소셜 네트워크에서 정확한 부호 예측을 위한 특징 선택

김병찬*, 최범석*, 이원창**, 이연창**, 김상욱*,[1]
*한양대학교 컴퓨터소프트웨어학부
**한양대학교 컴퓨터소프트웨어학과
{kbcbang, midi9848, wonchang24, lyc0324, wook}@hanyang.ac.kr

Feature Selection for Accurate Sign Prediction in Social Networks

Byung Chan Kim*, Beom Seok Choi*, Won-Chang Lee**, Yeon-Chang Lee**, Sang-Wook Kim*
*Dept. of Computer Science, Hanyang University
**Dept. of Computer and Software, Hanyang University

요 약

부호가 있는 소셜 네트워크는 친구, 호감, 동의의 긍정적인 관계와 적, 불호, 반대의 부정적인 관계가 함께 표현된 네트워크이다. 이러한 네트워크를 활용한 대표적인 애플리케이션으로, 각 사용자의 관계가 긍정적인 관계인지 부정적인 관계인지 예측하는 부호 예측 문제가 있다. 이러한 부호 예측 문제를 해결하는 대표적인 방안은 네트워크의 구조적 특징들을 활용하는 것이다. 본 논문에서는, 실세계 데이터 집합들을 활용한 실험을 통해 기존 부호 예측 방법들에서 활용하는 각 특징이 부호 예측 문제의 정확도에 얼마나 기여하는지 분석하고자 한다.

1. 서론

사용자들 간의 관계를 하나의 네트워크로 표현하면, 각 사용자를 하나의 정점(vertex)으로, 두 사용자들 간의 관계를 하나의 간선(edge)으로 간주할 수 있다. 이러한 네트워크 상에서 사용자들 간의 긍정/부정 관계가 함께 표현된 네트워크를 부호가 있는 소셜 네트워크(signed social network)라고 한다[1].

이러한 부호가 있는 소셜 네트워크를 통해, 우리는 사용자들 간의 내재된 복잡한 관계들을 좀 더 명확하게 이해할 수 있다[1]. 따라서, 최근 부호가 있는 소셜 네트워크를 활용하는 다양한 애플리케이션들이 등장하였다[2]. 대표적으로, 사용자들 간의 관계가 긍정적인 관계인지 부정적인 관계인지 예측하는 부호 예측(sign prediction) 문제가 있다[3].

부호 예측 문제를 위한 대표적인 접근 방안은 네트워크의 구조적 특징들(topological features)을 기반으로 각 사용자들 간의 관계를 정의하여 활용하는 것이다[3]. 이를 위해, Li et al. [3]은 총 26 개의 특징들로 관계를 정의하고, 각 특징에 대한 가중치를 학습하여 부호를 예측하는 방법을 제안했다.

그러나, [3]에서는 이러한 특징들을 모두 함께 활용하는 것의 효과만을 분석하고, 각 특징의 독립적인 효과를 분석하고 있지는 않다. 따라서, 본 논문의 목표는 부호 예측의 정확도에 크게 기여하는 특징들이 무엇인지 확인하는 것이다. 이를 위해, 우리는 실세계 부호가 있는 네트워크들을 활용한 실험을 수행하여, 각 특징이 정확도에 얼마나 기여하는지 분석하고자 한다.

2. 구조적 특징 기반 부호 예측

[3]에서, 저자들은 사용자들의 관계를 총 26 개의 특징으로 정의한다. 이러한 특징들은 다음과 같이 5 개의 유형으로 분류될 수 있다.

(1) 균형 이론(balance theory): 두 사용자 간 공통 친구(pp)/적(nn)이 많을수록 친구가 될 확률이 높아지는 반면, 공통 이웃에 대해 서로 다른 관계(pn)를 많이 가질 수록 적이 될 확률이 높아진다. 해당 유형에는 다음 6 개의 특징이 포함된다: pp, nn, pn, pp_ratio, nn_ratio, bal_diff.

(2) 상태 이론(status theory): 사용자 i 가 사용자 j 를 양의 간선으로 가리킨다면, j 의 상태/지위가 i 보다 더 높다. 해당 유형에는 다음 4 개의 특징이 포함된다: sta_diff, sta_diff_p, sta_diff_n, sta_diff_ratio.

[1] 교신 저자

(3) 호혜성(reciprocity): 두 사용자가 서로 간선이 존재할 때 두 간선은 같은 부호를 가지는 경향이 높다. 해당 유형에는 다음 1 개의 특징만 포함된다: reciprocity.

(4) Rich-get-richer: 네트워크에서 활발한 활동을 하는 사용자들끼리 간선이 생길 확률이 높다. 해당 유형에는 다음 10 개의 특징이 포함된다: out_p, out_n, in_p, in_n, out_p_ratio, in_p_ratio, prprs, prnrs, nrnrs, nrprs.

(5) 클러스터링(clustering): 두 사용자 간 공통 이웃이 많을수록 간선이 생길 확률이 높다. 해당 유형에는 다음 5 개의 특징들이 포함된다: cn, Katz, Jaccard coefficient, Preferntial_attachment, Status_similarity.

