19장 사회연결망네트워크 애널리틱스

1. 서론

Facebook, Instagram과 같은 소셜 네트워크의 성장 → 많은 양의 데이터 생성, 사람 간의 연결 정보를 알 수 있는 정보 제공

SNS는 개체(ex: 사람)와 개체들을 연결하는 선으로 구성

기본 구성요소 (노드 (node, vertices, vertexes)
선 혹은 호 (connection)

2. 방향/무방향 네트워크

노드와 노드 간의 연결 관에 방향성을 나타내는가?

그래프 내의 연결선은 가중치로 그 강도를 표현 가능.

3. 네트워크 분석 및 시각화

- 그래프 레이아웃

네트워크 그래프에서 X, Y축은 큰 의미가 없으며 이보다는 노드의 크기, 선의 굵기, 레이블, 화살표(방향) 등이 그래프의 특성을 보여주는데 중요한 역할을 한다. → 동일한 네트워크라도 완전히 다른 모양으로 표현 가능

그래프 레이아웃을 결정하는데는 많은 선택사항이 있음. 그래프를 이해하기 위해서 다음의 4가지 요소 강조.

① 모든 노드는 눈으로 확인 가능해야한다.

② 모든 노드에서 그것의 연결도(degree)를 수치화할 수 있어야한다.

③ 모든 연결은 시작점과 끝나는 점이 명확해야한다.

④ 노드의 군집과 이상노드는 확인 가능해야한다.

두 가지 간단한 레이아웃 - 원과 그리드

원 - 모든 노드가 원 안에 늘어선 형태. 그리드 - 사각형 그리드 내 선들이 만나는 교차점에 노드가 위치

네트워크의 군집이나 고립노드를 덜다 효과적으로 표현할 수 있는 레이아웃이 존재 → 다양한 알고리즘을 통해 구현

↳ 이러한 알고리즘은 일반적으로 고정 임의 시작 구조 (fixed arbitrary starting structure), 무작위 조정, 물리적 속성과의 유사성, 반복 순서 등을 고려해 작동

- 인접 리스트

네트워크 그래프는 인접 목록 혹은 연결선 목록으로 부르는 데이터 표로 요약 가능.

두 열의 개체는 노드, 각 행은 두 노드 간의 연결.

- 인접행렬

개체의 동일한 관계는 행렬 형태로 표현 가능

인접행렬의 각 셀은 가장 왼쪽 열에 있는 헤더로부터 최상단 행에 있는 헤더로의 방향으로 연결여부를 표시.

- 네트워크를 이용한 분류(범주예측)와 예측(수치예측)

앞에서 봤던 분류, 예측, 군집화는 모두 정형화된 행렬형태의 데이터.

네트워크 분석 시 정형 데이터를 사용하기도 하지만 많은 경우 비정형 혹은 일부만 정형화된 데이터 사용.

↳ 네트워크 분석을 위해서는 적절한 측정척도를 이용해 비정형데이터를 정형데이터로 변환할 수 있어야한다.

이 때, 측정척도는 네트워크의 속성을 설명할 뿐 아니라 일반적인 데이터분석에 있어 입력값으로도 사용될 수 있다.

4. 소셜 데이터의 측정척도 및 분류

기본적인 네트워크 용어

연결선 가중치 (Edge weight) - 서로 연결된 두 개 노드 사이의 관련성도를 나타냄.

경로 (Path)와 경로길이 (Path length) - 경로는 노드 A에서 노드 B로의 이동 길. 경로 길이는 이동 중에 거쳐는 경로의 개수.

↳ 경로 길이는 최단거리를 의미하는 것이 대부분이지만 최소 비용 거리가 될 수도 있음

연결 네트워크 (connected network) - 모든 노드가 연결된 네트워크 → 연결 네트워크(경로에 상관없이)

클릭 (Clique) - 네트워크 구조의 일종. 각각의 노드가 다른 모든 노드에 직접 연결된 경우

고립노드 (Singleton) - 다른 노드와 연결되지않고 홀로 존재하는 노드.

- 노드 관점에서의 중심척도

네트워크 내 특정 노드의 중요도를 파악할 수 있는 방법 중 하나.

