计算机科学
COMPUTER SCIENCE

基于动态负采样的图卷积协同过滤推荐模型

马汉达, 方雨清

引用本文

马汉达, 方雨清. 基于动态负采样的图卷积协同过滤推荐模型[J]. 计算机科学, 2023, 50(11A): 230200149-7.

MA Handa, FANG Yuqing. Dynamic Negative Sampling for Graph Convolution Network Based Collaborative Filtering Recommendation Model [J]. Computer Science, 2023, 50(11A): 230200149-7.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230200149

基于动态负采样的图卷积协同过滤推荐模型

马汉达　方雨清
江苏大学计算机科学与通信工程学院　江苏 镇江 212013

摘　要　负采样对协同过滤算法的准确性有很大的影响。针对现有的图卷积网络缺乏对负采样策略的探索这一问题,提出一种基于动态负采样的图卷积协同过滤推荐模型(Dynamic Negative Sampling-Based Graph Convolution Collaborative Filtering Recommendation Model,DGCCF)。首先,为了能更灵活地适应不同图数据的需求,在图卷积网络中引入归一化参数来调节节点进行信息传递时邻域对其的影响;其次,提出一种动态负采样策略,从用户未交互过的物品节点中选取负样本集,经过图卷积后得到负样本评分,选取评分最高的负样本作为难负样本;最后,将得到的难负样本和正样本作为样本对输入贝叶斯个性化排序函数,对模型进行优化。在 Gowalla,Yelp2018 和 Amazon-Book 3 个公开数据集上与基线模型进行的对比实验表明,DGCCF 在多个评价指标下均优于现有的基线方法,在 3 个数据集上,召回率分别比最优基线提升了 0.3%,9.4%和 10.6%。

关键词:协同过滤;图卷积神经网络;负采样;推荐系统;评分预测
中图法分类号　TP391

Dynamic Negative Sampling for Graph Convolution Network Based Collaborative Filtering Recommendation Model

MA Handa and FANG Yuqing
School of Computer Science and Communication Engineering,Jiangsu University,Zhenjiang,Jiangsu 212013,China

Abstract　Negative sampling has a great impact on the accuracy of collaborative filtering algorithms,to solve the problem that the existing graph convolutional network lacks the exploration of negative sampling strategies,dynamic negative sampling-based graph convolution collaborative filtering recommendation model(DGCCF) is proposed. Firstly,in order to adapt more flexibly to the needs of different graph data,a normalization parameter is introduced in the graph convolutional network to adjust the influence of the neighborhood. Secondly,a dynamic negative sampling strategy is proposed,which selects a set of negative samples from the item nodes that the user has not interacted with,and after graph convolution gets the negative sample score,selects the negative sample with the highest score as the hard negative sample,and finally uses the obtained hard negative sample and positive sample as samplesets to input the Bayesian personalized ranking function to optimize the model. Comparison experiments with the baseline model on the three public datasets Gowalla,Yelp2018 and Amazon-Book show that DGCCF is superior to existing baseline methods under multiple evaluation indicators. For example,compared to the optimal baseline,its recall rate increases by 0.3%,9.4%,and 10.6% respectively on three dataset.

Keywords　Collaborative filtering,Graph convolutional neural network,Negative sampling,Recommendation system,Score prediction

1　引言

大数据时代的到来,一方面为人们的工作和生活提供了广阔的视野,另一方面也造成了信息过载[1]的问题,人们很难从中快速提取真正有用的信息。在此背景下,推荐系统应运而生,它能够有效地对信息进行过滤和筛选,从而帮助用户寻找符合其需求的信息,因此被广泛应用于购物、社交媒体、广告和网络搜索领域。

协同过滤是推荐领域最有效的算法之一,它通过学习用户和物品之间的交互关系获取用户的潜在偏好,从而预测未来用户可能交互的项目。目前比较热门的协同过滤算法是通过嵌入的形式将用户和物品的交互投影到低维的向量空间,再利用传统的矩阵分解算法或者深度学习算法对向量进行建模,从而得到潜在的交互偏好。然而在将高维的交互关系投射到低维的过程中,信息的丢失是难以避免的。现在的图神经网络可以直接对图结构的数据进行学习,而用户和物品之间的交互关系可以表示成用户和物品之间的异构二部图,从而避免了从高维到低维的信息损失。因此,基于图神经网络的推荐模型已经在推荐领域得到了广泛的应用,并且取得了比已有模型更高的准确性[2]。

基于图卷积神经网络(Graph Convolution Network,GCN)的推荐模型通过信息传递对图节点和边的信息进行迭代聚合,从而捕捉节点与节点、节点与边之间的关系,最终对节点之间的交互关系进行预测[3]。Wang 等[4]提出的基于邻域的协同过滤(Neighborhood-Based Collaborative Filtering,NGCF)首次将 GCN 应用于协同过滤算法,通过消息传递显式地对用户物品交互的高维信息进行学习。He 等[5]在此基础上,根据协同过滤的特性删去了 GCN 过程中的特征变换

通信作者:马汉达(mahd@ujs.edu.cn)

和非线性激活部分，提出一种轻量级图卷积神经网络(Light Graph Convolution Network，LightGCN)，其取得了较 NGCF 更高的准确度。Ying 等[6]提出了一种基于随机游走的图卷积网络(Pin Graph with Sample and Aggregate，PinSAGE)，其具有高度的可扩展性，能够适配大规模的网络节点，实现了 GCN 在推荐领域工业级的应用。

