점진적 선호도 분해를 통한 그래프 기반 평점 예측 모델

이재현, 강성구, 유환조*
포항공과대학교

{jminy8, seongku, hwanjoyu}@postech.ac.kr

A Rating Prediction based on Graph Neural Network with Progressive Preference Decomposition

Jae Hyun Lee, SeongKu Kang, Hwanjo Yu*
POSTECH

요 약

본 논문은 최근 추천 시스템에서 높은 성능을 달성하는 그래프 신경망 모델들의 성능 하락 원인을 살펴보고, 이를 개선하는 모델을 제안하였다. 구체적으로, 최근 평점 예측 모델의 경우 사용자와 상품을 노드로, 관측된 평점을 링크로 하는 그래프를 구성한 뒤 그래프 신경망(Graph Neural Network, GNN) 적용해왔다. 그러나 해당 과정에서 평점 정보는 독립적인 관계 유형으로 취급되어 평점의 순서적 특성을 충분히 고려할 수 없었다. 본 연구는 평점 그래프를 사용자의 선호도를 고려하여 점진적으로 분해함으로써 사용자의 일반적인 관심에서부터 높은 선호도 정보를 순차적으로 고려할 수 있는 방법론을 제안한다. 그 결과 3 개의 실제 데이터 세트에 대한 실험에서 기존 연구들에 비해 높은 성능을 달성할 수 있음을 확인하였다.

Ⅰ. 서 론

현대 상업 시장에는 다양한 사용자들의 수요를 충족시키기 위해 수 없이 많은 상품이 생산되어지고 있으며, 때문에 사용자들이 이를 하나하나 비교하는 것은 불가능에 가깝다. 따라서 양질의 개인화 추천 시스템 개발은 개인의 편의성에서도, 기업의 수익성에서도 중요한 문제이다. 본 논문에서는 관측된 사용자와 상품 사이의 구매 및 평점 데이터를 이용해, 관측되지 않은 사용자와 상품 사이의 평점을 예측하는 평점 예측 모델을 다루고자 한다.

관련된 최근 연구들은 그래프 기반 모델들이 굉장히 효과적임을 시사하고 있는데, 이는 사용자와 상품을 노드로, 평점을 링크로 표현하는 그래프를 구성 한 뒤, GNN 을 적용하여 유사한 이웃들의 정보를 집계함으로써 각 노드의 표현을 이끌어낸다. 이러한 GNN 기반 모델은 링크를 단일 유형으로 취급하거나 [3], 독립적인 관계 유형으로 취급해왔다 [4,5].

기존 방법은 높은 성능을 달성했지만 다음과 같은 이유로 개선이 필요하다. (1) 일반적인 다중 유형 그래프(예: 감독, 배우)와 달리 평점 그래프에는 명확한 순서적 특성이 존재한다. 즉, 높은 평점은 해당 상품에 대한 사용자의 높은 선호도를, 반대의 경우 낮은 선호도를 의미한다. 때문에 GNN 의 메시지 전달(Message passing) 전략은 이러한 순차적 특성을 고려하여 모델링되어야한다. (2) 실제 시나리오에서는 평점 정보가 굉장히 희소하다. 추천 시스템에 활용되는 데이터는 기본적으르 희소성이 굉장히 높으며, 사용자가 상품을 클릭하고 구매한 뒤 최종적으로 평점을 매기는 시나리오에서 평점에 대한 정보는 수집되기 어렵다. 평점 정보가 제한되면, 각 평점 유형을 독립적으로 모델링하는 기존 방식은 소수의 데이터에 쉽게 편향될 수 있으며 이는 성능 저하로 이어진다. 때문에 소수의 평점 정보를 효율적으로 활용하는 것이 굉장히 중요하며, 동시에 평점 정보를 독립적으로 사용하기보다, 암시적 피드백 정보를 활용해 보완하는 과정이 필요하다.

본 논문에서는 기존 모델들의 성능 하락을 실험을 통해 확인하고, 평점 유형을 독립적으로 모델링하는 기존 방법론과 달리, 평점에 존재하는 순차적 관계를 고려한 모델을 제안한다. 제안 모델은 사용자의 명시적 피드백 정보가 부족한 환경에서 더 좋은 성능을 달성하여 현실 시나리오에서 적합한 모델임을 입증하였다.

Ⅱ. 본론

본 논문은, 어떤 사용자 i와 상품 j 사이의 상호 작용(예: 평점 1 점)을 원소(즉, M_{ij})로 갖는 평점 행렬 M이 존재할 때, 행렬의 누락된 항목을 예측하는 행렬 완성을 목표로 한다. 사용자는 구매한 상품 중 일부에만 평점을 매긴다는 것을 고려하여, 우리는 상호 작용을 2 가지로 정의한다. 첫째는 구매하였지만 평점은 매기지 않은 암시적 피드백 정보 (U)이며, 둘째는 명시적 피드백인 평점 $(\mathcal{R} = \{1, \dots R\})$이다. 이러한 행렬은 사용자와 상품 사이의 이분 그래프(Bipartite graph) G로써 표현할 수 있으며, GNN 기반 접근에서는 링크 예측(Link prediction)으로 표현된다.