↳ 특정 노드가 보유하고 있는 연결선의 개수인 연결도 (Degree) → 많은 연결이 있는 노드일수록 중요하다고 간주

노드의 중심정도를 파악할 수 있는 또 다른 척도 - 근접성 (closeness)

↳ 네트워크 내 한 노드가 다른 노드와 얼마나 가까이 있는지로 정의. → 특정 노드와 연결된 모든 노드와의 최단거리를 구한후 이를 평균내어 구할 수 있다.

또 다른 척도는 중개성 (Betweenness)

↳ 특정 노드가 다른 노드의 최단 경로 상에서 중개자 역할을 얼마나 하는가 → 특정 노드 A를 제외한 모든 두 개 노드 사이의 최단거리에 A가 속해있는 비율의 평균

고유벡터 중심성 - 특정 노드로부터의 링크 개수와 이 링크들로부터 뻗어 나가는 연결개수의 합

↳ 0 - 1 사이에 존재. 0의 경우 중심이 없는 것. 1의 경우 최대 중심성

중심성은 네트워크에서 노드의 크기로 표현 가능 → 클수록 중요한 노드

- 자기중심네트워크

개인 간 연결관계를 분석함으로써 유용한 정보를 얻을 수 있음.

자기중심 네트워크 – 개별 노드 중심으로 모여 있는 네트워크.

연결도가 1인 자기중심 네트워크 – 특정 개별 노드로 모든 연결선이 집중된 경우

- 네트워크 측정측도

연결도 분포 (Degree distribution) – 전반적으로 노드가 몇 개의 선으로 연결되었는지를 알 수 있는 측도

밀도 (Density) – 네트워크에서 전체적인 연결성을 설명하기 위한 또 다른 측도. 노드보다는 연결선을 기준.

↳ 노드 간 총 연결 가능한 선 중 실제 연결된 선의 개수가 몇 개인지를 비율. (n개의 노드로 구성된 방향 네트워크의 경우 최대 $n(n-1)$개 / 무방향 네트워크의 경우 최대 $n(n-1)/2$)

$$\Rightarrow \begin{cases} \text{density(directed)} = \dfrac{e}{n(n-1)} \\ \text{density(undirected)} = \dfrac{e}{n(n-1)/2} \end{cases}$$

→ e는 연결선의 개수. n은 노드의 개수.

밀도값의 범위는 0(밀도↓)보다는 크며 1(밀도 매우큼)보다는 작거나 같다.

5. 네트워크 측도를 이용한 예측과 분류

- 연결선 예측

"네트워크가 주어졌을 때, 다음 연결선을 어디에 형성할까?"

↳ 예측 알고리즘은 모든 노드 쌍에 유사도 함수를 부여한 후 가장 높은 쌍을 다음 연결선으로 정함. (미리 연결된 것은 제외)

(15장 군집분석에 사용된 방법 사용)

유사도 계산에는 네트워크 측도뿐만 아니라 일반적인 데이터분석에 사용되는 변수도 함께 사용될 수 있다.

네트워크 연결선 예측에 사용되는 주요 측정 측도 : 최단 경로 공통된 이웃의 개수 연결선 가중치

- 개체 해석

여러 데이터분석에 출현한 인물이 동일인인지에 대한 여부는 거리측도를 이용한 최근접이웃방법이나 군집분석을 통해 확인 가능. (동일 개체가 여러 네트워크에서 동시 출현)

↳ 유클리드 거리는 특정 인물을 식별하는 네트워크 분석에 특화된 거리측도는 아니지만 인물의 속성을 설명하는 예측변수에 대한 거리를 측정한다고 보면 비슷하게 적용. 예측변수들에 대한 개체 해석 시 도메인지식은 각 변수의 중요도를 결정하는데 유용하게 활용될 수 있다.

네트워크 속성을 이용하면 개인 프로파일에 기반을 둔 거리 계산 가능. 나아가 전체적인 네트워크의 형태를 알 수 있다.

개체 해석은 고객 정보관리와 범죄에 유용하게 사용될 수 있다.

- 협업필터링

유사도 측도를 기반으로 비슷한 사람들을 찾고 이들의 속성을 이용해 특정인에게 특정 상품이나 서비스를 추천하는 기법.

↳ 네트워크 내 비슷한 객체끼리 묶거나, 영향력이 큰 개체를 탐지하거나 질병 혹은 컴퓨터의 확산경로를 밝혀준다.