现有的方法大多只考虑了图卷积网络的模型架构，却忽略了难负样本在协同过滤算法中起到的作用。现有的协同过滤推荐常常采用正负样本配对形式的损失函数，如贝叶斯个性化排序[7](Bayesian Personalized Ranking，BPR)，这种损失函数通过学习正负样本之间的差异来学习用户的偏好，因此负样本的质量直接影响了最终模型的训练性能，合适的负采样策略可以选择出对模型训练贡献更大的难负样本。基于此，本文提出一种基于动态负采样的图卷积协同过滤推荐模型。贡献如下：

(1)提出一种新的图卷积神经网络来进行协同过滤推荐任务，引入了归一化参数来灵活调节信息传递过程中节点的邻域对其影响的大小，从而使图卷积网络能够灵活适应不同规模的图数据集。

(2)提出一种新的负采样策略，从用户未交互过的物品节点中随机选取负样本集，通过比较负样本集中的样本评分选出难负样本，利用 BPR 损失函数计算正负样本之间的差值作为损失用于模型优化。

(3)在 3 个公开数据集 Gowalla，Yelp2018 和 Amazon-book 上进行了实验，实验结果表明所提出的模型相比基准模型取得了显著的提升，从而验证了所提图卷积网络和负采样策略的有效性。

2 相关工作

2.1 基于传统协同过滤算法推荐模型

协同过滤算法通过物品和用户的交互关系，计算用户或物品的相似度，并基于相似度向用户进行推荐。矩阵分解[8](Matrix Factorization，MF)算法是被广泛应用的协同过滤算法，它将用户和物品交互的稀疏的共现矩阵分解为稠密的用户矩阵和物品矩阵，通过计算用户矩阵和物品矩阵的内积得到用户对物品的评分。深度学习的发展进一步改进了协同过滤算法。Kim 等[9]提出一种基于卷积的神经网络模型，利用卷积神经网络学习用户和物品之间的交互特征，取得了比传统推荐算法更好的推荐效果。He 等[10]提出了神经协同过滤(Neural Collaborative Filtering，NeuMF)，把矩阵分解中用户矩阵和物品矩阵的内积操作换成了多层的神经网络，使得用户向量和物品向量可以得到充分的交叉。Ebesu 等[11]提出一种协同记忆网络(Collaborative Memory Network，CMN)，混合记忆网络和注意力机制来学习用户和物品的邻域信息，从而通过邻域来为用户进行推荐。

上述这些工作都是将物品用户交互信息转换成欧氏空间的嵌入来计算的，然而实际上用户物品的交互信息更近似于图结构，这就导致传统的协同过滤算法无法显式地捕捉交互信息，只能将交互信息作为模型训练的监督信号来隐式地捕捉。

2.2 基于图模型的协同过滤推荐模型

协同过滤算法的输入为用户和物品的交互信息，从图的角度可以看作是用户物品交互的异构二部图，图神经网络对用户物品交互的二部图进行信息传递，利用二部图的高阶联通性，学习到用户或者项目的偏好。基于图模型的协同过滤推荐可以分为 4 个阶段。

(1)图嵌入。图的结构对于图神经网络的信息传递起到至关重要的作用，原始的二部图由物品用户节点及表示它们之间交互关系的边组成。在大规模图中，想要同时计算整张图中所有的节点和交互是不切实际的，因此大规模图计算中广泛应用采样策略来缩减计算量。

(2)邻域聚合。邻域聚合决定了在每一次消息传递的过程中，邻居节点的消息对节点自身的影响。其中最直接的方法是平均聚合[12-13]，如式(1)所示：

$$e_u^{(k)}=\frac{1}{|N_u|}W^{(k)}h_i^{(k)} \tag{1}$$

其中，$e_u^{(k)}$ 表示在第 k 层节点 u 聚合邻居节点的信息，N_u 表示节点 u 的邻居集合，$W^{(k)}$ 表示第 k 层的变换矩阵，$h_i^{(k)}$ 表示第 k 层的输入。

平均聚合在邻居节点的重要性有显著不同的情况下就不再适用。NGCF 等一些工作加入了邻居节点的度来进行归一化，使得邻域聚合的过程可以根据邻居节点的重要性灵活调整。加入归一化后的聚合如式(2)所示：

$$e_u^{(k)}=\sum_{i\in N_u}\frac{1}{\sqrt{|N_u||N_i|}}W^{(k)}h_i^{(k)} \tag{2}$$

其中，N_i 表示物品 i 的邻居集合，i 属于节点 u 的邻居。这种方式仅仅考虑到了邻居节点的重要程度和节点的热门程度成负相关关系，却忽略了不同应用场景下的差异性，导致在一些情况下无法达到较好的效果。Zhang 等[14]提出一种交互式图卷积网络(Inter Active Graph ConvolutionNetwork，IA-GCN)，通过相似性来判断邻居节点的重要程度，从而使图卷积操作能更有针对性地提炼信息。

(3)信息更新。在聚合了邻域的信息后，就要对节点的信息进行更新。NGCF 选择在聚合得到的信息中加入节点自身的信息，再经过非线性激活函数得到新的节点信息。而 LightGCN 通过实验证明自连接和非线性激活对基于图卷积的协同过滤算法的贡献是微乎其微的，因此采取了更简单的信息更新方式，仅保留邻域聚合的信息和归一化的部分。

