

【电子与信息科学 / Electronics and Information Science】

基于分离嵌入交叉网络的推荐模型

封书蕾¹, 蒋中云²

1) 上海海洋大学信息学院, 上海201306; 2) 上海建桥学院信息技术学院, 上海201306

摘要: 针对现有深度学习推荐模型中的特征交叉方法存在无法充分利用嵌入向量信息与预测的精度不够的问题, 提出一种基于分离嵌入交叉网络(separated embedding interaction networks, SEIN)的深度学习推荐模型. 该模型先采用嵌入神经网络层将稀疏的特征向量转化为稠密的嵌入向量, 再对不同维度的特征矩阵分离进行特征交叉, 并且通过分离嵌入交叉网络层数显式控制特征交叉的阶数, 最终将得到的各隐藏层矩阵求和池化, 并通过预测层得到最后的输出. 在Criteo、AutoML与Movielens公开数据集上, 以推荐结果的曲线下面积、对数损失、准确率与召回率作为评估指标进行了点击率预测与top- k 推荐实验. 结果表明, 与点击率预测基线模型DeepFM、Deep&Cross与xDeepFM对比, SEIN模型在Criteo数据集上的曲线下面积分别提升了2.38%、2.31%与2.35%, 对数损失分别下降了1.81%、1.99%与1.85%; 在AutoML数据集上曲线下面积分别提高了1.17%、2.60%与0.57%, 对数损失分别下降了0.66%、2.53%与0.35%. 与基于图神经网络的推荐模型HeteGraph、IR-Rec、GSIRec、KGNCf-RNN和ITRA相比, 基于SEIN的推荐模型的准确率($k=5$)分别提升1.27%、0.47%、0.48%、0.56%与2.59%. 基于SEIN的推荐模型能够有效解决无法充分利用嵌入向量信息的问题, 提高推荐准确度.

关键词: 人工智能; 推荐技术; 深度学习; 特征交叉; 向量稠密化; 数据挖掘; 点击率预测

中图分类号: TP181; TP183; TP391 **文献标志码:** A **DOI:** 10.3724/SP.J.1249.2023.04513

Recommendation model based on separated embedding interaction networks

FENG Shulei¹ and JIANG Zhongyun²

1) School of Information, Shanghai Ocean University, Shanghai 201306, P. R. China

2) School of Information Technology, Shanghai Jian Qiao University, Shanghai 201306, P. R. China

Abstract: Aiming at the problems that the existing feature interaction methods in deep learning recommendation models cannot fully utilize the embedding vector information and thus have the insufficient accuracy, we propose a deep learning recommendation model based on separated embedding interaction networks (SEIN). This model first uses the embedding neural network layer to convert the sparse feature vectors into dense embedding vectors, then separates the feature matrices of different dimensions for feature interaction, and explicitly controls the order of feature interaction through the number of SEIN layers. Finally, the obtained hidden layer matrices are pooled by summation, and the final output is obtained through the prediction layer. In public datasets of Criteo, AutoML and Movielens, click-through rate prediction and top- k recommendation experiments are carried out by using the area under the curve (AUC), log-loss, accuracy and recall rate as evaluation indicators. The experimental results demonstrate that compared with the baseline models for click-through rate prediction, namely DeepFM, Deep&Cross

Received: 2022-11-25; **Accepted:** 2023-04-02; **Online (CNKI):** 2023-06-05

Foundation: Construction Project of Applied Undergraduate Pilot Program in Shanghai Universities (Z32004-17-84); Special Project of Shanghai Municipal Education Commission for the Construction of First-Class Undergraduate Majors (JYLB202002)

Corresponding author: Associate professor JIANG Zhongyun. E-mail: jianqiao_jzy@163.com

Citation: FENG Shulei, JIANG Zhongyun. Recommendation model based on separated embedding interaction networks [J]. Journal of Shenzhen University Science and Engineering, 2023, 40(4): 513-520. (in Chinese)



and xDeepFM, the SEIN model improves the AUC by 2.38%, 2.31% and 2.35% on the Criteo dataset, and reduces the log-loss by 1.81%, 1.99% and 1.85%, respectively. On the AutoML dataset, the SEIN model increases the AUC by 1.17%, 2.60% and 0.57%, and reduces the logloss by 0.66%, 2.53% and 0.35%, respectively. Compared with the recommendation models based on graph neural networks, specifically HeteGraph, IR-Rec, GSIRec, KGNCf-RNN, and ITRA, the recommendation model based on SEIN improves the accuracy ($k = 5$) by 1.27%, 0.47%, 0.48%, 0.56% and 2.59%, respectively. Therefore, the SEIN model can effectively solve the problem of not fully utilizing the embedding vector information and thus improve the accuracy of recommendation.

