Node2vec

: Scalable Feature Learning for Networks

이혜승

Contents

[*Introduction*]

Introduction

- 네트워크 분석
  1. **node classification**: network에서, node에 가장 가능성 있는 label을 배정하고자 함.
  2. **Link prediction**: network 의 a pair of nodes가 edge로 연결되는지 여부를 예측하고자 함. (connectivity 파악)
- Issue: 네트워크의 예측 문제에서, **feature vector representation** 이 필요.

  But 현재 기술: 손수 feature 추출함. 일반화 X (Feature engineering 에 많은 노력이 필요.)

  → Aim: feature 를 학습하는 objective function(목적함수)를 정의 & optimize it.(계산효율성과 예측 정확도의 밸런스를 맞춤)
- 현재의 Classic 방법들: 합리적인 목적함수 최적화x. 선형/비선형 차원축소 – expensive, bad 성능.

Introduction

- 대안: nodes의 지역 이웃의 보존하고자 하는 objective function 을 설계.(preserve local neigborhoods of nodes) SGD로 효율적으로 optimization 가능.
- 최근의 시도들이 효율적인 알고리즘을 제안하고있으나, network neighborhood에 대한 rigid notion에 기반함.

1. Homophily: 네트워크의 노드들은 그 노드들이 속한 Communities 기반
2. Structural equivalence: 네트워크에서 노드의 구조적 역할 기반

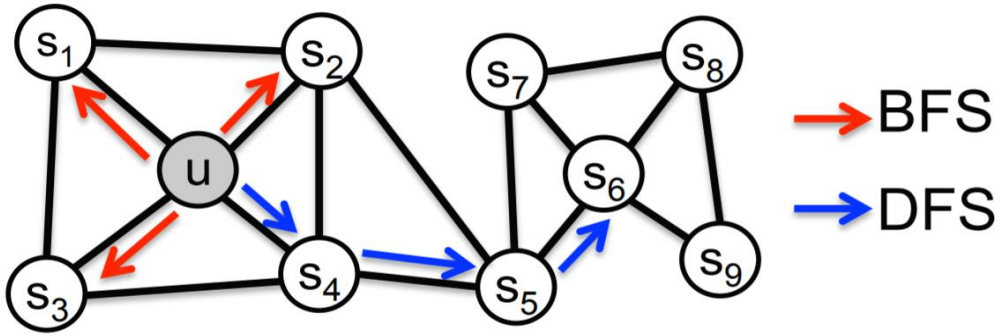


Figure 1: BFS and DFS search strategies from node u ($k = 3$).

1. 노드 u와 노드 s1은 같은 커뮤니티에 속함.
2. 노드 u와 노드 s6은 hub node로써의 구조적 역할.

Introduction

- Real-world network: 두 종류의 equivalence(homophily & structural)의 혼합을 나타냄.
- Thus, **두 가지 원리를 모두 준수하는 node representation을 학습하는 flexible algorithm**이 필요.

→ Feature 학습 알고리즘의 일반화 가능!

원리 1. 같은 네트워크 커뮤니티의 노드들을 가까이에 embed(homophily 기반)

2. 비슷한 역할을 공유하는 노드들이 비슷하게 embed(structural equivalence 기반)

→ Node2vec

Introduction

- 풀고자 하는 문제

: homophily와 structural equivalence, 두 가지 원리를 모두 준수하는, 효율적이고 자동화된 feature vector representation algorithm 의 개발.

- 관련 연구들은 이 문제를 어떻게 풀어왔는지

: 차원 축소(성능 bad). BFS, DFS

- 이전 연구들과 달리, 어떻게 문제를 해결했는지

: 유연한 feature learning 알고리즘 사용

[*Feature Learning framework*]

Feature Learning framework

- 두가지 standard 가정(최적화 문제를 쉽게 다루기 위해)

1. **조건부 독립**: 노드 u가 주어졌을 때, 다른 이웃 노드들이 관측될 확률은 모두 독립적.