(4)节点表示。最终表示层需要将各层的嵌入聚合在一起，一般常用的聚合方式有平均聚合[15]、加和聚合[16]、加权聚合[17]以及级联聚合[18-19]。

2.3 负采样策略

在模型训练的过程中，常常会同时给模型提供正样本和负样本，然后通过损失函数增大正负样本之间的差值，来学习正负样本之间的差别，从而学习用户的偏好。负采样是指获取这些与正样本相比较的负样本的策略。比较常见的负采样方式是进行随机负采样[20-21]，Cai 等[22]提出约 95%的负样本是易负样本，它们与正样本并不相似，仅用易负样本不足以训练出一个好的模型。其次，约 5%的负例是难负样本，它们与正样本相似但有区别，这些难负样本几乎决定了模型的结果，在训练中发挥了关键作用。

为了提高负采样的质量，Zhang 等[23]提出了一种应用在协同过滤推荐任务中的动态负采样策略，根据当前生成的排名列表动态选择负样本。Yang 等[24]提出了一种用于图表示学习的负采样策略，提出负样本的分布应与正样本正相关，通过

自对比估计逼近正采样分布，并应用马尔可夫链加速计算。Wang 等[25]提出一种基于知识图谱的负采样策略，利用强化学习探索知识图谱上的辅助信息，从而获取高质量的负样本。

总结上述工作中的负采样策略，可以看出在负采样过程中最重要的是如何选定合适的负样本。本文基于图卷积神经网络的特点，提出一种应用在图卷积网络上的动态负采样策略，通过图卷积计算负样本评分，以选择高质量的难负样本，提高推荐性能。

3 模型描述

本节对提出的基于动态负采样的图卷积协同过滤推荐模型及其模型训练方式做出了详细的介绍。首先是对本模型所用图卷积神经网络的介绍，如图 1(a)所示，图卷积神经网络主要分为图嵌入、消息传递和节点表示 3 个模块，然后是对本模型的负采样策略的描述，如图 1(b)所示，最后详细说明了本文在训练过程中是如何根据模型的变化动态地进行负采样的。

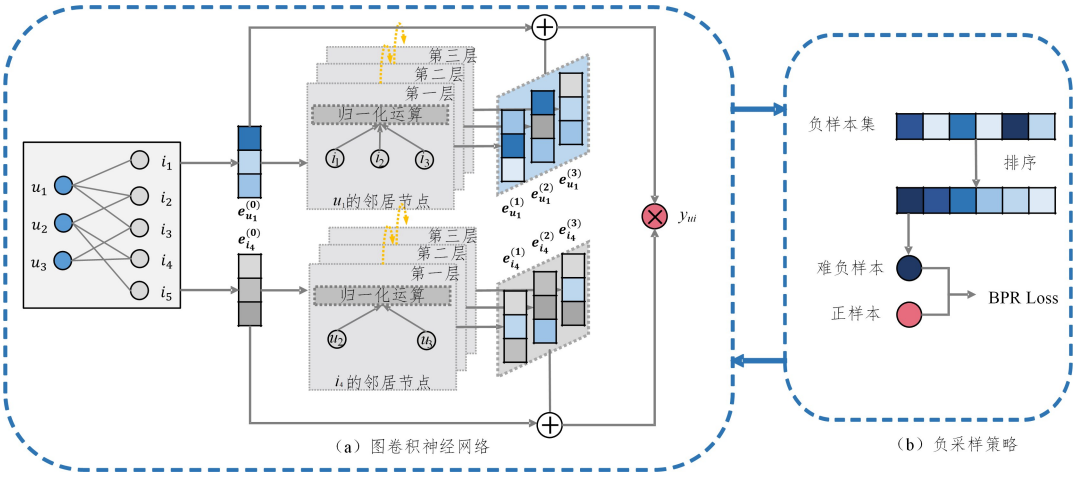


图 1 基于动态负采样的图卷积协同过滤推荐模型

Fig. 1 Dynamic negative sampling for graph convolution network based collaborative filtering recommendation model

3.1 图卷积神经网络

常见的图卷积神经网络可以分为图嵌入、消息传递和节点表示 3 个部分。输入的用户物品交互二部图首先通过图嵌入的处理变为嵌入的形式，然后将生成的嵌入输入到网络中，经过 K 层的邻域聚合和信息更新，就获得了 K 个经过消息传递后的层嵌入，将这些层嵌入通过表示层进行层聚合就可以得到节点的最终嵌入，再通过内积等方式将节点进行加权运算，就可以得到最终的评分。本模型也遵循上述的图卷积过程，下面将分 3 部分详细介绍本模型的图卷积神经网络的细节和原理。

(1)图嵌入

参考 NGCF 中的图嵌入方式，将一个用户 u 或物品 i 的第 k 层的嵌入表示为$e_u^{(k)} \in R^d$或$e_i^{(k)} \in R^d$，其中 R 表示用户和物品之间的交互矩阵，d 表示嵌入的维度。用户 u 和物品 i 的嵌入构成嵌入向量表$E^{(k)}$，如式(3)所示：

$$E^{(k)} = [e_{u_1}^{(k)}, \cdots, e_{u_N}^{(k)}, e_{i_1}^{(k)}, \cdots, e_{i_M}^{(k)}] \tag{3}$$