Key words: artificial intelligence; recommendation technology; deep learning; feature interaction; vector densification; data mining; click through rate prediction

随着大数据时代的到来,数据信息量正呈爆炸式增长,引发了信息过载问题,这令用户难以从海量信息中找到自己感兴趣的内容.推荐系统作为解决信息过载问题的首要手段,已被广泛用于视频、电影和广告等领域.从用户角度来看,推荐系统能令用户在海量信息中高效获得感兴趣的信息;而从企业角度来看,推荐系统能最大限度地吸引和留存用户、增加用户黏性、提高用户转化率,促进企业商业目标连续增长,从而实现信息消费者与生产者的双赢.

传统的推荐方法主要分为基于协同过滤^[1-2]、基于内容^[3]和混合推荐方法^[4].近年来,深度学习技术被广泛用于推荐系统^[5-14],其本质是对大规模数据特征进行更深层次的挖掘,学习有效的特征表示及映射以优化推荐质量.

推荐系统模型的核心是特征之间的交叉,因子分解机(factorization machines, FM)模型^[15]的提出,首次考虑了特征之间的二阶交叉,弥补了逻辑回归表达能力差的缺陷,也解决了在稀疏数据场景下模型参数难以训练的问题.自编码器推荐(auto-encoders recommender, AutoRec)模型^[16]将自编码器的思想和协同过滤相结合,是一种单隐藏层神经网络推荐模型. AutoRec模型的提出,拉开了使用深度学习解决推荐问题的序幕,为复杂深度学习网络的构建提供了思路. SHAN等^[5]提出的Deep crossing模型是深度学习架构在推荐系统的一次完整应用,解决了从特征工程、稀疏向量稠密化和多层神经网络进行优化目标拟合等一系列应用问题,可通过调整神经网络的深度进行特征之间的“深度交叉”. QU等^[6]提出基于向量点积的神经网络(product-based neural networks, PNN)模型,用乘积层(product layer)代替了Deep crossing模型中的堆叠(stack)层,令不同特征的嵌入向量不再是简单的拼接,而是用product操作进行两两交互,更有

针对性地获取特征之间的交叉信息. CHENG等^[7]提出的Wide&Deep模型是由单层的wide部分和多层的deep部分组成的混合模型,wide部分使模型具有较强的记忆能力,deep部分让模型具有泛化能力,因此该模型能够快速处理并记忆大量历史行为特征,且表达能力强. HE等^[8]提出神经协同过滤(neural collaborative filtering, NeuralCF)模型,采用多层神经网络+输出层的结构代替矩阵分解模型中简单的内积操作,让用户向量和物品向量充分交叉,得到更多有价值的特征组合信息,并引入更多的非线性特征让模型的表达能力更强. GUO等^[9]提出深度因子分解机(deep factorization machines, DeepFM)模型,用FM替代Wide&Deep模型的wide部分,加强了浅层网络部分特征的组合能力. HE等^[10]提出神经因子分解机(neural factorization machines, NFM),将FM的二阶交叉特征作为多层感知机的输入,结合了FM的线性与多层感知机的非线性. XIAO等^[11]提出注意力因子分解机(attentional factorization machines, AFM)模型,引入了注意力机制,并假设不同的交叉特征对于结果的影响程度不同. LIAN等^[12]提出极深因子分解机(extreme deep factorization machine, xDeepFM)模型,在Wide&Deep的基础上添加了压缩交互网络(compressed interaction network, CIN)显式的构造有限阶特征组合. 阿里巴巴提出深度兴趣网络(deep interest network, DIN)^[13]和深度兴趣演化网络(deep interest evolution network, DIEN)^[14]模型,前者侧重点是在利用候选商品和历史行为商品之间的相关性生成注意力评分,后者则是通过门循环单元(gated recurrent unit, GRU)来捕捉用户行为序列中时序关系,并且结合注意力机制,利用注意力评分来修正更新门,控制隐藏层的更新,在演化过程中弱化不相关的兴趣.