   가정 하에서, 이웃 노드 집합($N_S(u)$)에 있는 모든 노드들의 각각확률을 곱한 것으로 쪼개짐!

$$Pr(N_S(u)|f(u)) = \prod_{n_i \in N_S(u)} Pr(n_i|f(u)).$$

2. **Feature space에서 대칭**: 노드 사이의 관계는 상호 같음.

   따라서, 모든 source – neighborhood node 쌍의 조건부 likelihood를 softmax로 모델링!

$$Pr(n_i|f(u)) = \frac{\exp(f(n_i) \cdot f(u))}{\sum_{v \in V} \exp(f(v) \cdot f(u))}.$$

Feature Learning framework

- 두 개의 가정 하에, objective function이 간소화된다.

$$\max_f \sum_{u \in V} \log(N_S(u)|f(u)) = \boldsymbol{\max_f} \sum_{\boldsymbol{u \in V}} \left[-\boldsymbol{\log Z_u} + \sum_{\boldsymbol{n_i \in N_S(u)}} \boldsymbol{f(n_i) f(u)} \right]$$

(좌변) skip-gram 아키텍처를 네트워크로 확장.

- skip-gram이 중심단어로부터 주변단어가 나올 log-likelihood를 최대화하는 것
- node2vec 에서는 중심 노드로부터 이웃 노드들이 나올 log-likelihood 를 최대화 하는 것

(우변) 간소화된 objective function.

노드별 분할 함수 Z_n은 큰 네트워크에서 계산 비용이 큼. 그래서 네거티브 샘플링을 통해 이를 근사함.

목적함수를 f 피처를 정의하는 모델 파라미터에 대해 SGA(Stochastic Gradient Ascent)를 사용해 최적화함.

Feature Learning framework

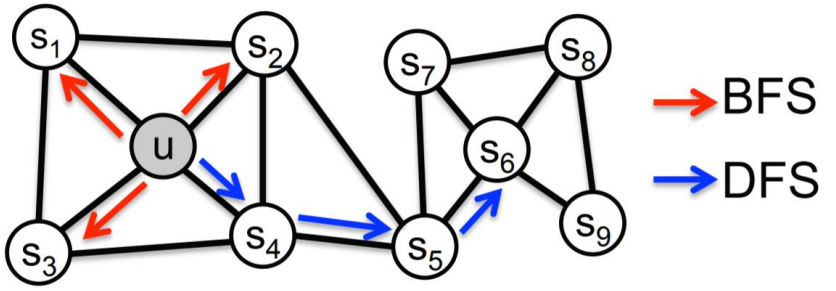


Figure 1: BFS and DFS search strategies from node u ($k = 3$).

- Sampling strategies

1. BFS(Breath-first Sampling): 너비 우선 탐색

   Structural equivalence 관점 반영

   $N_S(u)$를 소스 노드의 직접적인 이웃에 제한. 그림1에서 $N_S(u)$ = {s1, s2, s3} , k=3

2. DFS(Depth-first Sampling): 깊이 우선 탐색

   Homophily 관점 반영

   소스 노드에서 점점 먼 거리에서 순차적으로 샘플링되는 노드들. 그림1에서 $N_S(u)$ = {s4, s5, s6}, k=3

→ BFS와 DFS는 search space관점에서 서로 다른 시나리오. But real-world에서는 두 개의 동등성 개념이 배타적X

→ 본 논문에서는, 유연한 neighbor sampling strategies를 설계

[*Node2vec*]

node2vec

- 본 논문에서는 두 가지 원리를 준수하는 유연한 알고리즘으로 node2vec을 제안.
- 네트워크에서 확장가능한 feature 학습을 위한 semi-supervised알고리즘.
- 직관적으로, d차원 feature 공간에서, nodes의 네트워크 이웃들을 보존하는 가능성(우도)을 극대화하는 feature representation을 반환할 것임. 우리는 2차 랜덤워크 접근방식을 사용하여, 노드의 네트워크 환경을 생성.
- Key contribute

Node의 네트워크이웃의 유연한 개념을 정의하는 것.적절한 개념을 고름으로써, node2vec은 노드들이 속하는 커뮤니티나 네트워크의 역할에 기반하여 nodes를 구성하는 representation을 학습. 효율적으로 다양한 이웃을 탐색하는 biased 랜덤워크의 family를 개발함으로써 이 방법을 성취. 이 알고리즘은 유연하고, 조정가능한 파라미터를 통해 탐색 공가간을 제어할 수 있게함.