다양한 평점 유형의 효과적인 사용은 높은 성능을 달성하는데 중요하며, 해당 섹션에서는 기존 GNN 기반 모델들이 이를 어떻게 모델링하였는지에 대해 간략히 살펴보고자 한다. 첫째로 LGCN [3]의 경우, 다양한 평점을 단일 유형으로 취급하는 것으로, 메시지 전달 레이어는 다음과 같은 형태로 구성된다.

$$e^{l+1}[i] = \sum_{j \in N(i)} \frac{1}{c_{ij}} e^{l}[j] \cdots (1).$$

$e^{l+1}[i]$는 l번째 레이어를 통과하고 난 뒤의 노드 i의 표현을, $N(i)$는 노드 i의 이웃 노드 집합을, c_{ij}는 정규화 상수를 의미하며, 본 논문에서는 대칭 정규화 $(N_t(i)||N_t(j))$ 를 적용하였다. 둘째로 [4, 5]의 경우, 각 평점을 독립적인 관계 유형으로 취급하는 Rating-specific transformation 전략을 사용하며 수식은 다음과 같다.

$$e^{l+1}[i] = \sum_{r\in\mathcal{R}} \sum_{j\in N_r(i)} \frac{1}{c_{ij}} W_r^l e^l[j] \cdots (2).$$

W_r^l 은 등급별 변환 행렬을 의미하며, $N_r(i)$ 는 노드 i 와 r 유형으로 연결된 이웃 노드 집합을 의미한다. 명칭대로 평점별로 다른 매개변수 및 이웃을 사용함으로써 유형별 패턴 정보를 학습할 수 있도록 구성되어져 있다.

우리는 사용자 선호도를 기반으로 점진적으로 분해된 그래프를 활용하는 GNN 기반의 추천 모델을 제안한다. 이는 평점 유형을 독립적으로 취급하는 기존 방식과 달리 평점의 순차적 특성에 따른 포함 관계를 직접적으로 활용하기 위하여, 주어진 이분 그래프 G 를 하위 그래프 집합 $\{G_r\}_{r\in\mathcal{R}}$ 으로 분해한다. G_r 는 평점이 r 이상인 링크로 구성되며, 각 하위 그래프들은 모두 동일한 노드 집합을 갖는다. 이 후, 각 그래프에 수식 (4)의 메시지 전달을 적용한다.

$$e_t^{l+1}[i] = \sum_{j\in N_t(i)} \frac{1}{c_{ij}} e_t^l[j] \cdots (3)$$

$e_t^{l+1}[i]$는 링크 유형 t에 대한 노드 i의 표현을 의미한다. L 개의 전파 레이어(Propagation layer)를 통과시킨 뒤, 우리는 각 레이어의 표현을 집계한 노드 표현을 획득한다 (즉, $h_t[i] = AGG(\{e_t^l[i] | 0 \le l \le L\})$). 기존 연구들과 달리 우리가 제안하는 점진적 분해 기법은 각 하위 그래프에 대해 더 많은 링크를 전달한다. 이는 데이터 희소성 문제를 완화시켜주며, 순차적 특성에 기반한 데이터 증강 기법으로도 해석할 수 있다.

추가적으로 선호도뿐 아니라, 사용자의 일반적인 관심 정보를 활용하기 위하여 우리는 U 유형의 링크을 이용해 G_I 를 구성한다. 구체적으로, G_I 에 대해 수식 (4)를 계산함으로써 사용자의 관심 정보를 획득하고, 이를 선호도 학습에서의 가이드라인으로 활용하며 수식은 다음과 같다.

$$\mathcal{L}_{IR} = \sum_{1\le i\le N} \frac{1}{|\mathcal{R}|} \sum_{r\in\mathcal{R}} \|h_r[i] - h_I[i]\|_2 \cdots (4).$$

N 은 전체 사용자와 상품 수를 의미한다. 이를 관심 정규화으로 지칭하며, 각 평점 유형을 통해 학습된 표현이 사용자의 근본적인 관심을 공유하도록하는 역할을 수행함으로써, $\{G_r\}_{r\in\mathcal{R}}$ 이 갖는 희소성 문제를 U 를 통해 보완한다.

우리는 최종 평점 예측을 위해 이중 선형 디코더를 사용한다. 우선 $\{h_r[i]\}_{r\in\mathcal{R}}$의 총합을 노드 i의 최종 표현으로 정의한 뒤, 다음과 같은 수식을 이용해 평점별 확률 분포를 연산한다.

$$z_{ij}^r = h[i]^T Q_r h[j] \cdots (5)$$

$$p(\widehat{M}_{ij}) = \frac{\exp(z_{ij}^r)}{\sum_{r\in\mathcal{R}} \exp(z_{ij}^r)} \cdots (6)$$

Q_r은 평점별 매개변수 행렬이며, 최종 평점은 확의 평균으로 정의한다 (즉, $\widehat{M}_{ij} = \sum_{r\in\mathcal{R}} rp(\widehat{M}_{ij} = r)$).