3. 실험

3.1 데이터

본 논문에서는 부호가 있는 소셜 네트워크 데이터 집합인 Slashdot 과 Wikipedia 를 사용하여 실험을 수행하였다. Slashdot 은 82,140 개의 정점과 549,202 개의 간선(양: 425,072; 음: 124,130)을 가지며, Wikipedia 는 7,118 개의 정점과 193,694 개의 간선(양: 151,925; 음: 41,769)을 가진다.

3.2 실험 방법

먼저, 우리는 데이터 집합을 9:1 의 비율로 트레이닝 집합과 테스트 집합으로 구분하였다. 그 후, 트레이닝 집합만을 이용하여 사용자들 간의 관계를 앞서 소개한 특징들로 학습하고, 테스트 집합에 존재하는 사용자들 간의 관계를 예측하였다. 각 특징의 효과를 분석하기 위해, 학습을 수행할 때 사용하는 특징 조합을 바꿔가며 실험을 수행하였다.

성능 측정을 위해, 우리는 정확도(ACC)와 AUC 를 사용하였다. 정확도는 테스트 간선 중 올바르게 부호를 예측한 간선의 수로 계산된다. 또한, AUC 는 다음과 같이 계산된다:

$$AUC = \frac{1}{|\mathbb{P}||\mathbb{N}|}\sum_{a_i\in\mathbb{P}}\sum_{a_j\in\mathbb{N}} I\left(f(a_i) > f(a_j)\right),$$

여기서, $\mathbb{P}$와 $\mathbb{N}$은 각각 양의 간선들과 음의 간선들의 집합을 나타낸다. $f(a_i)$는 간선 a_i의 점수, $I(x)$는 x가 참이면 1, 거짓이면 0 을 갖는 함수를 의미한다.

3.3 실험 결과

우리는 모든 특징 조합 별 실험을 수행하였으나, 각 데이터 별로 유의미한 결과를 보이는 특징 조합들에 대한 결과만 선별하여 보인다.

표 1 과 표 2 는 각각 Slashdot 과 Wikipedia 에서의 실험 결과를 보여준다. 우리는 두 데이터셋에서 일관적으로 다음 2 개의 특징들이 정확도 향상에 기여를 많이 한다는 것을 확인하였다: prprs(양의 간선을 가리키는 비율과 양의 간선을 받는 비율), nrnrs(양의 간선을 가리키는 비율과 음의 간선을 받는 비율). 또한, Slashdot 에서는 다음 2 가지 특징들 또한 정확도 개선에 도움을 주었다: out_p_ratio(다른 사용자들을 가리키는 간선의 비율), reciprocity. 반면, Wikipedia 에서는 in_p_ratio(양의 간선을 받은 비율)이 도움을 준다는 것을 확인하였다. Wikipedia 에서는 두 사용자가 서로 간선을 가지는 경우가 적어 reciprocity 가 유의미한 효과를 보이지 못했다.

두 데이터의 주요 특징 조합은 다르나, 우리는 두 데이터 모두 일부의 주요 특징 조합만을 사용하더라도 모든 특징 조합(Full)을 사용하는 것과 유사한 정확도를 보인다는 것을 확인하였다.

표 1. Slashdot 에서 주요 특징 별 AUC 와 ACC

특징	AUC	ACC
out_p_ratio+prprs+nrnrs	0.873	0.81
out_p_ratio+prprs+nrnrs +reciprocity	0.89	0.829
full	0.899	0.834

표 2. Wikipedia 에서 주요 특징 별 AUC 와 ACC

특징	AUC	ACC
prprs	0.895	0.862
nrnrs	0.883	0.869
in_p_ratio	0.876	0.853
prprs+nrnrs+in_p_ratio	0.907	0.865
full	0.912	0.867

4. 결론

본 논문에서는 특징 기반 부호 예측 방법에서 사용하는 구조적 특징들의 효과를 분석하였다. 실세계 데이터를 활용한 실험을 통해, 우리는 각 네트워크의 일부 주요 특징 조합만을 사용하더라도 모든 특징 조합을 사용하는 것에 준하는 결과를 얻을 수 있다는 것을 확인하였다.

Acknowledgements

본 연구는 (1) 정부(과학기술정보통신부)의 재원으로 한국연구재단(No. NRF-2020R1A2B5B03001960), (2) 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학지원사업(2016-0-00023), 그리고 (3) 정부(과학기술정보통신부)의 재원으로 한국연구재단-차세대정보·컴퓨팅기술개발사업(No. NRF-2017M3C4A7083678)의 지원을 받아 수행된 연구임.

참고문헌

[1] J. Leskovec et al., “Signed Networks in Social Media”, In Proc. of CHI, 2010.

[2] J. Tang et al., “A Survey of Signed Network Mining in Social Media”, ACM Computing Surveys, 2016.

[3] X. Li et al., “Rethinking the Link Prediction Problem in Signed Social Networks”, In Proc. of AAAI, 2017.