2018年后,推荐系统的研究方向大多偏向于利

用辅助信息, 如基于知识图谱(knowledge graph, KG)的推荐系统和基于图神经网络的推荐系统等. KG是大量实体与实体之间关系的集合, 且可通过结构化数据形式将各种信息进行关联. 引入KG可更有效地利用辅助信息, 有效缓解传统推荐系统中存在的冷启动等问题. 图神经网络借鉴了卷积、循环网络和自编码器的思想, 利用图来反映对象之间的复杂关系, 在学习图形结构数据方面表现出色, 近年来被广泛用于推荐系统^[17-23]. HUANG等^[17]提出了知识增强的序列推荐(knowledge-enhanced sequential recommender, KSR)模型, 利用GRU删除无效信息, 保留重要信息, 并得到隐状态向量, 通过翻译嵌入(translating embeddings, TransE)模型^[24]对实体与实体之间的关系进行学习表征, 再由Key-Value记忆网络模块^[25]得到用户的偏好向量. SHA等^[18]提出基于分层注意力知识图谱嵌入(hierarchical attentive knowledge graph embedding, HAKG)模型, 利用KG中用户-物品组的子图表征它们之间的连通性, 子图包含了KG丰富的语义和拓扑结构, 且只保留与特定用户-物品连通性相关的实体和关系, 避免了引入来自整个KG的噪声. TRAN等^[19]提出基于图卷积网络(graph convolution networks, GCN)的HeteGraph框架, 通过一种抽样技术和一个图卷积运算来学习高质量图的节点嵌入. CHEN等^[20]将提出一种基于KG的交互式规则引导推荐(interactive rules-guided recommender, IR-Rec)框架. LI等^[21]利用图的边信息提出深层端到端推荐(graph side information for recommendation, GSIRec)框架, 利用图注意力神经网络来提高推荐效果, 能够解决数据稀疏和冷启动问题. 针对KG无法自动捕获实体对推荐的长期依赖关系, SANG等^[22]提出知识图谱增强神经协同过滤残差递归网络(knowledge graph enhanced neural collaborative filtering with residual recurrent network, KGNCN-RRN)的双通道神经交互方法, 既能捕捉丰富的语义信息, 还能捕捉用于推荐的用户与项目之间复杂的隐含关系. ZHENG等^[23]提出一种基于变分自动编码器(variational auto-encoder, VAE)的隐式信任关系感知社交推荐模型(implicit trust relation-aware model, ITRA), 通过注意模块将加权信任嵌入信息反馈到继承的非线性VAE结构中, 并重建了一个社交矩阵进行推荐.

为提高模型的表达能力, 基于深度学习的推荐

模型采用了不同的特征交叉方式, 但仍存在特征交叉时无法充分利用嵌入向量信息的问题. 为更好地利用嵌入向量信息, 本研究提出一种基于分离嵌入交叉网络(separated embedding interaction networks, SEIN)的模型, 它具有显式特征交叉的特点, 通过对不同维度的特征进行交叉, 能够比其他深度学习模型挖掘出更深的隐藏信息, 更充分地利用嵌入层的信息, 使模型表征能力更好.

1 SEIN模型

如图1, SEIN模型主要包括特征嵌入层、分离嵌入交叉网络和预测层. 其中, $\text{field}_1, \text{field}_2, \dots, \text{field}_m$ 为特征域, 即独热编码后的原始数据.

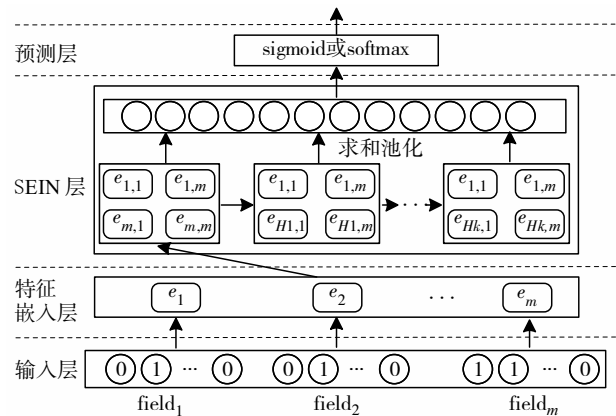


图1 SEIN模型结构

Fig. 1 The architecture of SEIN model.