제안 모델은 선호도와 관심 정보를 모두 활용하기 때문에 평점 예측 손실 [5]와 쌍별 순위 손실 [2]를 함께 사용한다. 후자는 관측된 상호 작용과 그렇지 않은 상호 작용을 구분함으로써 사용자의 일반적인 관심을 효과적으로 학습할 수 있게 하며, 전자는 비교적 소수의 상품에 대한 세밀한 평점 예측 능력을 학습할 수 있게한다. 이를 위해 학습 데이터셋 $\mathcal{D} = \{(i,j,k) | M_{ij} \in \mathcal{R} \wedge M_{ik} \notin \mathcal{R} \cup \{U\}\}$ 와 다음과 같은 손실 함수를 정의한다.

$\mathcal{L} = -\sum_{(i,j,k)\in\mathcal{D}} (\sum_{r\in\mathcal{R}} I(r = M_{ij}) \log p(\widehat{M}_{ij} = r) + \alpha \log\sigma(o_{ij} - o_{ik}))$ … (7)

o_{ij} 는 사용자 i와 상품 j의 관심 정보를 의미하며 z_{ij}^r 의 총합으로 정의한다. 최종적으로 제안 모델은 수식 (5)와 (8)의 가중합으로 학습된다.

우리는 GroupLens 의 오픈 데이터셋 ML-100K, ML-1M 을 이용하여 제안 모델의 성능을 실험하였다. 해당 데이터는 영화에 대한 평점을 수집한 것으로 본 연구에서는 실제 시나리오를 반영하기 위해 학습 과정에서 평점 데이터의 일부 비율만을 필터링하여 활용하였다. (즉, 평점 비율={25%, 50%, 100%}). 제안 모델의 우수성을 입증하기 위해 MF[1]와 다양한 GNN 기반 모델 [3,4] 을 비교군으로 하여, RMSE(Root Mean Squared Error)를 측정하였다. 다만, 기존 연구들의 경우 U 유형의 링크를 고려하지 않기 때문에 실험의 공정성을 위하여 해당 유형을 추가한 모델(LGCN+, RGCM+)을 추가로 실험하였다. 표 1 에서 제안 모델은 가장 낮은 RMSE 를 달성하였으며, 경쟁 모델 대비 절반의 평점 정보만으로도 경쟁력있는 성능을 보인다. 이는 제안 모델이 현실 시나리오에서 큰 장점을 갖는 것으로 해석할 수 있다.

Table 1. RMSE results.

Rating-Frac	Method	ML-100K	ML-1M
100%	MF	0.9125 ± 0.0100	0.8994 ± 0.0018
	RGCN	0.9012 ± 0.0083	0.8417 ± 0.0004
	LGCN+	0.9431 ± 0.0061	0.8984 ± 0.0005
	RGCN+	0.9095 ± 0.0181	0.8613 ± 0.0161
	Proposed	**0.8726 ± 0.0009**	**0.8340 ± 0.0035**
	$Gain_{best}$	0.0286	0.0077
50%	MF	0.9495 ± 0.0070	0.9188 ± 0.0012
	RGCN	0.9610 ± 0.0110	0.8755 ± 0.0025
	LGCN+	0.9539 ± 0.0094	0.9104 ± 0.0011
	RGCN+	0.9450 ± 0.0185	0.8831 ± 0.0106
	Proposed	**0.9042 ± 0.0036**	**0.8561 ± 0.0021**
	$Gain_{best}$	0.0408	0.0194
25%	MF	1.0870 ± 0.0087	0.9374 ± 0.0007
	RGCN	0.9879 ± 0.0050	0.9135 ± 0.0025
	LGCN+	0.9732 ± 0.0114	0.9243 ± 0.0006
	RGCN+	0.9939 ± 0.0299	0.9122 ± 0.0024
	Proposed	**0.9317 ± 0.0025**	**0.8781 ± 0.0016**
	$Gain_{best}$	0.0415	0.0341

Ⅲ. 결론

각 평점 유형을 독립적으로 취급하는 기존 방법론들과 달리, 본 논문에서는 평점의 순차적 특징을 고려한 그래프 분해를 적용함으로써 GNN 이 평점을 효과적으로 활용할 수 있게 하였다. 또한 관심 정규화 과정을 이용해 학습 과정을 개선하였으며, 이러한 제안 방법의 효과를 다양한 실험을 통해 검증하였다. 특히, 평점 정보가 부족한 환경에서의 실험 결과는 현실적인 시나리오에서 본 모델이 굉장히 효과적으로 작동할 수 있음을 시사하며, 추후 평점 뿐 아니라 사용자의 다양한 행동(예: 장바구니, 리뷰)을 이용해 성능을 개선시킬 수 있을 것으로 보여진다.

참 고 문 헌

[1] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. Computer 42, 8 (2009), 30–37.

[2] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In UAI.

[3] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. LightGCN: Simplifying and Powering Graph Convolution Network.

[4] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In European semantic web conference. Springer, 593–607.

[5] Muhan Zhang and Yixin Chen. 2020. Inductive matrix completion based on graph neural networks. In ICLR.