其中，$E^{(k)}$ 表示第 k 层的嵌入向量表，在图卷积网络中，嵌入表以端到端的方式进行优化。

(2)消息传递

在消息传递过程中，每个节点需要先进行邻域聚合，获取邻域的信息，然后再进行信息更新，将邻域信息和自身节点信息经过处理作为新的节点信息。在邻域聚合的部分，本算法考虑到不同规模的图数据中邻域对节点自身信息的影响有着很大的区别，对邻域聚合过程进行归一化处理，并且在归一化中引入了归一化参数来控制邻域大小对聚合过程的影响。在信息更新的过程中，本算法参考了 LightGCN 对节点更新进行的简化处理，删除了特征变换和非线性激活函数，从而使图卷积网络更加适合协同过滤推荐任务。本图卷积网络的信息传递过程如式(4)、式(5)所示：

$$e_u^{(k+1)} = \sum_{i \in N_u} N_u^{-\frac{1}{2}} N_i^{-\frac{1}{2}\alpha} e_i^{(k)} \tag{4}$$

$$e_i^{(k+1)} = \sum_{u \in N_i} N_i^{-\frac{1}{2}} N_u^{-\frac{1}{2}\alpha} e_u^{(k)} \tag{5}$$

其中，$e_i^{(k)}$ 和$e_u^{(k)}$ 分别表示用户 i 和物品 u 经过第 k 层信息传递后的嵌入，N_u 和N_i 分别表示用户 u 和物品 i 的邻居集合。$N_u^{-\frac{1}{2}} N_i^{-\frac{1}{2}\alpha}$为归一化项，用来避免由于卷积的原因导致嵌入的规模增大，其中 α 为超参数，通过调节 α 可以灵活地调整图卷积中自身节点和邻居节点的影响占比，从而使算法更加适应不同规模的图数据。

(3)节点表示

经过 K 层的信息传递后，可以得到各层的嵌入，再将各层的嵌入和初始的嵌入通过加权聚合进行表示，可以得到用户 u 和物品 i 的最终嵌入e_u和e_i，如式(6)、式(7)所示：

$$e_u = \sum_{k=0}^{L} \beta_k e_u^{(k)} \tag{6}$$

$$e_i = \sum_{k=0}^{L} \beta_k e_i^{(k)} \tag{7}$$

其中，β_k作为一个超参数，用来调节层聚合时各层的权重。关于β_k的取值，尝试使用多层感知机计算各层的权重，从而进行加权聚合，但结果并没有提升，最后参考文献[3]将β_k固定为$\frac{1}{1+K}$，即平均聚合。计算得到最后输出的评分$\hat{y}_{ui}$如式(8)所示：

$$\hat{y}_{ui} = e_u^{\mathrm{T}} e_i \tag{8}$$

其中，e_u表示用户 u 的最终嵌入，e_i表示物品 i 的最终嵌入。

3.2 负采样策略

通过图卷积操作可以得到正样本最终的评分$\hat{y}_{ui}$，同理也可以对负样本进行图卷积操作，从而得到负样本的最终评分。本模型采用动态负采样策略，通过负样本评分找出负样本中

的难负样本，即与正样本相似度高但有区别的负样本。在协同过滤推荐任务中，用户评分高的物品会被认为是可能与用户产生交互的物品推荐给用户，在选取难负样本的时候也可以利用相同的思想。评分高的负样本可以被认为是接近正样本的负样本，也就是所谓的难负样本。因此通过随机选取与目标节点 u 之间没有交互的 n 个节点 j 产生负样本集 S_j，对 S_j 的样本进行图卷积，计算负样本集的评分并进行排序，选取最高的评分作为难负样本的评分 $\hat{y}_{uj}$，如式(9)所示：

$$\hat{y}_{uj}=\arg\max_{j_i\in S_j} e_u^{\mathrm{T}}e_{j_i} \tag{9}$$

其中，e_u 表示用户 u 的最终嵌入，e_{j_i} 表示负样本集中第 i 个负样本 j_i 的最终嵌入。这种选取负样本的方式可能会将用户潜在的兴趣物品作为负样本，从而导致推荐性能的下降，但只要选取合适的负样本集大小，就可以减少这种伪负样本对性能的影响。在下一节的超参数设置中会具体讨论负采样取值与模型推荐性能之间的关系。

为了对难负样本和正样本之间的偏好差别进行有效的计算，本文采用了贝叶斯个性化排序对模型进行优化，其基本思路是对样本进行两两比较，构建偏序对，从比较中学习排序。其公式如式(10)所示：

$$L_{bpr}=-\sum_{u=1}^{M}\sum_{i\in N}\sum_{j\notin N}\ln\sigma(\hat{y}_{ui}-\hat{y}_{uj})+\lambda\|E^{(0)}\|^2 \tag{10}$$

其中，λ 控制 L2 正则化的强度，$E^{(0)}$ 是第 0 层的输入 $e_u^{(0)}$ 和 $e_i^{(0)}$ 组成的嵌入表，N 表示用户的邻域，M 是用户数量，u 表示用户节点，i 表示物品节点，j 表示负样本节点，$\hat{y}_{ui}$ 为正样本评分，$\hat{y}_{uj}$ 为负样本评分，σ 表示 sigmoid 激活函数。

3.3 模型训练

动态负采样指的是随着模型训练而变化的负采样策略。DGCCF 的训练过程如算法 1 所示，在每个 epoch 的训练过程中，模型首先先对损失函数初始化，然后对每一个用户 u 随机选择其未交互过的 n 个节点 j 作为负样本集 S_j，用户集、物品集和负样本集都会被送入图卷积网络，经过 K 层神经网络得到其 K 层的层嵌入，并通过层聚合得到用户集、物品集和负样本集的最终嵌入，将用户集分别与物品集和负样本集进行内积运算，得到正样本评分和负样本集评分。接着通过上节叙述的负采样策略从负样本集中选取评分最高的负样本作为难负样本，保留难负样本的评分，与正样本评分一同通过 BPR 损失函数计算出损失，再通过 Adam 优化器来对梯度下降算法进行优化，更新模型的参数。