1.1 特征嵌入层

在计算机视觉领域, 输入基本上是图像或稠密的数据, 其空间局部相关性往往很强. 因此, 可直接通过全连接层或卷积层处理计算机视觉领域的数[26-28]. 推荐系统中的样本往往有成百万甚至上千万个, 根据特征进行独热(one-hot)编码后, 矩阵往是高维度的稀疏矩阵, 若直接使用全连接层或卷积层来处理输入数据, 会产生过多的网络参数, 导致训练效率过低. 因此, 通常是将原始数据学习表示成隐嵌入, 一个隐嵌入向量表示一个特征字段. 性质和来源相同的字段可归结为同一字段, 如性别和户籍等, 且同一字段的数据类型一致, 把数据的特征按照字段的方式拼接起来, 便组成了一段以0和1编码的特征.

嵌入层的作用是把每个特征都投影到一个固定长度的隐嵌入向量, 因此原本稀疏离散、维度不确定的特征经嵌入层处理后变成维度一样, 稠密连续

的数值向量. 嵌入层的输出为一个 $m \times D$ 的矩阵, 表示为 $\mathbf{e} = [\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_m]^T$. 其中, $\mathbf{e}_i \in \mathbf{R}^D$ 为第 i 个字段的嵌入表示; D 为字段嵌入的维度.

1.2 SEIN 模型

SEIN 模型主要有以下特点: ① 特征间的交互发生在向量级, 而非元素级上; ② 不同特征嵌入层是分离的, 交互阶数可通过网络层数来控制; ③ 交叉网络所需的参数和计算复杂度不会随着交互阶数呈指数级增长.

在 SEIN 结构中, 为把特征嵌入层中的每个字段的隐嵌入向量当作一个整体单元来处理, 采用 xDeepFM 模型中压缩交互网络的做法, 避免同一个向量内部的元素之间发生交互. 因此, 特征嵌入层的宽向量形式化表示成矩阵 $\mathbf{X}^0 \in \mathbf{R}^{m \times D}$, m 和 D 分别为特征字段的个数和嵌入向量的维度. 该矩阵中的第 d 列为第 d 个维度的隐嵌入向量, 用 $\mathbf{X}_d^0$ 表示, 将 $\mathbf{X}^0$ 中不同维度的隐嵌入向量与自身的转置做矩阵乘法, 可得第 d 维的初始交叉矩阵为

$$\mathbf{X}_d^1 = \mathbf{X}_d^0 \cdot (\mathbf{X}_d^0)^T \quad (1)$$

其中, $\mathbf{X}_d^1 \in \mathbf{R}^{m \times m}$. 在每个维度重复上述操作后得到一个三维张量 $\mathbf{X}^1 \in \mathbf{R}^{m \times m \times D}$, D 为嵌入为维度. 计算的详细过程见图 2. 其中, $\otimes$ 表示向量外积操作.

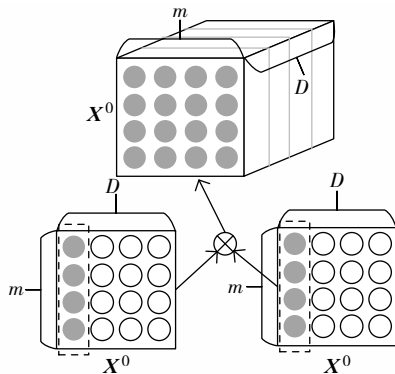


图 2 初始交叉张量计算过程

Fig. 2 Calculation process of initial interaction tensor.

令 $\mathbf{X}^k$ 表示 SEIN 中第 k 层隐藏层的状态, 则 $\mathbf{X}^k \in \mathbf{R}^{m \times H_k \times D}$ 也是一个张量, H_k 为第 k 层隐藏层的特征个数. $\mathbf{X}^k$ 可看作由 D 个二维矩阵拼接而成, 其中第 d 个嵌入维度上的隐嵌入矩阵为 $\mathbf{X}_d^k \in \mathbf{R}^{m \times H_k}$. 于是 SEIN 中第 k 层 ($k \geq 2$) 隐藏层中第 d 维的交叉矩阵状态为

$$\mathbf{X}_d^k = \mathbf{W}_d \cdot \mathbf{X}_d^{k-1} \quad (2)$$

其中, $\mathbf{W}_d \in \mathbf{R}^{H_{k-1} \times H_k}$ 是第 k 层隐藏层中 d 维嵌入矩阵的参数矩阵. 将 D 个交叉矩阵拼接起来, 便能得到

