그래프의 링크 구조 및 노드 특성을 고려한 클러스터링 알고리즘

이우중[1] 황지영[2†]
[1]성균관대학교 수학과
[2]성균관대학교 컴퓨터공학과

{wj.lee, jjwhang}@skku.edu

Graph Clustering with Link Structure and Node Attributes

Woojoong Lee[1] Joyce Jiyoung Whang[2]
[1]Department of Mathematics, Sungkyunkwan University
[2]Department of Computer Science and Engineering, Sungkyunkwan University

요 약

그래프(Graph)란 노드(node)의 집합과 간선(edge)의 집합으로 이루어진 구조를 말한다. 사람을 노드로, 사람간의 관계를 간선으로 나타낸 그래프 모델로 소셜 그래프(Social Graph)를 들 수 있다. 그래프의 구조를 이해하는 데 있어 그래프 군집화(clustering)는 중요한 역할을 한다. 그래프 군집화에 대한 관련 연구에서는 노드의 특성(node attributes)만 고려하거나 링크의 구조(link structure)만 고려하여 그래프 군집화를 진행하였다. 하지만 노드의 특성과 링크의 구조를 따로 분류해서 고려하였을 경우에 완벽하게 군집화 하지 못한다는 단점이 있다. 이에 대한 해결 방안으로 노드의 특성과 링크의 구조 모두를 고려하는 방법이 있다. 본 논문에서는 노드의 특성과 링크의 구조 모두를 활용하여 새로운 군집화 알고리즘을 제안한다. 소셜 네트워크 등을 나타내는 그래프 모델에서 노드와 노드 사이에 간선이 형성되는 이유를 노드 쌍 사이 특정 특성의 유사도가 높기 때문이라고 생각할 수 있다. 따라서 간선이 형성 되어있는 노드 쌍에서 특성의 유사도를 구하고 이를 간선이 형성되어 있지 않은 노드에 적용하여 그래프에 반영되지 않은 간선을 추가하여 새로운 그래프를 형성한다. 새로 생성된 그래프를 기존의 커뮤니티 탐지 알고리즘에 적용하여 보다 나은 군집화를 진행한다.

1. 서 론

그래프란 노드의 집합과 간선의 집합으로 이루어진 구조를 말한다. 사람을 노드로, 사람간의 관계를 간선으로 나타낸 그래프 모델로 소셜 그래프를 들 수 있다. 이와 같은 그래프 모델은 크기가 거대하여 신속하고 정확한 분석을 하는 것이 현대 사회에서 중요한 이슈(issue)이다. 그래프를 분석하는 방법 중 하나로 그래프 군집화(clustering)을 꼽을 수 있다. 데이터를 나타낸 그래프에는 비슷한 특징을 공유하는 노드들 간의 응집력 있는 그룹이 존재하며 이는 군집(cluster)으로 불린다. 결국 그래프 군집화는 크기가 큰 그래프를 조밀(dense)하게 내부에서 연결 된 노드들의 집합으로 나누는 것을 뜻한다.

현재까지 진행 된 그래프 군집화 방법은 크게 두 가지로 나눌 수 있다. 하나는 노드의 특성만(node attributes)을 이용하여 군집화 하는 방법이고, 나머지는 링크 구조(link structure)를 이용하여 그래프를 군집화 하는 방법이다. 하지만 이 두 가지를 독립적으로 보고 따로 군집화 하는 것은 그래프의 단면만을 보고 군집화 하는 것이기 때문에 보다 정확한 군집화를 기대하기 어렵다. 따라서 그래프의 구조뿐만 아니라 노드의 특성 모두를 고려한 그래프 군집화는 노드의 특성과 그래프의 구조를 독립적으로 보고 진행한 그래프 군집화보다 더 나은 그래프 군집화를 할 것으로 기대된다.