→ 결과: 일반화, 두 가지의 equivalence 모델링, 직관적인 해석가능

node2vec

- NLP의 feature learning 발전

→ word와 같은 어떠한 discrete object들의 feature learning 을 가능하게 함.

- Ex) skip-gram: 단어 feature representations는 negative sampling으로 SGD를 활용하여 likelihood 목적함수를 최적화함으로써 학습하여, 비슷한 단어의 주변에서 비슷한 단어가 나타나는 경향.

- 네트워크를 "문서"로 표현함으로써 네트워크의 유추를 확립. 문서와 비슷한 방법은 순서가 정해진 단어들의 시퀀스인데, 네트워크로부터 노드 순서를 샘플링하여 네트워크를 순서가 정해진 노드로 바꿀 수 있음. 그러나 노드에 대해 가능한 샘플링 전략이 많아서, 학습된 feature representation마다 다른 결과를 보여줌. 사실, 우리가 보여주어야할 것 처럼, 모든 예측, 분류 문제에서 명확하게 잘되는 것은 없음. 이는 네트워크 노드 샘플링에 유연성을 제공하지못하는 prior work의 단점. 우리의 알고리즘 node2vec은 특정 샘플링 전략에 묶이지않고, 탐색하는 서치공간을 조절하게하는 파라미터를 제공함으로써, 이러한 한계를 극복.

node2vec

Random Walk

$$P(c_i = x \mid c_{i-1} = v) = \begin{cases} \frac{\pi_{vx}}{Z} & \text{if } (v, x) \in E \\ 0 & \text{otherwise} \end{cases}$$

소스 노드 u가 주어졌을 때, 고정된 길이 ℓ 로 Random walk를 시뮬레이션.

- c_i: walk의 i-번째 노드
- π_{vx} : 노드v와 노드x간의 비정규화된 transition probability(전이확률: state가 v에서 x로 이동할 확률)
- Z: 정규화 상수

node2vec

Parameters

2^{nd} order randomwalk를 2개의 파라미터 p, q로 정의한다.

1. Return parameter, **p**
   : 파라미터 p는 walk에서 즉시 재방문하는 노드의 likelihood를 제어
   - parameter p ↑, 새로운 노드로 가는 경향(이미 방문한 노드를 샘플링할 가능성 ↓)
   - parameter p ↓, 소스 노드에서 walk를 가깝게 유지

2. In-out parameter, **q**
   : 파라미터 q는 탐색이 안쪽으로 향하는 노드와 바깥으로 향하는 노드 간의 구별을 가능하게 함.
   - q > 1, 시작 노드에 대한 그래프의 지역적인 관점 얻음(BFS 관점)
     - 시작 노드에 치우쳐져 있음.
   - q < 1, 시작 노드에서 멀리 떨어진 노드를 방문하는 경향(DFS 관점)