算法 1 基于动态负采样的图卷积协同过滤推荐模型训练过程

输入：数据集，嵌入维度 d，学习率 lr，图卷积层数 K，L_2 正则化强度 λ，负采样集大小 n，归一化参数 α

输出：预测物品集 R

1. 初始化神经网络中的各项参数。
2. 划分训练集和测试集。
3. FOR e ←1 to epoch do
4. 初始化损失函数 $L_{bpr}=0$。
5. 根据正样本 u 选取负样本集 S_j。
6. FOR k ←1 to K do
7. 通过式(4)、式(5)获取用户集、物品集和负样本集的各层嵌入。
8. END FOR
9. 通过式(6)、式(7)进行层聚合，得到物品集、用户集、负样本集的最终嵌入。
10. 计算出正样本的评分 $\hat{y}_{ui}$ 和负样本集的评分集。
11. 在负样本集中选出难负样本，得到难负样本评分 $\hat{y}_{uj}$。
12. 更新损失函数 L_{bpr}，通过 Adam 优化器优化梯度下降算法对参数进行更新。
13. END FOR

4 实验验证

本节首先介绍了实验所用的数据集、评价指标、实验环境等相关实验设置，然后给出本模型在 3 个数据集上的实验结果，并与其他 6 个模型进行推荐性能的对比，接着针对本模型进行消融实验以验证各个模块的有效性，最后给出了模型相关的超参数的实验结果。

4.1 实验设置

4.1.1 数据集

本实验采用 3 个公开的真实数据集 Gowalla，Yelp2018 和 Amazon-book。Gowalla 是用户签到数据集，Yelp2018 是 yelp 网站公开的餐厅评价数据，Amazon-book 是亚马逊购物评论数据集中的书籍子数据集。为了保证数据集的质量，对这 3 个数据集进行了预处理操作，仅保留拥有至少 10 次交互记录的节点。3 个数据集的统计情况如表 1 所列，可以看出 3 个数据集的规模各不相同。

表 1 数据集情况

Table 1 Dataset statistics

数据集	# User	# Item	# Interaction
Gowalla	29858	40981	1027370
Yelp2018	31668	38048	1561406
Amazon-Book	52643	91599	2984108

注：数据集均来自 https://github.com/gusye1234/LightGCN-PyTorch。

4.1.2 评价指标

本实验选择了两个评价指标 Recall@Q 和 NDCG@Q。Recall@Q 即召回率，衡量的是 TopQ 推荐中，推荐列表中用户有过交互行为的商品数量占测试集中用户所有有过交互行为的商品数量的比例。NDCG@Q 即归一化折损累计增益，衡量了 TopQ 推荐中，推荐列表中不同位置推荐结果的相关性得分，与用户相关性越高的推荐商品排序越靠前，其推荐效果越好得分越高。计算方式如式(11)、式(12)所示。

$$Recall@Q=\frac{1}{U}\sum_{u=1}^{U}\frac{|L_u\cap L_u^{\text{test}}|}{|L_u^{\text{test}}|} \tag{11}$$

$$NDCG@Q=\frac{1}{U}\sum_{u=1}^{U}\frac{\sum_{i=1}^{K}\frac{2^{r_u(i)}-1}{\log_2(i+1)}}{\sum_{i=1}^{L_u^{\text{test}}}\frac{1}{\log_2(i+1)}} \tag{12}$$

其中，U 是测试集中的用户数量，L 是给用户推荐的前 Q 个物品的列表，L_u^{test} 是用户有过交互的前 Q 个物品的列表，$r_u(i)$ 表示推荐列表上第 i 个位置上的物品是否是用户 u 交互过的物品。在本实验中，参考文献[4]和文献[5]取 Q 的值为 20。

4.1.3 对比模型

为了验证本模型的有效性，分别从传统协同过滤模型、基于深度学习的推荐模型和基于图神经网络的推荐模型中挑选了公认表现较好的模型作为基准模型。

(1)MF[8]：传统推荐模型中的经典算法，其基本思想是将协同过滤中的用户物品交互矩阵分解为用户的隐向量矩阵和物品的隐向量矩阵两个稠密矩阵，通过这两个

矩阵来预测用户对物品的评分。

(2)CMN[11]:是一种将基于潜在因素模型的全局结构和基于邻域的局部结构用非线性方法进行统一实现的深度学习模型。

(3)PinSage[6]:是 Pinterest 公司基于 GraphSAGE 实现的召回算法,将图卷积经典算法 GraphSAGE 应用于推荐领域,并实现了图卷积在大规模网络上的落地应用。

(4)NGCF[4]:通过构建用户物品交互二部图,将图卷积神经网络应用于协同过滤推荐任务中,以捕捉用户和物品之间的高维交互关系。

(5)LightGCN[5]:对 NGCF 进行消融实验,删去了其中的特征变换和非线性激活部分,使网络结构更加轻量级并且更加适合协同过滤推荐任务,取得了更好的推荐效果。

(6)IA-GCN[14]:通过学习邻居节点与目标节点之间的相似性,为与目标节点更相似的邻居节点赋予更高的注意力权重,从而使得图卷积操作能够更有效地提炼出与目标节点有关的信息。

4.1.4 实验环境及参数

本实验的实验环境为 Python3.7,Pytorch1.11.0,NVIDIA GeForce RTX 3060,随机将数据集中的 80%划分为训练集,剩下的 20%为训练集。参数设置如表 2 所列,其中 d,lr,K,λ,Q,n,α 分别为嵌入维度、学习率、图卷积层数、L_2 正则化强度、物品列表长度、负采样集大小、归一化参数。

