**논문 요약**

**날짜: 2020.12.02**

**이름: 전우진**

|  |  |
| --- | --- |
| 논문제목 | An evolutionary algorithm approach to link prediction in dynamicsocial networks |
| 논문 요약 | 시간에 따라 변화하는 소셜 네트워크는 시간이 지남에 따라 역동성이 변화하는 그룹을 모델화하는 데 사용될 수 있다.  노드로 대표되는 개인은 네트워크에 들어가거나 나갈 수 있으며, 링크로 재전송되는 상호작용은 강화되거나 약화될 수 있다.  대부분의 네트워크 성장 모델들은 글로벌 자산을 포착하지만, 향후 누구와 연결될 것인가와 같은 특정 지역적 특성들을 포착하지는 않는다.  그러나, 국가 안보, 온라인 소셜 네트워크 작업 사이트(알고 있을 수 있는 사람), 조직 연구(잠재적인 협력자 예측)와 같은 애플리케이션에서 가장 가치가 있는 것은 바로 이러한 유형의 정보다.  링크 예측 전략은 크게 유사성 기반 전략, 최대우도성 예측 전략 및 확률론적 모델의 세 그룹으로 분류할 수 있다.  연결 예측을 위한 훈련에서, 연구자들은 지원 벡터 머신[27], 의사결정 나무[4], 임의의 숲[17], 감독되는 무작위 보행[6], 다중 계층 수용자 등을 포함한 감독된 학습을 사용해 왔다.  특히 알 하산 외. [27] 위상학적 특징과 노드별 특징을 모두 사용하여 여러 개의 슈퍼바이저 학습 알고리즘을 비교한다.  그들은 지원 벡터 머신(SVM)이 미래 링크 예측에 가장 잘 수행한다는 것을 발견했다.  SVM은 종종 예술 감독 학습 모델의 상태로 간주되지만, 주요 단점 중 하나는 커널 선택과 관련이 있다[28].  위와 같은 동기에서, 우리는 위상학 및 사용자별 정보를 포괄하는 링크 포디텍터를 제공하는 것을 목표로 한다. 이 정보는 빠른 정합성을 보이며 계산 복잡성으로 인해 매개변수 임계값이나 과소 샘플링이 필요하지 않다.  이 논문에서는 주변 유사성 측정과 노드별 데이터를 결합하기 위한 선형 모델을 수정하고, 정확하게 예측된 링크의 비율을 최적화하는 계수를 찾기 위해 진화 알고리즘을 사용한다.  모든 심-일률 지수가 동일한 중요성을 갖는다는 것을 미리 보장하기보다는, 이 선형 조합의 가중치가 공분산 행렬 적응형 진화 전략(CMA-ES)[29]을 사용하여 조정할 수 있도록 허용한다.  우리는 거대하고 역동적인 소셜 네트워크에 대한 링크 예측 문제에 초점을 맞추지만, 우리의 방법은 네트워크 유형과 독립적이며 다양한 생물학적, 인프라적, 소셜 및 가상 네트워크에 적용될 수 있다.  여기서 일반적으로 사용되는 유사도 지수를 16개 보여주지만, 다른 유사도 지수를 상호 교환하거나 이 연구에 포함된 지수에 추가할 수 있다는 점을 강조한다. 최근 몇 년 사이 소셜네트워크 분석 렌즈를 통한 뷰트위터 활동에 대한 관심이 급증하고 있다.  우리는 노드가 사용자를 대표하고 분석 기간 동안 왕복된 회답의 재송부 증거를 링크하는, 몇 주 단위로 구성된 이러한 네트워크의 진화를 검토한다.  우리의 데이터 세트는 2008년 9월 9일부터 12월 1일까지 비아테 트위터 가든호스 API 서비스를 수집한 5천 1백만 개 이상의 트윗으로 구성되어 있다.  이 컬렉션은 이 기간 동안 전송된 모든 메시지의 약 40%를 나타낸다(표 A1).  Bliss 외 연구진[34]에 의해 정의된 기준을 사용하여, 우리는 상호 응답 네트워크를 분석 대상 주 동안 그리고 이러한 개인이 상호 응답을 보이는 경우에만 노드 u와 v 사이에 링크가 존재하는 가중되지 않은 비방향 네트워크로 구축한다(그림 1).  이러한 네트워크는 N = 78, 296에서 N = 155, 753 노드(표 A2)까지 범위가 다양하다.  우리의 전체적인 결과는 16개의 유사성 지수로 구성된 진화된 예측 변수가 주어진 주의 RRN에 훈련이 발생했을 때 훈련 데이터의 다른 모든 조합 및 개별 지수를 능가한다는 것이다. 임의로 선택한 새 링크보다 개선의 중요한 요인을 관찰하며, 일반적으로 상위 N<20(그림 10)의 경우 104의 순서에 따른다.  교육 중 상위 5개 링크를 선택하기 위해 격리된 상태에서 사용할 경우 리소스 할당이 다른 유사성보다 뛰어나다는 것을 알 수 있으며 교차 검증(예측자)RA) 상위 10개(또는 그 이하) 링크 선택을 위한 단계.  