소셜 네트워크에서의 노드 속성을 고려한 중첩 커뮤니티 탐지

모용철[1] 이범석[02] 이한빈[2] 황지영[2†]
[1]성균관대학교 수학과
[2]성균관대학교 컴퓨터공학과
{orcescout, scarletbs, mafp12, jjwhang}@skku.edu

Overlapping Community Detection in Social Networks with Node Attributes

Yongcheol Mo[1] Beomseok Lee[02] Hanbin Lee[2] Joyce Jiyoung Whang[2†]
[1]Department of Mathematics, Sungkyunkwan University
[2]Department of Computer Science and Engineering, Sungkyunkwan University

요 약

소셜 네트워크(social networks)는 노드와 간선으로 표현되는 그래프 모델로 나타낼 수 있다. 커뮤니티(community)란 주어진 그래프에서 노드 간에 비슷한 특징을 공유하는 응집력 있는 노드들의 부분 집합으로 정의되며, 그래프의 구조를 이해하는 데 있어 커뮤니티 탐지(community detection)는 중요한 역할을 한다. 기존의 커뮤니티 탐지 알고리즘들은 대부분 그래프 구조와 노드의 속성 중 한 가지만을 이용하였다. 본 논문에서는 그래프 구조와 노드의 속성을 모두 활용하는 새로운 커뮤니티 탐지 알고리즘을 제안한다. 이 방법론은 변수 중요도를 제공하는 지도학습 알고리즘으로 두 노드 사이의 속성 차이로부터 노드 유사도를 구하고, 이를 바탕으로 그래프에 간선을 추가한 다음, 그래프 구조를 이용하는 기존의 커뮤니티 탐지 알고리즘을 적용한다.

1. 서 론

소셜 네트워크(social networks)는 사용자를 노드(node)로, 사용자들의 간의 연결 관계(친구 관계 등)를 간선(edge)으로 나타내어 그래프 모델로 표현할 수 있다. 이러한 소셜 네트워크상에서는 비슷한 특징을 공유하는 사용자들 간의 응집력 있는 그룹이 존재하며, 이는 커뮤니티(community) 또는 클러스터(cluster)라고 불린다. 즉, 커뮤니티는 외부로의 연결보다 내부로의 연결이 더 많은 응집력 있는 정점의 집합을 의미한다.

커뮤니티 탐지는 네트워크 분석에서 그래프의 구조적 특성에 대한 정보를 제공하기 때문에 큰 의미가 있다. 또한, 그래프 시각화, 네트워크 형태의 빅 데이터 처리, 소셜 네트워크 분석에서의 링크 예측(link prediction) 등 다양한 응용 분야에서 활용된다. 특히, 거대 그래프를 분석하기 위해 분할 정복 방식(divide and conquer)을 채택할 경우, 커뮤니티 탐지는 분할(divide) 기법으로 활용될 수 있으므로 그래프 마이닝 및 빅 데이터 분석에 근간이 되는 기술 중 하나로 여겨진다.

커뮤니티 탐지는 그래프 상의 노드들에 대한 클러스터링(clustering)으로 볼 수 있으며, 그래프의 구조(graph structure)와 노드의 속성(node attributes)이 클러스터링에 이용된다. 지금까지의 연구들은 주로 노드의 연결 구조만 고려하거나, 노드의 속성만을 고려하는 방향으로 이루어져 왔으며, 두 가지 요소에 대해서 상호 독립적으로 커뮤니티를 탐지하였다. 그 이유로는, 인접 행렬(adjacency matrix)로 주어지는 그래프 구조와 벡터로 표현되는 노드의 속성이라는 두 이질적인 정보를 통합하기 위한 표준적인 체계가 없었고, 매우 큰 규모의 네트워크에 대한 적용 가능성, 즉 확장성(scalability)을 위한 계산 복잡도 개선 문제 등이 있기 때문이다.