node2vec

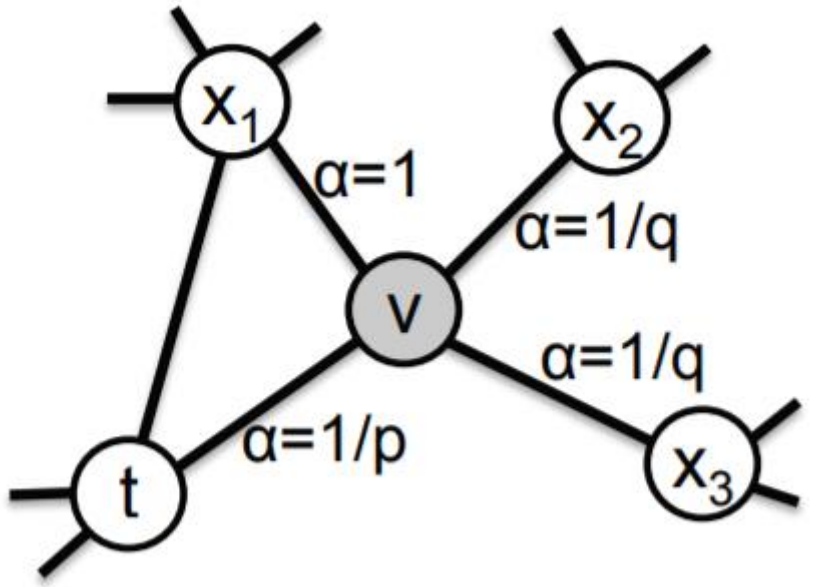
Parameters

t노드에서 2nd random walk진행

1. V에서 t로 전환
2. 다음 스텝으로는 t(내부 재방문), x1, x2, x3가능
   t로 돌아가는 경우, d(t, t) = 0
   d(t, x1) = 1
   d(t, x2) = d(t, x3) = 2

즉, p는 t로 되돌아오는 경우를 제어

q는 t에서 멀어지는 방향으로 갈 확률을 제어



역수로 계산되는 p가 작거나 q가 커질 경우, sampling이 근처 노드에서만.
P가 크거나 q가 작아질 경우, 멀리 있는 노드로 뻗어 나가는 식으로 sampling 진행.

"p와 q의 설정에 따라 node2vec은 다른 결과를 반환"

node2vec

benefits of random walks.

1. 공간적 효율성

   : 그래프의 모든 노드의 직접적인 이웃들을 저장하는 **공간복잡도**는 O(|E|).
   2차 순서 랜덤워크 - O(a^2|V|)

2. 시간적 효율성
   : 샘플 생성 과정에서 그래프 연결을 부과함으로써, 랜덤 워크는 편리한 매커니즘을 제공.
   샘플들을 재사용 → 샘플링 속도를 효과적으로 높임.

node2vec

Node2vec algorithm

랜덤워크에서, 소스 노드 u의 선택으로 인한 암묵적인 bias가 있었음. 우리는 모든 노드에 대한 representations를 학습하기 때문에, 이 bias를 상쇄한다.

모든 노드로부터 시작하는 고정길이 l의 랜덤워크를 시뮬레이션 함으로써. walk의 모든 스텝마다, 샘플링은 전이 확률 π_vs에 기반해서 이뤄짐. π_vs는 사전 계산이 가능하므로, walk가 진행되는 동안에는 노드의 샘플링이 O(1)시간내에 효율적으로 수행됨.

node2vec의 세가지 단계:
1. 전이확률을 계산하는 전처리
2.랜덤워크 시뮬레이션
3. SGD를 이용한 최적화.
: 순차적으로 수행됨. 각자 단계는 비동기적으로 병렬처리 및 실행되어 node2vec의 전체 확장성에 기여.

Algorithm 1 The *node2vec* algorithm.

LearnFeatures (Graph $G = (V, E, W)$, Dimensions d, Walks per node r, Walk length l, Context size k, Return p, In-out q)
  π = PreprocessModifiedWeights(G, p, q)
  $G' = (V, E, \pi)$
  Initialize $walks$ to Empty
  for $iter = 1$ **to** r **do**
    for all nodes $u \in V$ **do**
      $walk$ = node2vecWalk(G', u, l)
      Append $walk$ to $walks$
  f = StochasticGradientDescent(k, d, $walks$)
  return f

node2vecWalk (Graph $G' = (V, E, \pi)$, Start node u, Length l)
  Inititalize $walk$ to $[u]$
  for $walk\_iter = 1$ **to** l **do**
    $curr = walk[-1]$
    V_{curr} = GetNeighbors($curr$, G')
    s = AliasSample(V_{curr}, π)
    Append s to $walk$
  return $walk$

Introduction

- 풀고자 하는 문제

: homophily와 structural equivalence, 두 가지 원리를 모두 준수하는, 효율적이고 자동화된 feature vector representation algorithm 의 개발.