본 논문에서는 노드의 특성과 링크의 구조 모두를 적용한 군집화 알고리즘을 제안한다. 간선이 있는 노드 쌍에서 간선이 형성되는 이유를 노드와 노드 사이에 특정한 특성의 유사도가 높기 때문이라고 생각하고 간선이 있는 노드 쌍에서 간선의 형성에 중요한 영향을 미치는 특성을 찾아낸다. 또한 그 특성들이 간선 형성에 영향을 미치는 정도를 수치화 한다. 이렇게 수치화 된 정보를 이용하여 간선이 없는 노드 쌍에 적용함으로서 새로운 간선을 찾아내고 이를 기존 그래프에 추가하여 새로운 그래프를 생성한다. 얻어진 새로운 그래프 구조를 기존의 커뮤니티 탐지 알고리즘에 적용하여 보다 정확한 군집화 알고리즘을 제시한다.

본 논문의 순서는 다음과 같다. 섹션2에서는 본 논문과 관련된 연구에 대하여 설명한다. 섹션3에서는 링크의 구조 및 노드 특성을 고려한 그래프 군집화 방법을 설명한다. 섹션4에서는 본 논문에서 사용한 데이터에 대해 설명하고 노드 특성간의 유사도를 구한 방법에 대해 설명한다. 섹션5에서는 결론을 맺고 향후 연구 방향에 대해 설명한다.

2. 관련 연구

NEO-K-Means[1]는 그래프 군집화 알고리즘 중에서도 중첩된 군집(overlapping cluster)을 탐지하고 이상점 탐지(outlier detection)를 동시에 진행 가능한 알고리즘 이다. 이 알고리즘에서는 군집의 중첩 정도를 나타내는 매개변수 알파(α)와 이상점의 비율을 나타내는 베타(β)를 사용한다. 이 알고리즘은 그래프의 노트 특성만을 이용하여 군집화 작업을 할 수 있고, 그래프의 구조만으로도 그래프 군집화 작업을 할 수 있다.

Graph Clustering Based on Structural/Attributes Similarities [2]는 노드의 특성을 반영하기 위해 특성에 대한 노드 만들고 노드 쌍에서 그 특성이 겹치면 두 노드와 특성노드를 연결하는 간선을 만든다. 이렇게 만들어진 그래프를 원래의 그래프와 융합한 융합그래프(augmented graph)를 만들고 무작위 행보(random walk)를 사용하여 노드간의 거리를 측정한다. 이후에 K-Medoids 군집화 방법을 이용하여 그래프 군집화를 한다.

Community Detection in Networks with Node Attributes [3]는 노드의 특성과 그래프 구조가 상호 의존한다는 관점에서 그래프 군집화를 진행한다. 이 논문에서 사용하는 CESNA알고리즘은 하나의 노드가 어떤 커뮤니티에 속할 때 다른 커뮤니티의 노드와 연결 될 확률이 낮다는 관념을 버리고 확률 모델을 구

† 교신저자(Corresponding author)

이 논문은 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (2016R1D1A1B03934766, NRF-2010-0020210). 또한, 과학기술정보통신부 및 정보통신기술진흥센터의 SW중심대학지원사업의 연구결과로 수행되었음(2015-0-00914).

한다. 이를 토대로 그래프 구조와 노드 특성을 이용하여 노이즈에 영향을 덜 받는 그래프 군집화를 한다.

Jaccard Similarity[4]는 이진수로 표현된 데이터의 유사도를 구하는 방법이다. 1로 표현된 값이 0으로 표현된 값보다 중요하다는 관점에서 두 대상이 모두 1의 값을 가진 것에 비중을 더 두어 유사도를 측정한다. 유사도를 측정하는 많은 방법 중에서 이진수로 표현된 데이터의 유사도를 측정하는 통용된 방법이다.

3. 링크 구조 및 노드 특성을 고려한 그래프 군집화

그림1은 본 논문에서 사용하는 그래프 군집화 알고리즘의 전반적인 흐름을 보여준다.

3.1 노드 쌍에서 특성의 유사도 측정

간선이 존재하는 노드 쌍에서 노드 쌍이 갖는 세부 특성의 유사도와 모든 노드 쌍에서 노드 쌍이 갖는 세부특성의 유사도의 차이로 각 세부특성이 간선 형성에 영향을 미치는 정도를 파악하고 이를 수치화한다.

3.1.1 간선이 존재하는 노드 쌍에서 세부 특성의 유사도 측정

간선이 존재하는 노드 쌍에서 중복되지 않는 세부 특성의 유사도를 측정한다. 각 세부특성이 '1'로 같은 노드 쌍의 개수를 구하고 세부특성사이의 비율을 수치화 한다.