우리는 조합된 예측 변수가 상위 N 링크 예측의 대부분의 선택에서 고립에 사용된 지수를 능가한다는 것을 관찰한다.  최근 네트워크 흐름 측정 방법 사용에 관심이 있기 때문에, 우리는 또한 우리의 예측 변수를 리히텐왈터 등이 제안한 방법인 길이 2의 경로로 프롭플로 제한되는 것과 비교한다. [17].  우리의 방법은 이 지수를 크게 능가한다.  마지막으로, 우리는 우리의 결과를 예비 의사결정 나무 분류기를 훈련하여 얻은 결과와 비교한다.  8 전형적으로, 균형 잡힌 클래스는 불균형 클래스와 관련된 문제를 극복하기 위해 이진 의사결정 트리를 훈련하는 데 사용된다[17,44,45].  링크 예측 방법은 경로 길이 2(예: 매우 불균형한 클래스)로 구분된 모든 노데노드 쌍에서 작동하므로, 균형 및 불균형 클래스와 관련된 문제와의 비교에 혼동을 피하기 위해 불균형 클래스에 대한 이항 결정 트리를 교육한다.  또한, 우리는 상위 N=7417links를 선택하도록 우리의 방법을 설정했는데, 이 링크는 이진 결정 트리 분류기로 식별된 것과 거의 동일한 수의 진위를 제공한다.  표 2는 이 비교의 결과를 보여준다.  이러한 상위 N 선택으로, 우리의 접근법은 정확도, 회수 등과 같은 몇 가지 지표에 걸쳐 약간 더 나은 성능을 발휘한다.  가장 주목할 만한 것은 우리의 정밀도는 우리의 이항 결정 트리에서 얻은 정밀도보다 거의 세 배나 높다.  우리의 거짓 발견률은 이항 결정률에 대해 얻어진 것보다 낮으며 이는 단순히 상위 N 링크를 조정하여 예측하는 잘못된 긍정의 수를 본질적으로 제한하는 상위 N 접근법 때문일 수 있다.  여러 연구에서 위상학적 유사성 지수와 노드 특정 유사성 지수의 조합이 링크 예측 노력을 크게 향상시킬 수 있다고 제안하였다[2,4,22–25,27].  우리는 트위터의 상호 리플렉션 네트워크 작업에서 이 주장에 대한 지지를 발견한다.  특정 주에 훈련이 발생한 실험의 경우, 우리는 결합된 "all16" 예측 변수가 위상학적으로만 예측 변수 "topo12"를 능가한다는 것을 발견했고 상위 N < 20에서 이 차이가 가장 뚜렷하다는 것을 발견했다.  우리의 조치는 트위터 링크 예측 분야에서 일하는 다른 조사원들에 비해 상당히 좋은 성과를 거두고 있다.  로위트 알. [23] 다음 행동을 예측하기 위한 노력의 일환으로 위상학적이고 개별적인 특정 유사성(단어와 주제 유사성)을 탐구한다. 우리는 우리의 결과를 프롭플로우나 이진 의사결정 나무와 같은 다른 접근법과 비교한다.  다른 사람들이 제안하고 관찰한 바와 같이, 크고 희박한 네트워크에서 링크 예측은 불균형한 클래스와 관련된 문제로 인해 어려움을 겪는다.  이와 같이, 우리는 이항 의사결정 나무와 같은 산업 표준과의 비교에서 우리의 결과를 해석할 것을 경고한다.  우리의 작업이 지도 학습의 척도를 결합한 다른 작업에 대한 개선과 비교될 수 있다는 것을 제안하는 것 외에, 우리는 투명하고 전달 가능한 방법을 제시한다.  미래 작업은 링크를 예측하기 위한 지리공간 자료[47] 또는 지역사회 구조의 포함을 포함할 수 있다.  시간의 경과에 따른 고리의 지속성 또는 쇠퇴, 또는 유량의 불일치를 고려하는 노력도 결실을 맺을 수 있다[48]. |
| 논문 의의 | 네트워크 클래스에 대한 가정이나 분석 대상 시스템에 대한 사전 지식이 필요하지 않다는 점에서 이는 우리 방법의 강점이다.  우리의 방법은 투명성과 향후 연결에 대한 좋은 예측 변수로서 기능하는 지수의 탐지를 제공하여 시간이 지남에 따라 네트워크의 진화를 촉진할 수 있는 가능한 메커니즘을 규명하는 데 도움을 준다.  이 연구에서 가장 흥미로운 측면 중 하나는 우리의 연결 예측 변수에서 크고 긍정적인 가중치를 가지도록 진화하는 유사성 지수의 검출이다. |
| 논문의 한계점 및 기타 | 최적의 모형 조합 유사성 지수는 선형적이지 않을 수 있으며, 이 모형 구조에 대한 우리의 가정은 우리 작업의 한계다. CMAES에서 계수를 초월한 세대간의 진화에서 균형 잡힌 클래스를 사용함으로써 우리의 방법론에 대해 개선될 수 있다.  이러한 전략과 다른 전략들을 통합하면 우리의 방법과 감독되는 다른 학습 접근법 사이에 더 통찰력 있는 비교를 할 수 있을 것이다. |