본 논문에서는 그래프 구조와 노드의 속성 모두를 고려한 커뮤니티 탐지 알고리즘을 제안한다. 변수중요도를 제공하는 지도학습 알고리즘(LASSO[1], Ridge Regression[7], Random Forest[2] 등)을 사용하여, 두 노드 사이의 속성 차이로 간선의 유무를 예측하는 모델을 훈련한다. 이 과정에서 얻어진 변수 중요도와 노드 간 속성 차이로 간선 가중치(edge weight)를 구하고, 기존 그래프의 인접 행렬과 결합하여 새로운 융합 행렬을 만든다. 이렇게 얻어진 융합 행렬에 그래프 구조를 이용하는 기존의 커뮤니티 탐지 알고리즘을 적용함으로써 간단한 방식으로 커뮤니티 탐지에 네트워크 구조와 노드의 속성이 모두 반영될 수 있도록 하는 통합적인 방법론을 제시한다.

2. 관련 연구

본 논문에서 제시하는 방법에 사용된 클러스터링 알고리즘인 NEO-K-Means는 기존 클러스터링 알고리즘에서는 별개의 작업으로 이루어진 이상점 탐지(outlier detection)와 중첩 커뮤니티 탐지(overlapping community detection)를 하나의 클러스터링 작업으로 통합한 알고리즘이다[3]. 또한, 네트워크의 각 노드의 속성 값만으로도 군집을 찾는 클러스터링 작업을 수행할 수 있으며 네트워크의 구조만으로도 군집을 찾는 커뮤니티 탐지 또한 가능하다.

† 교신저자 (Corresponding author)
본 연구는 미래창조과학부 및 정보통신기술진흥센터의 SW중심대학지원사업의 연구결과로 수행되었음 (R2215-16-1005)

NISE는 컨덕턴스 커뮤니티 점수(conductance community score)를 최적화하는 개인화 페이지랭크 클러스터링(personalized PageRank clustering) 기법을 기반으로 시딩(seeding)을 하는 단계를 거쳐 시드가 인근의 노드들을 대표하도록 하는 인접 인플레이션(neighborhood inflation) 단계를 통해 시드 팽창(seed expansion)을 하여 커뮤니티를 효과적으로 탐지하는 방법이다[4].

그밖에 네트워크 구조에 노이즈가 있어도 노드의 속성 값을 고려하여 네트워크에서 중첩 커뮤니티를 정확하고, 네트워크의 크기에 유연하게 탐지할 수 있는 알고리즘으로 CESNA 알고리즘[5]이 있으며 그래프 데이터에서 링크기반 거리(link-based distance)를 얻기 위해 간선(edge)이 존재하는 각 노드의 거리 메트릭(distance metric)을 학습하고 이를 이용하여 그래프 상의 모든 노드의 거리를 추정하는 DSHRINK 방법 등의 연구가 진행되었다[6].

3. 노드 속성을 고려한 커뮤니티 탐지

3.1 변수 중요도 측정

그래프의 모든 노드 쌍(node pair)에 대하여, 두 노드의 속성 차이와 두 노드 사이에 간선 존재 여부를 포함하는 노드 쌍 데이터를 만든다. 노드 쌍 데이터에서 변수 중요도를 제공하는 지도학습 알고리즘으로 속성 차이로 간선 존재 여부를 예측하는 모델을 훈련한다. 이때 모델의 정확성은 중요하지 않기 때문에 테스트 세트를 따로 분리하지는 않는다. 본 논문에서는 LASSO, Ridge regression, Random forest 등 3가지 알고리즘을 사용하였으며, 알고리즘에 대한 소개는 후술하였다.

3.1.1 LASSO Regression

일반적인 로지스틱 회귀 (logistic regression) 목적 함수에 균일화를 위해 coefficient norm을 추가한다. L1 norm의 성질로 인해 계수가 희소해진다는 장점이 있으나, 대신 상호 연관성이 매우 높은 변수들의 모임에서는 하나의 변수를 제외하고 계수의 값이 0이 되는 경향을 보인다.

3.1.2 Ridge Regression

Ridge regression은 L2 norm 을 사용하기 때문에 LASSO가 가지는 문제점이 없으며 소수의 계수가 큰 값을 갖는 것을 막는 경향이 있다.

3.1.3 Random Forest

결정 트리(decision tree) 기반의 알고리즘으로써, 배깅(bagging; bootstrap aggregating), random subspace sampling method를 이용해 조금씩 다른 결정 트리를 여러 개 생성하여, 각 트리의 예측 결과를 majority vote를 통해서 취합 후 최종 예측 결과를 생성한다. 변수마다 트리의 노드에서 분류 기준으로 쓰일 때 평균적으로 감소시킨 불순도를 측정해 변수 중요도를 얻는다.