3.1.2 모든 노드 쌍에서 세부 특성의 유사도 측정

모든 노드 쌍에 대해서도 위와 같이 노드 쌍에서 중복되지 않는 세부 특성의 유사도를 측정한다. 이 또한 각 세부특성이 '1'로 같은 노드 쌍의 개수를 구하고 세부 특성사이의 비율을 수치화한다.

3.2 간선 생성에 영향을 미치는 정도 측정

3.1.1과 3.1.2에서 각 세부 특성에서의 비율차이를 이용하여 세부 특성이 '1'로 같을 때 간선의 생성에 영향을 미치는 정도를 측정한다. 그 중에서 대부분의 영향을 차지하는 상위 25%의 세부 특성만을 고른다. 이 때 3.1.1에서 구한 비율과 3.1.2에서 구한 비율의 차이 값은 각 세부특성이 간선 형성에 미치는 영향이다.

3.3 임계 값 이상의 간선을 이용한 새로운 그래프 생성

각 노드 쌍이 그 세부특성 값을 모두 '1'로 가질 때 3.2에서 구한 세부특성이 간선 형성에 영향을 미치는 정도를 더하여 노드 쌍의 유사도를 구한다. 각 노드 쌍의 유사도가 임계 값(threshold)이상일 때 그 노드 쌍에 간선을 추가하여 새로운 그래프를 만들고 이에 대해 기존의 클러스터링 알고리즘을 적용한다.

3.3.1 노드 쌍의 유사도 측정

3.2에서의 각 세부특성이 간선 생성에 영향을 미치는 정도를 노드 쌍 모두 그 세부 특성을 '1'로 가질 때 합하여 각 노드 쌍의 유사도를 측정한다. 앞에서 구한 유사도를 간선이 없는 노드 쌍에 간선이 생성될 확률로 본다.

3.3.2 임계 값(threshold)이상의 유사도를 갖는 노드 쌍에 새로운 간선 추가

3.3.1에서 구한 확률이 임계 값(threshold)이상일 때 간선을 새로 추가하여 새로운 그래프를 완성한다.

3.4 NEO-K-Means를 이용한 그래프 군집화

3.3에서 새로 생성된 그래프를 NEO-K-Means[1]에 적용하여 보다 나은 그래프 군집화를 진행한다. NEO-K-Means는 비고갈(non-exhaustive) 및 중첩(overlapping)된 그래프 군집화에 적용할 수 있는 알고리즘이다. 하지만 비고갈 및 중첩 데이터가 아닌 데이터에 적용할 시 다른 그래프 군집화 알고리즘에 적용할 수 있다.

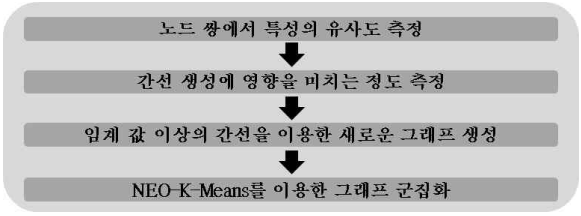

그림 1 링크 구조 및 노드 특성을 고려한 그래프 군집화 방법의 전반적인 흐름

4. 실험

4.1 실험 데이터 설명

본 논문에서 사용한 데이터는 Stanford Network Analysis Platform(SNAP)에서 제공하는 Facebook network의 일부인 ego414 network를 사용하였다. Ego414 network에서 노드의 개수는 160개, 간선의 개수는 1852개, 그리고 ground truth 군집 개수는 7개이다. 각 노드에 대한 분류된(categorical) 큰 특성이 존재하고 특성마다 세부 특성이 존재하는데 세부특성의 개수는 105개이고 이 세부특성은 익명화 되어있다. 세부특성은 세부특성이 있으면 '1', 없으면 '0'의 값을 갖는다. 이 때 '0'은 세부특성이 존재하지 않음을 의미할 뿐 다른 의미는 없다.