表 2 参数设置

Table 2 Parameters setting

数据集	d	lr	K	λ	Q	n	α
Gowalla	64	0.0005	3	0.0001	20	48	0.7
Yelp2018	64	0.0001	3	0.0001	20	48	0.4
Amazon-book	64	0.0001	3	0.0001	20	48	0.1

4.2 实验结果

4.2.1 模型推荐效果

本文在 3 个不同的数据集上对包括本文提出的模型在内的 7 个模型进行了对比实验,实验结果如表 3 所列,其中下划线表示最佳基线结果,粗体字表示最优结果。

表 3 本模型与对比模型在 3 个数据集上的 Recall@20 和 NDCG@20 结果

Table 3 Recall@20 and NDCG@20 of DGCCF and comparison models on 3 data sets

数据集	Gowalla		Yelp2018		Amazon-book	
方法	*Recall*@20	*NDCG*@20	*Recall*@20	*NDCG*@20	*Recall*@20	*NDCG*@20
MF	0.1291	0.1109	0.0433	0.0354	0.0250	0.0196
CMN	0.1405	0.1221	0.0457	0.0369	0.0267	0.0218
PinSage	0.1380	0.1196	0.0471	0.0393	0.0282	0.0219
NGCF	0.1570	0.1327	0.0579	0.0477	0.0344	0.0263
LightGCN	0.1762	0.1505	0.0643	0.0524	0.0408	0.0318
IA-GCN	0.1839	**0.1562**	0.0659	0.0537	0.0472	0.0373
本模型	**0.1845**	0.1557	**0.0721**	**0.0592**	**0.0522**	**0.0408**
提升/%	0.3	−0.3	9.4	10.2	10.6	9.4

通过分析实验结果可以得到如下结论:

(1)本模型在所有数据集的所有评价指标下均取得了最优性能。以最佳基线在 Recall@20 上的结果为准,本模型分别在 3 个数据集上取得了 0.3%,9.4%,10.6%的提升,足以证明本模型的有效性。

(2)CMN 的结果优于 MF。可以看出,基于深度学习实现的推荐模型可以比传统推荐模型挖掘出更多的隐含信息,从而获得更好的推荐结果。

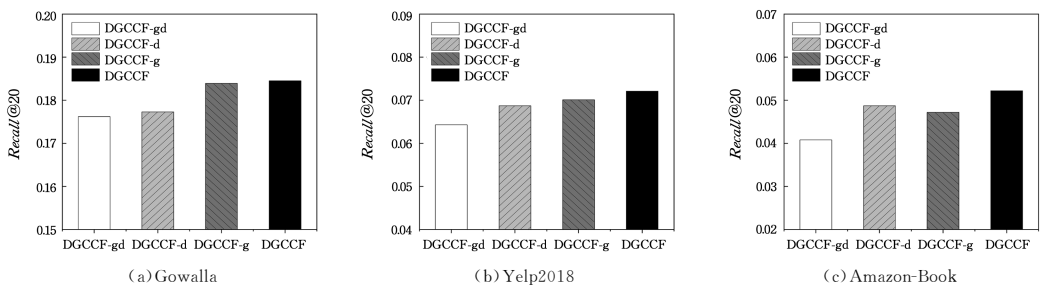
(3)虽然同为图卷积模型,但 PinSage 比 CMN 结果更差,可能是由于 PinSage 采用的图结构并不适用于协同过滤推荐任务,而 NGCF 和 LightGCN 适应协同过滤推荐任务的要求改进了图卷积模型,取得了明显优于深度学习模型的结果。

(4)可以看出,随着数据集交互规模的增加,本模型取得的提升越来越显著,这体现了本模型能更加有效地应对交互数多的大数据集。

(5)对比最新的图卷积模型 IA-GCN,本模型在 Gowalla 数据集上的提升不大,但是在 Yelp2018 和 Amazon-book 上取得了较大提升,佐证了本模型对于大数据集的效果,以及本模型在小数据集上仍有提升空间。

4.2.2 消融分析

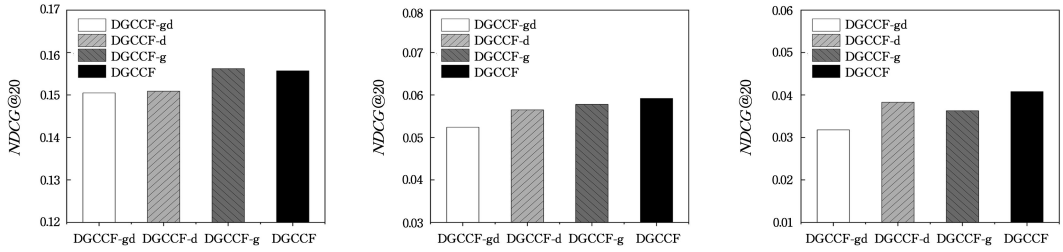
为了进一步验证本模型中图卷积模块和负采样模块的有效性,本文设计了 3 种不同情况的消融实验,其中 DGCCF-g 表示将 DGCCF 中的归一化参数去除后的实验结果,DGCCF-d 表示将 DGCCF 中的负采样策略更换为随机负采样后的实验结果,DGCCF-gd 表示将 DGCCF 中的归一化参数去除并设置负采样策略为随机负采样后的实验结果。在 3 个数据集上进行的消融实验结果如图 2 和图 3 所示。