3.2 융합 그래프 생성

3.1에서 예측모델을 훈련하며 변수중요도를 얻은 후, 노드 쌍 데이터에서 각 노드 쌍마다 속성 차이 값에 해당하는 변수중요도를 더하여 각 노드 쌍마다 간선 가중치를 구한다. 구해진 간선 가중치로부터 가중 그래프(weighted graph)를 얻는다.

이렇게 구해진 가중 그래프를 기존 그래프에 아무 처리 없이 합한다면, 그래프의 인접 행렬이 매우 조밀(dense)해져 알고리즘의 계산복잡도가 매우 증가하는 문제가 발생한다. 따라서 간선 가중치를 이진화(thresholding)한다. 즉, 노드 쌍의 간선 가중치가 특정한 임계 값 이상일 때만 기존 그래프의 해당 노드 쌍에 간선을 추가하여 노드 속성을 그래프 구조에 반영한다. 이로써 기존 그래프에 노드 속성으로부터 추출한 간선 가중치를 병합한 융합 그래프를 생성한다.

3.3 융합 그래프 상에서 커뮤니티 탐지

3.2에서 얻어진 융합 그래프에 기존의 그래프 기반 커뮤니티 탐지 알고리즘을 적용한다. 본 논문에서는 NEO-K-Means 알고리즘을 사용하였는데, 정확성과 확장성 측면에서 우수하고, K-Means 알고리즘에 기반하여 해당 알고리즘에 관한 연구 결과들을 적용하기 쉬우며, 비고갈(non-exhaustive) 및 중첩(overlapping) 커뮤니티 탐지를 지원하는 장점 때문이다. 하지만 실험 목적에 따라서 다른 커뮤니티 탐지 알고리즘 또한 적용할 수 있다.

4. 실험 결과

4.1 실험 데이터

본 연구를 수행하기 위해서 Stanford Network Analysis Platform(SNAP)에서 제공하는 페이스북 에고 네트워크(ego-network) 데이터를 사용하였다. 그 중에서도 414 에고 네트워크를 사용하였다. 표1에서 볼 수 있듯이, 414 에고 네트워크에는 각 노드의 속성이 총 18개 항목의 105차원의 데이터로 구성되어 있다.

명칭	노드 수	간선 수	속성 종류 수
414	160	3386	105

표 1 ego network 데이터 집합

4.2 속성별 중요도 측정

1) 서로 다른 두 노드의 공통된 속성과 간선 유무를 데이터 집합(이하 노드 쌍)으로 기록한다.
2) 각각의 노드 쌍의 각 카테고리 별로 동일하게 1의 값을 갖는 세부 속성의 개수를 조사한다. (예. Birthday 카테고리의 세부 속성 중에서 동일하게 1의 값을 갖는 세부 속성의 개수)

3) 본인 노드와 쌍을 이루는 경우를 제외한 노드 쌍 항목에 대해 중복된 속성의 개수로 이루어진 데이터 집합을 생성한다.
4) 목표 값으로써, 모든 노드 쌍에 대해 간선이 있으면 1, 없으면 0을 갖는 데이터를 생성한다.
5) 변수 중요도를 제공하는 지도학습 알고리즘을 통해 두 노드의 공통된 속성과 간선 유무를 예측하는 모델을 생성한다.
6) 각 속성과 간선 유무의 연관성을 구한다.

표2는 Random forest와 Ridge regression 모델을 사용하여 구한 변수의 중요도 중 가장 높은 중요도를 보인 5개의 속성을 보여준다.

ranking	random forest	ridge regression
1	edu-school	edu-school
2	locale	locale
3	edu-type	birthday
4	birthday	edu-concentration
5	location	location

표 2 중요도가 가장 높게 측정된 5개 속성

4.3 클러스터링 실험 결과

실험에서 지도학습 알고리즘으로 LASSO, Ridge regression, Random forest 3가지 알고리즘을 적용하였고, 대조군으로 그래프 구조만을 이용한 커뮤니티 탐지 결과를 설정하였다.