4.2 중복되는 세부 특성 삭제

Jaccard Similarity[4]를 이용하여 각 노드별로 세부특성의 유사도가 매우 높은(90%이상) 세부특성 중 하나를 삭제하였다. 이 때 삭제된 세부특성은 'Hometown'을 나타내는 전체에서 69번째 세부특성과 'name'을 나타내는 전체에서 89번째 세부특성이다. 따라서 중복되지 않는 세부특성은 103개이다.

4.3 간선형성에 중요 역할을 하는 세부특성 탐지

4.3.1 세부특성의 간선형성에 영향을 미치는 정도 측정

간선이 있는 노드 쌍으로부터 각 세부특성별로 세부특성이 '1'로 같은 것의 개수를 측정하고 비율(유사도)을 구한다. 모든 노드 쌍에 대해서도 이와 같이 측정한다. 식 1, 표1, 그리고 표2는 간선형성에 중요 역할을 하는 세부특성 탐지과정을 나타낸다.

$$S_{ij}(A_k) = \begin{cases} 1 & V_i(A_k) = 1 \text{ and } V_j(A_k) = 1 \\ 0 & otherwise \end{cases} \quad \cdots (1)$$

$V_i(A_k)$는 노드 i의 세부특성 A_k의 값을 의미한다.

표 1 간선이 있는 노드 쌍과 모든 노드 쌍에서 세부특성 A_k가 모두 '1'로 같은 것의 개수(N은 노드의 개수)

세부특성	A_k
간선이 있는 노드 쌍	$\sum_{(i,j)\in E} S_{ij}(A_k)$
모든 노드 쌍	$\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} S_{ij}(A_k)$

표2에서 간선이 있는 노드 쌍에서의 세부 특성이 모두 '1'인 것의 비율과 모든 노드 쌍에서의 세부 특성이 모두 '1'인 비율의 차이를 식(2)와 식(3)을 통하여 구한 후 정규화(Normalization)하여 이를 이용해 세부특성이 간선형성에 얼마나 영향을 미치는지 식(4)을 이용하여 측정한다.

표 2 간선이 있는 노드 쌍과 모든 노드 쌍에서 세부특성 A_k가 모두 '1' 로 같은 것의 비율. (d는 세부 특성의 개수, N은 노드의 개수)

세부특성	A_k
간선이 있는 노드 쌍	$\frac{\sum_{(i,j)\in E} S_{ij}(A_k)}{\sum_{k=1}^{d}\sum_{(i,j)\in E} S_{ij}(A_k)}$
모든 노드 쌍	$\frac{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} S_{ij}(A_k)}{\sum_{k=1}^{d}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} S_{ij}(A_k)}$

$$\frac{\sum_{(i,j)\in E} S_{ij}(A_k)}{\sum_{k=1}^{d}\sum_{(i,j)\in E} S_{ij}(A_k)} = E_k,\ \frac{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} S_{ij}(A_k)}{\sum_{k=1}^{d}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} S_{ij}(A_k)} = \overline{E}_k \cdots (2)$$

$$E_k - \overline{E}_k = a_k \quad \cdots (3)$$

$a_k > 0$인 k중에서 a_k의 값이 상위 25%인 k선택
상위 25%의 세부특성 A_k는 A_{k_m}으로 나타낸다.
N은 노드의 개수, d는 세부특성의 개수를 나타낸다.
d' 은 상위 25%로 선택된 세부특성의 개수를 나타낸다.

$$a'_{k_m} = \frac{a_{k_m}}{\sum_{m=1}^{d'} a_{k_m}} \quad \cdots (4)$$

4.3.2 간선 형성에 대부분의 기여도를 차지하는 상위 25%의 특성 추출

4.3.1에서 측정된 값 중에 간선 형성에 대부분의 기여도를 차지하는 상위 25%의 특성만 채택한다. 이 때 상위 25%의 특성이 간선 형성에 미치는 영향은 95%이다.

4.4 간선이 없는 노드 쌍의 유사도 측정 및 임계 값 이상의 유사도를 갖는 노드 쌍에 새로운 간선 추가

간선이 없는 노드 쌍에서 4.3.2에서 채택된 세부특성의 값이 모두 '1' 인 세부특성의 식(4)에서 구한 값을 모두 더하여 각 노드 쌍에서 간선이 생성될 확률, 즉 유사도를 식(5)를 이용하여 구한다. 식(5)에서 구해진 유사도가 임계 값(threshold)보다 높으면 새로운 간선을 추가한다.

$$\overline{S}(i,j) = \sum_{m=1}^{d'} V_i(A_{k_m}) \times V_j(A_{k_m}) \times a'_{k_m} \cdots (5)$$

$\overline{S}(i,j)$는 결국 노드 i와 노드 j사이에 간선이 생길 확률, 즉, 노드 i와 노드 j의 유사도를 나타낸다.

