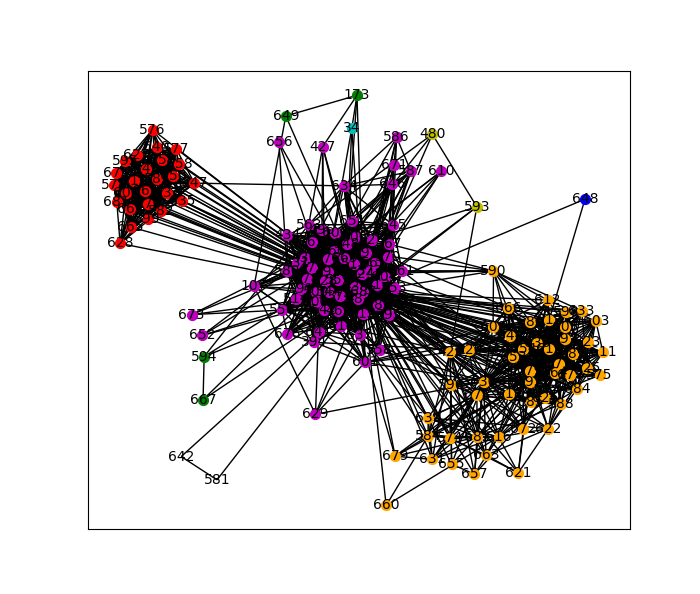
<그림2>

4-3) Girvan-Newman method 결과

Girvan-Newman method의 경우 Community가 총 7개가 되도록 설정한 후 진행하였다. 각 각의 노드 별로 Betweenness를 계산하여 edge를 끊어내는 작업을 수행하여야 했기에 실행시간이 길었다. 결과는 아래 <그림3>과 같았다. Graph 구현단계에서 직관적으로 볼 수 있었던 큰 3개의 경우로 군집화가 진행된 것을 확인 할 수 있다. 하지만 <표2>에 보이는 바와 같이 Ground truth에 존재하는 3개의 군집에 대해서는 전혀 찾아 내지 못하는 결과를 보였다.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| F1-score | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 | Cluster 6 | Cluster 7 | Avg  F1-Score |
| Girvan-Newman | 0 | 0.8780 | 0.4705 | 0 | 0.9804 | 0 | 0.9074 | 0.4622 |

<표2>



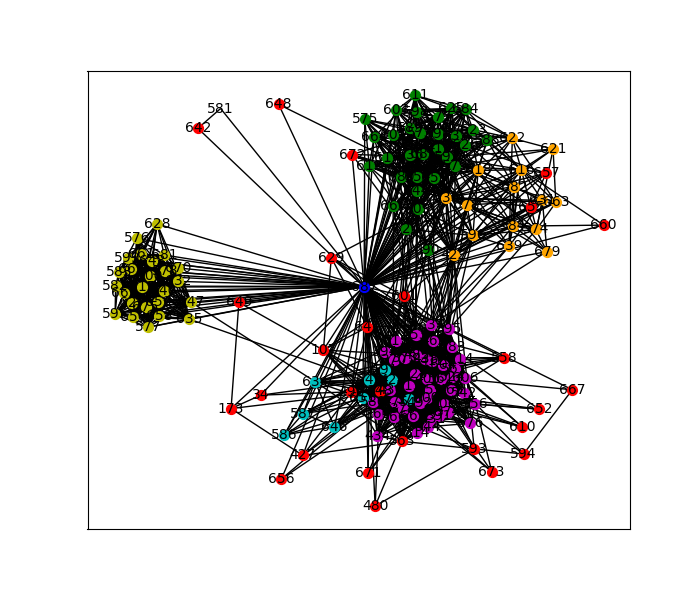
<그림3>

4-4) Spectral clustering 결과

Spectral clustering의 경우 총 군집의 개수는 7개로 설정한 후 얻은 결과는 아래의 <표3>과 <그림4>와 같다. <그림4>에 빨갛게 표현된 많은 노드들이 실제 Ground truth 상에 존재하지 않는 noise이다. 하지만 Spectral clustering의 경우 noise를 하나의 군집으로 묶는 현상이 발생하며 특정한 4번 군집에 대해서는 찾아내지 못하는 결과를 보였다.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 | Cluster 6 | Cluster 7 | Avg  F1-Score |
| Spectral  Clustering | 0.1176 | 0.8485 | 0.45 | 0 | 0.9804 | 0.5833 | 0.6966 | 0.5252 |