- 관련 연구들은 이 문제를 어떻게 풀어왔는지

: 차원 축소(성능 bad). BFS, DFS

- 이전 연구들과 달리, 어떻게 문제를 해결했는지

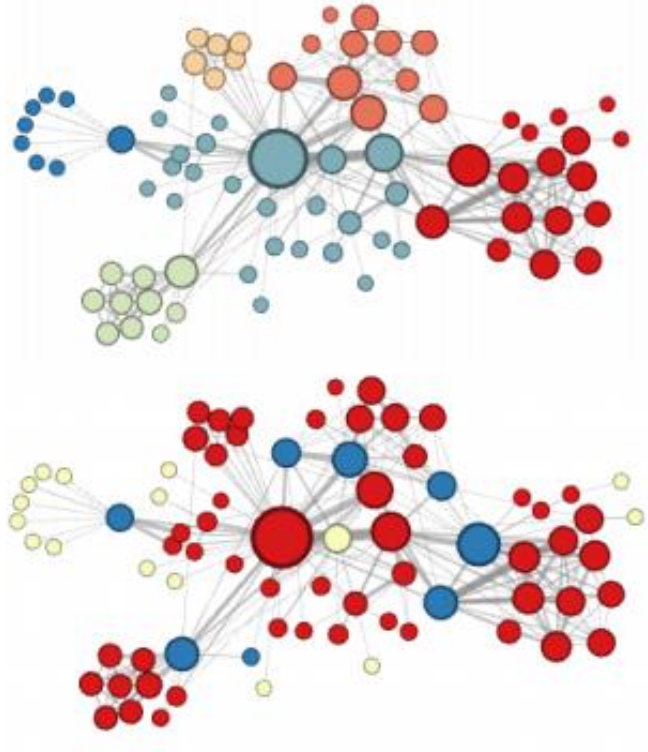
: 유연한 feature learning 알고리즘 사용, Node2vec

[*Experiments*]

Experiments

Case study 1.

- 레미제라블 소설에 '동시에' 등장하는 인물들에 대해 network를 모델링 → node2vec 적용 → feature representation → k-means 클러스터링