4.5 그래프 군집화

4.4에서 새롭게 생성한 간선으로부터 얻어진 새로운 그래프를 NEO-K-Means[1]에 적용하여 그래프 군집화를 진행한다. 이 때 NEO-K-Means에서 정의하는 베타(β)값은 0.013으로 고정하여 그래프 군집화를 한다. 기존 그래프에서 알파(α)값에 따른 F1 Score와 노드 쌍 유사도의 상위25%, 상위30%와 평균값 이상의 유사도를 갖는 노드 쌍에만 간선을 추가한 그래프에

[표 3] 유사도별 간선 추가 결과의 F1 score
Ground truth alpha = 0.15625 (단위 : %)

alpha	기존 그래프	상위 25% edge 추가	상위 30% edge 추가	평균 값 이상의 edge 추가
0.05	**55.2**	52.4	54.9	51.2
0.1	50.9	**53.3**	51.4	52.6
0.14	51.8	51.9	51.1	**55.4**
0.15	50.3	51.1	51.0	**52.8**
0.16	50.3	49.1	51.2	**54.6**
0.2	50.7	51.4	51.5	**52.5**
0.25	51.9	53.1	53.3	**54.1**
0.3	51.5	51.3	51.3	**53.4**

대해서도 기존 그래프와 같이 그래프 군집화를 진행한 F1 Score를 구한다. 표3에서 볼 수 있듯이 유사도가 평균값 이상인 노드 쌍에 간선을 추가하였을 때 실제 알파 값 근처에서 보다 나은 그래프 군집화를 이뤘음을 알 수 있다.

5. 결론 및 향후 연구

본 논문에서는 그래프의 두 요소인 노드의 특성과 그래프 구조 모두를 이용하여 그래프 군집화를 진행하였다. 노드 쌍이 유사할 때 노드 쌍 사이에 간선이 생긴다는 가정 하에서 노드가 갖고 있는 세부특성이 간선에 얼마나 영향을 미치는지 계산했다. 앞서 구한 수치를 이용하여 간선이 없는 노드 쌍의 유사도를 구하고 유사도가 임계 값 이상인 노드 쌍에 간선을 추가하여 새로운 그래프를 형성하였다. 새로운 그래프를 군집화 알고리즘인 NEO-K-Means에 적용하여 보다 정확한 그래프 군집화를 이루었다.

향후에는 이 알고리즘을 이용하여 소셜네트워크 뿐만 아니라 생체 네트워크, 테러 네트워크 등에 적용하여 새로운 노드가 생성 되었을 때 이 노드의 특성을 이용하여 보다 빠르게 어떤 군집에 들어갈지 예측할 수 있을 것으로 기대한다. 또한 개선된 알고리즘을 통하여 거대한 그래프까지 확장 가능한 알고리즘으로 개선하는 것과 노드의 특성이 일부 빠져있는 그래프에서도 적용될 수 있는 알고리즘으로 개선하는 것을 목표로 하고 있다.

참 고 문 헌

[1] "Non-exhaustive, Overlapping k-means," Joyce Jiyoung Whang, Inderjit S. Dhillon, David F. Gleich, SIAM International Conference on Data Mining(*SDM*), 2015.

[2] "Graph Clustering Based on Structural/Attribute Similarities," Yang Zhou, Hong Cheng, Jeffrey Xu Yu, *VLDB* 09, August 24-28, 2009, Lyon, France.

[3] "Community Detection in Networks with Node Attributes," Jaewon Yang, Julian McAuley, Jure Leskovec, The IEEE International Conference on Data Mining(*ICDM*),

[4] "A NEW SIMILARITY INDEX BASED ON PROBABILITY," David W. Goodall, International Biometric Society, Vol. 22, No. 4 (Dec., 1966), pp. 882-907