<표3>



<그림4>

4-5) 기법들에 대한 결과 분석

위의 3가지 Clustering 기법에 대한 결과를 각각 분석해 보면 평균 F1-scroe가 가장 큰 기법은 Spectral Clustering으로써 0.5252의 값을 기록하였고 그 다음으로 Girvan-Newman method가 0.4622, 마지막으로 Neo-k-means가 0.4251의 F1-score 값을 기록하였다. 본래 군집 간의 중복이 있어야 함으로 Neo-k-means가 제일 높은 F1-score를 기록할 것이라고 예상이 되었지만 실제 data상의 이상치가 너무 많이 존재하며 이로 인해 이상치들을 제거하지 못하고 중복이 되는 노드들을 중복이 허용되지 못한 채로 군집화 하여서 이런 결과가 나타내게 된 것으로 생각이 되어진다. 하지만 다른 기법과는 다르게 모든 Ground truth에 대해 군집이 존재하게 결과 값이 도출되었으므로 안 좋은 기법이라고 할 수는 없다. 그러므로 세분화된 군집화를 하기 위해서는 Neo-k-means 기법이 오히려 더 좋아 보인다고 할 수 있다. 이와 반대로 Spectral clustering의 경우 noise로 취급되어지는 data에 대해 하나로 군집화 함으로써 F1-score 면에서는 매우 높은 값을 나타내고 있지만 이로 인해 Ground truth의 하나의 군집을 전혀 잡아내지 못하는 현상을 보이고 있다. 그러므로 모든 군집에 대해 군집화를 하지 못하더라도 높은 F1-score를 기대하고자 한다면 Spectral clustering을 하는 것이 바람직하다고 볼 수 있다. 마지막으로 Girvan-Newman method의 경우 Ground truth와 비교하여 절반에 해당하는 군집만을 나타낸다. 이는 그래프 상에 존재하는 간선들의 betweenness만을 가지고는 세분화하여 군집을 나타내지 못하고 큰 군집에 대해서만 대략적인 군집화를 시켜준다는 것을 알 수 있는 결과이다.

5. 결론

본 작품에서는 다른 노드와 노드 간의 연결의 유무만을 따져 군집화를 진행하였다. 이로 인해 원래 예상과는 매우 상반된 결과가 도출되어진 듯하다. 하지만 각 노드의 특성, 중요 간선의 유무, 노드 간의 Homophily를 모두 고려하여 간선을 추가하여 연구를 진행한다면 위의 결과 값들이 더욱 향상되어 나올 수 있다고 생각되어진다.

향후에는 Graph의 간선이 매우 적지만 노드의 개수가 많은 data, 혹은 노드의 개수는 매우 많지만 간선이 매우 적은 data에 대해 실제 적용해본다면 더 정확히 graph의 구조에 따라 적용해야 할 군집화 기법을 알 수 있을 것으로 기대되어진다.

6. 후기

데이터베이스, 소프트웨어 특강의 수업을 들으면 이론적으로 들었던 그래프 클러스터링 기법들을 실제 구현해보고 실행하기 위해 데이터를 그래프화 시키는 작업부터 보다 이상적인 수치를 얻기 위해 파라미터를 변형하고 실제 데이터와 비교 분석하는 작업은 생각보다 큰 시간을 필요로 하는 작업이었습니다. 하지만 이 과정 중 습득하는 여러 지식과 인내의 시간은 매우 뜻 깊었습니다. 앞으로 사회에 나아가 더 큰 data와 시간을 들이는 작업에 대해서도 이번에 수행한 작품작업이 매우 도움이 될 것이라 생각됩니다. 그리고 data를 필요에 맞게 군집화 하는 것에 대해 매우 큰 흥미를 느낄 수 있는 좋은 기회였던 것 같습니다.

6. 참고문헌

[1] MacQueen, J. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, 281--297, University of California Press, Berkeley, Calif., 1967.

[2] Non-exhaustive, Overlapping k-means J. J. Whang, I. S. Dhillon, and D. F. Gleich, SIAM International Conference on Data Mining (SDM), 2015.

[3] Girvan M. and Newman M. E. J., Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA 99, 7821–7826 (2002)

[4] Barnard, S., Pothen, A., and Simon, H. (1995). A spectral algorithm for envelope reduction of sparse matrices. Numerical Linear Algebra with Applications, 2 (4), 317 – 334.