(a)Gowalla (b)Yelp2018 (c)Amazon-Book

图 2 在 3 个数据集上 DGCCF 及其变体的 *Recall*@20 实验结果

Fig. 2 *Recall*@20 of DGCCF and its variants on three datasets



(a) Gowalla (b) Yelp2018 (c) Amazon-Book

图 3 在 3 个数据集上 DGCCF 及其变体的 *NDCG*@20 实验结果

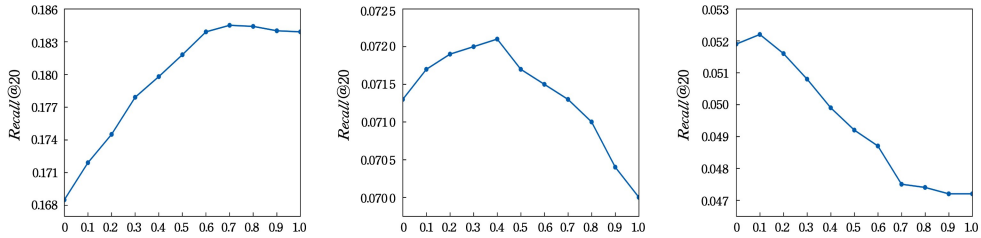
Fig. 3 *NDCG*@20 of DGCCF and its variants on three datasets

从实验结果可以得出，本模型的图卷积模块和负采样模块均对推荐准确度的提升起到了贡献作用。在 Gowalla 数据集中，起到主要作用的是动态负采样模块，归一化参数起到的作用比较小；在 Yelp2018 中，动态负采样起到的作用略高于归一化参数；在 Amazon-Book 数据集上，归一化参数起到的作用略高于动态负采样。通过二者之间贡献占比的变化，可以得出随着网络规模的增大，归一化参数的重要性越来越凸显，而在小规模网络中，合适的负采样策略可以对模型性能提升起到更大的作用。

4.3 超参数设置

4.3.1 归一化参数

归一化参数 α 可以调节邻域聚合过程中邻居节点和节点自身对聚合结果影响的占比。实验从 α 为 0 开始每次增加 0.1，直到 α 为 1 停止训练，记录每次训练后的结果，选取其中最优值作为参数的取值保存下来，实验结果如图 4 所示。由于 Recall@20 和 NDCG@20 之间有相似的变化趋势，且 Recall@20 的波动较明显，本实验仅展示 Recall@20 的实验结果。



(a) Gowalla (b) Yelp2018 (c) Amazon-Book

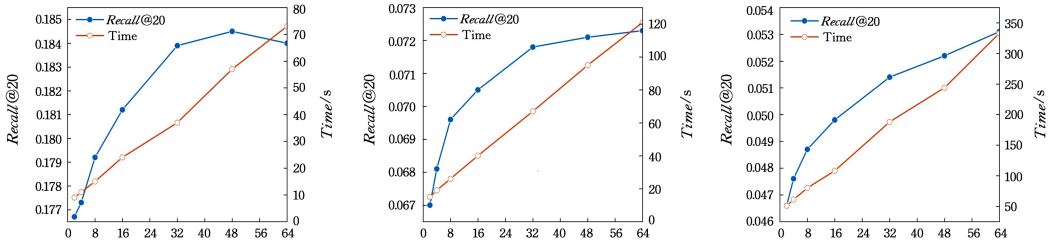
图 4 在 3 个数据集上不同 α 取值对应的 *Recall*@20 实验结果

Fig. 4 *Recall*@20 corresponding to different α values on three data sets

从图 4 中可以看出，Recall@20 随着归一化参数 α 的变化而变化，整体的变化趋势为抛物线。根据实验结果可以确定 α 在 3 个数据集上的取值分别为 0.7，0.4，0.1，通过实验可以看出，随着数据集规模的扩大，α 的取值逐渐减小，即在越大规模的网络中，邻居节点的权重越小，其预测结果越好。

4.3.2 负采样集大小

负采样集大小 n 决定了在动态负采样过程中，选择多大规模的负样本集作为候选。需要注意的是，随着负样本集的扩大，模型的训练时间也会增加，所以选择负样本集大小 n 的时候，需要考虑到其对性能的提升和训练时长的增加两方面因素。对超参数 n 分别取值{2，4，8，16，32，48，64}进行实验，选取 10 个 epoch 的耗时进行平均之后得到实验的平均耗时。实验结果如图 5 所示，左轴表示实验结果 Recall@20，右轴表示实验平均耗时(单位为 s)。



(a) Gowalla (b) Yelp2018 (c) Amazon-Book

图 5 在 3 个数据集上不同 n 取值对应的 *Recall*@20 实验结果和实验平均耗时

Fig. 5 *Recall*@20 and average time-consuming corresponding to different n values on three data sets

从图 5 中可以看出，实验的平均耗时随着 n 的增大而增大，且近似于线性相关；但 Recall@20 近似于抛物线，且随着 n 的增大其增速逐渐放缓，在 Gowalla 数据集上甚至出现了拐点，这可能是因为随着负样本集的增加，采样到伪负样本即用户潜在感兴趣物品的概率增加，导致了结果的下降，并且图规模越小，采样到伪负样本的概率就越大，拐点出现得就越早。综合实验结果和实验平均耗时来考虑，选定 48 作为本实验 n 的取值。特别指出，此处 n 的取值不是定值，随着数据集或者实验设备的变化，n 的取值也应相应改变。