Node: 캐릭터

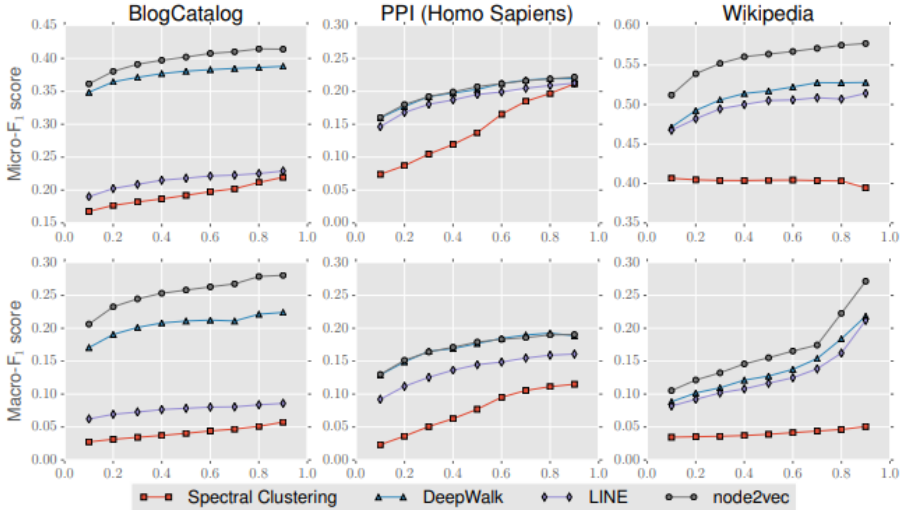
Edge: 같이 등장하면 edge로 노드 간 연결

1) p=1 q=0.5: homophily에 연관됨. 소설의 주요 plot에서 자주 상호작용하는 캐릭터들의 커뮤니티를 발견할 수 있음. 즉, 같이 등장하는 캐릭터들끼리 즉, community 내부의 캐릭터들이 비슷하게 배정된다.
2) P=1 q=2: 노드의 구조적 역할을 발견할 수 있음.소설의 다른 plot에서 각각 맡은 역할이 비슷하면 비슷하게 배정됨. 파랑색: sub-plots의 캐릭터와 브릿지 역할, 노란색: 주변부.

Experiments

Multi-label classification

"spectral clustering, DeepWalk, LINE과 Node2vec을 multi-label classification 문제에 적용해보니, **Node2vec의 성능이 좋았음.**"



- BlogCatalog에서, train:test = 7:3인 경우 deepwalk보다 성능이 가장 크게 우수.
- PPI에서, deepwalk와 성능 차이 별로 없음. p=4(이미 방문한 노드의 재방문을 피함)로 두어, 조금 우위에 있을 뿐. LINE에 비해서는 상승률을 보임.

Experiments

Link prediction

- Setting for Link prediction : 네트워크의 connectivity를 유지한 상태로 50%의 edge를 cut.

- 데이터:

- Facebook: node-사용자, edge-친구여부
- PPI: node- 단백질, edge-생물학적 상호작용
- arXiv: node- 과학자, edge-협업 논문 여부

- 실험결과: **Node2vec best.**
  :모든 네트워크에서 Deepwalk와 LINE을 능가
  Hadamard 연산자 사용시, 가장 안정성이 뛰어남

Op	Algorithm	Dataset		
		Facebook	PPI	arXiv
	Common Neighbors	0.8100	0.7142	0.8153
	Jaccard's Coefficient	0.8880	0.7018	0.8067
	Adamic-Adar	0.8289	0.7126	0.8315
	Pref. Attachment	0.7137	0.6670	0.6996
(a)	Spectral Clustering	0.5960	0.6588	0.5812
	DeepWalk	0.7238	0.6923	0.7066
	LINE	0.7029	0.6330	0.6516
	node2vec	0.7266	0.7543	0.7221
(b)	Spectral Clustering	0.6192	0.4920	0.5740
	DeepWalk	**0.9680**	0.7441	0.9340
	LINE	0.9490	0.7249	0.8902
	node2vec	**0.9680**	**0.7719**	**0.9366**
(c)	Spectral Clustering	0.7200	0.6356	0.7099
	DeepWalk	0.9574	0.6026	0.8282
	LINE	0.9483	0.7024	0.8809
	node2vec	0.9602	0.6292	0.8468
(d)	Spectral Clustering	0.7107	0.6026	0.6765
	DeepWalk	0.9584	0.6118	0.8305
	LINE	0.9460	0.7106	0.8862
	node2vec	0.9606	0.6236	0.8477

Table 4: Area Under Curve (AUC) scores for link prediction. Comparison with popular baselines and embedding based methods bootstapped using binary operators: (a) Average, (b) Hadamard, (c) Weighted-L1, and (d) Weighted-L2 (See Table 1 for definitions).

Reference

- 논문 원본, https://cs.stanford.edu/~jure/pubs/node2vec-kdd16.pdf
- 논문 리뷰 블로그 1, https://towardsdatascience.com/node2vec-embeddings-for-graph-data-32a866340fef
- 논문 리뷰 블로그 2, https://frhyme.github.io/machine-learning/node2vec/#hyper-parameter-p-and-q
- 논문 리뷰 블로그 3, https://analysisbugs.tistory.com/m/214

$$[\ Q\ n\ A\ ]$$