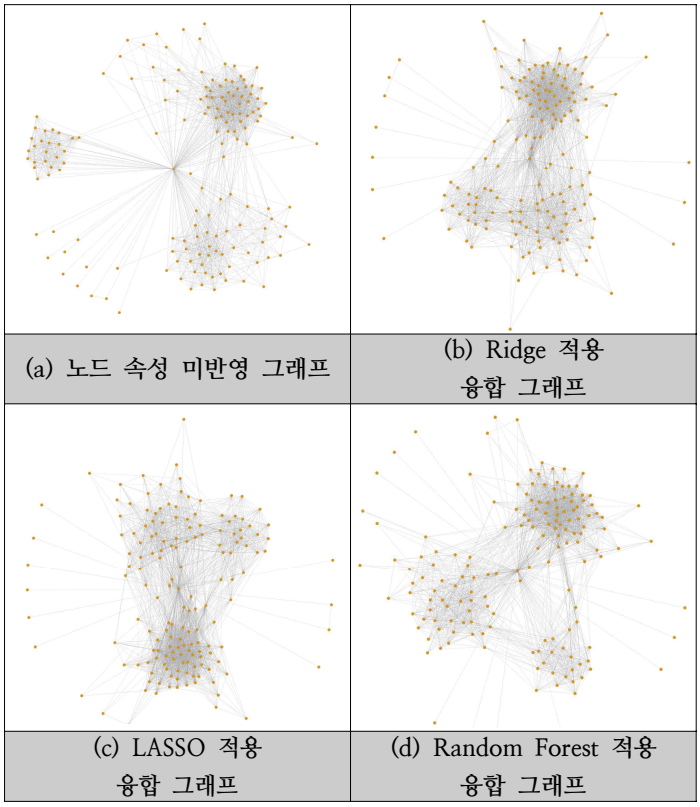

그림 1 각 regression 방법별 융합 그래프와 대조군의 비교

평가 기준은 SNAP에서 제공하는 실측 커뮤니티 자료(ground-truth community)에 대한 F1 점수(F1 score)로 설정하였다. alpha 값에 따른 F1 score 도 비교하였으며, alpha 값이 클수록 커뮤니티의 중첩을 많이 허용한다. F1 점수는 0에서 1 사이의 값을 가지며, 1에 가까울수록 실측 커뮤니티 자료와 결과와 가까운 것을 의미한다. 표3에서 볼 수 있듯이 본 논문에서 제시한 융합 그래프를 사용함으로써 보다 정확한 커뮤니티를 찾아낼 수 있었다.

alpha	노드 속성 미반영 그래프	Ridge 적용	LASSO 적용	Random Forest 적용
0.05	0.525	0.496	**0.531**	**0.559**
0.1	0.524	0.511	**0.526**	0.523
0.15	0.523	0.485	**0.526**	0.512
0.2	0.524	**0.545**	**0.527**	**0.525**
0.25	0.515	0.495	0.505	**0.537**
0.3	0.522	0.496	0.513	**0.54**

표 3 알고리즘별 NEO-K-Means 결과의 F1 score

5. 결론 및 향후 연구

본 논문에서는, 지도학습을 통해 노드 속성별 중요도를 측정한 후 이를 통해 노드 속성을 그래프 구조에 반영함으로써, 새롭게 생성된 융합 그래프에 기존의 확장성 높은 커뮤니티 알고리즘을 적용하여 노드 속성과 구조를 병합할 수 있는 간단한 형태의 중첩 커뮤니티 탐지 방법론을 개발하였다. 추후 알고리즘의 확장성을 개선하여 궁극적으로는 거대 소셜 네트워크 등 빅 데이터에도 적용할 수 있도록 하는 것을 목표로 하고 있다.

참 고 문 헌

[1] "Regression Shrinkage and Selection via the lasso," R. Tibshirani, *Journal of the Royal Statistical Society, Series B*, 1996.
[2] "Random Decision Forests," T. Ho, *ICDAR*, 1995.
[3] "Non-exhaustive, Overlapping k-means," J. Whang, I. Dhillon, and D. Gleich, *SDM*, 2015.
[4] "Overlapping Community Detection Using Neighborhood-Inflated Seed Expansion," J. Whang, D. Gleich, and I. Dhillon, *TKDE*, 2016.
[5] "Community Detection in Networks with Node Attributes," J. Yang, J. McAuley and J. Leskovec, *ICDM*, 2013
[6] "Community Detection in Incomplete Information Network," W. Lin et al., *WWW*, 2012.
[7] "Ridge regression: biased estimation for nonorthogonal problems," A. Hoerl and R. Kennard, *Technometrics*, 2000.