结束语 本文提出了一种基于动态负采样的图卷积协同过滤推荐模型，通过图卷积网络来捕捉用户和物品之间的高维交互信息，以完成协同过滤任务。首先，在图卷积过程中，

引入了归一化参数,用来调节信息传递过程中邻域对节点的影响程度,从而使图卷积网络更加适应不同规模的图数据;其次,提出一种动态负采样策略,根据图卷积网络的特点来选择负样本,具体是从用户未交互过的物品节点中选取负样本集,通过图卷积得到负样本集的评分,选取评分最高的负样本作为难负样本;最后,将难负样本和正样本作为样本对加入训练中来提升推荐模型的准确度。经过实验证明,本模型的推荐性能优于基准模型,取得了较高的准确度。

本模型是基于协同过滤的模型,且对于大规模的图数据集效果更佳,所以存在对用户交互数据依赖度较高的问题,更适用于已经成熟运作具有较多历史交互数据的推荐系统中。下一步工作中,可以通过添加辅助信息的手段,如结合知识图谱添加用户和物品的内容信息,或者添加用户社交信息等,形成混合推荐模型,以缓解冷启动问题,从而扩大模型的应用范围。

参 考 文 献

[1] MA H D,JING D. A microblog friend recomm endation algorithm-based on SSD and timing model[J]. Computer Engineering & Science,2021,43(7):1291-1298.

[2] CHENG Z T,ZHONG T,ZHANG S M,et al. Survey of Recommender Systems Based on Graph Learning[J]. Computer Science,2022,49(9):1-13.

[3] WU S,SUN F,ZHANG W,et al. Graph neural networks in recommender systems: a survey[J]. ACM Computing Surveys,2022,55(5):1-37.

[4] WANG X,HE X,WANG M,et al. Neural graph collaborative filtering[C]// Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2019:165-174.

[5] HE X,DENG K,WANG X,et al. Lightgcn:Simplifying and powering graph convolution network for recommendation[C]// Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020:639-648.

[6] YING R,HE R,CHEN K,et al. Graph convolutional neural networks for web-scale recommender systems[C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018:974-983.

[7] RENDLE S,FREUDENTHALER C,GANTNER Z,et al. Bayesian personalized ranking from implicit feedback[C]// Proc. of Uncertainty in Artificial Intelligence. 2014:452-461.

[8] KOREN Y,BELL R,VOLINSKY C. Matrix factorization techniques for recommender systems[J]. Computer,2009,42(8):30-37.

[9] KIM D,PARK C,OH J,et al. Convolutional matrix factorization for document context-aware recommendation[C]// Proceedings of the 10th ACM Conference on Recommender Systems. 2016:233-240.

[10] HE X,LIAO L,ZHANG H,et al. Neural collaborative filtering[C]// Proceedings of the 26th International Conference on World Wide Web. 2017:173-182.

[11] EBESU T,SHEN B,FANG Y. Collaborative memory network for recommendation systems[C]// The 41st international ACM SIGIR Conference on Research & Development in Information Retrieval. 2018:515-524.

[12] SUN J,ZHANG Y,MA C,et al. Multi-graph convolution collaborative filtering[C]// 2019 IEEE International Conference on Data Mining(ICDM). IEEE,2019:1306-1311.

[13] TAN Q,LIU N,ZHAO X,et al. Learning to hash with graph neural networks for recommender systems[C]// Proceedings of The Web Conference 2020. 2020:1988-1998.

[14] ZHANG Y,WANG P,ZHAO X,et al. IA-GCN:Interactive Graph Convolutional Network for Recommendation[J]. arXiv:2204.03827,2022.

[15] GONG K,SONG X,WANG S,et al. ITSM-GCN:Informative Training Sample Mining for Graph Convolutional Network-based Collaborative Filtering[C]// Proceedings of the 31st ACM International Conference on Information & Knowledge Management. 2022:614-623.

[16] WANG X,JIN H,ZHANG A,et al. Disentangled graph collaborative filtering[C]// Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020:1001-1010.

[17] WU J,HE X,WANG X,et al. Graph convolution machine for context-aware recommender system[J]. Frontiers of Computer Science,2022,16(6):1-12.

[18] JIN B,GAO C,HE X,et al. Multi-behavior recommendation with graph convolutional networks[C]// Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020:659-668.

[19] CHEN L,WU L,HONG R,et al. Revisiting graph based collaborative filtering:A linear residual graph convolutional network approach[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020:27-34.

[20] DIAZ-AVILES E,DRUMOND L,SCHMIDT-THIEME L,et al. Real-time top-n recommendation in social streams[C]// Proceedings of the Sixth ACM Conference on Recommender Systems. 2012:59-66.

[21] CUI P,LIU S,ZHU W. General knowledge embedded image representation learning[J]. IEEE Transactions on Multimedia,2017,20(1):198-207.

[22] CAI T T,FRANKLE J,SCHWAB D J,et al. Are all negatives created equal in contrastive instance discrimination? [J]. arXiv:2010.06682,2020.

[23] ZHANG W,CHEN T,WANG J,et al. Optimizing top-n collaborative filtering via dynamic negative item sampling[C]// Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2013:785-788.

[24] YANG Z,DING M,ZHOU C,et al. Understanding negative sampling in graph representation learning[C]// Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020:1666-1676.

[25] WANG Y,LIU Z,FAN Z,et al. Dskreg:Differentiable sampling on knowledge graph for recommendation with relational gnn[C]// Proceedings of the 30th ACM International Conference on Information & Knowledge Management. 2021:3513-3517.

MA Handa, born in 1966, master, professor, is a member of China Computer Federation. His main research interests include data mining, and big data procesing technology &its application.