$\mathbf{X}^k$. 因为 $\mathbf{X}^k$ 是基于 $\mathbf{X}^{k-1}$ 得到的, 因此 $\mathbf{X}^k$ 的阶数比 $\mathbf{X}^{k-1}$ 高一阶. 可见, SEIN 每增加 1 层隐藏层, 能获得的特征交叉的最大阶数就增加 1, 图 3 为隐藏层的具体计算过程. 从结构来看, SEIN 是由多个卷积神经网络 (convolutional neural network, CNN) 或全连接网络组成, 且分别对不同维度的嵌入隐向量进行交叉, 每次交叉的状态都与前一层有关, 通过式 (1) 与式 (2) 的转换实现向量级特征交互, 并与不同维度的嵌入层的分离, 加强了模型的表达能力.

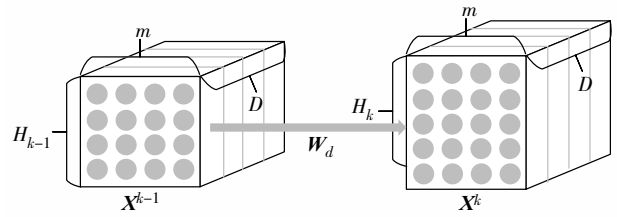


图 3 SEIN 隐藏层计算过程

Fig. 3 Calculation process of hidden layer of SEIN.

实际上, 每个隐藏层中每个嵌入维度都可铺平为 1 个一维向量, 这对计算过程并不会产生影响. 求得全部 k 层隐藏层的状态后, 对所有隐藏层都进行求和池化操作则得到该隐藏层的特征交叉结果, 即

$$\mathbf{p}^k = \sum_{d=1}^D \mathbf{X}_{d,h}^{k-1} \quad (3)$$

其中, $h \in [1, m \times H_k]$. 通过把隐藏层里每个维度的池化结果拼接起来, 得到该隐藏层的新向量为 $\mathbf{p}^k = [p_1^k, p_2^k, \dots, p_{m \times H_k}^k]$. 将所有隐藏层的池化向量水平拼接起来得到一个宽向量, 作为 SEIN 的最终输出 $\mathbf{p}^+ = [p^1, p^2, \dots, p^k]$. SEIN 的宏观架构如图 4.

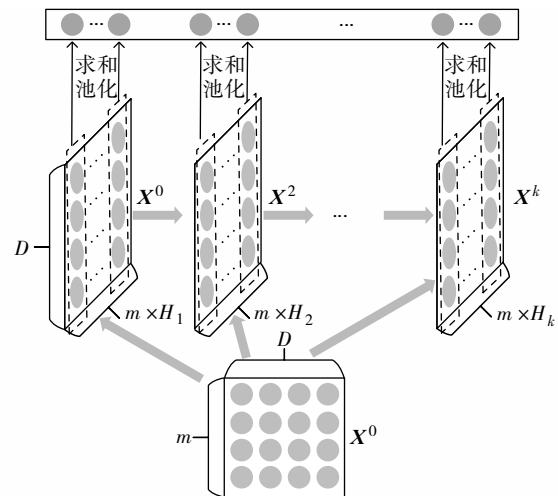


图 4 SEIN 宏观架构

Fig. 4 The macroscopic architecture of SEIN.

将SEIN用于广告点击率预测或商品推荐任务,即二分类与多分类任务,预测层采用sigmoid或softmax函数,最终的预测结果可表示为 $y = \text{sigmoid/softmax}(p^+)$.

2 实验

2.1 数据集

为验证SEIN模型的有效性和泛用性,本研究设置两个实验:实验1为在Criteo和AutoML数据集上进行广告点击率预估(click through rate, CTR). CTR指网络广告的点击通过率,该实验本质上是一个二分类任务,通过预测用户是否点击广告来对用户进行广告推荐.实验2在MovieLens-1M数据集上进行top- k 推荐实验.Top- k 指从最后的按得分排序的推荐列表中返回前 k 个评分最高的结果.Criteo数据集包含其7 d内的部分流量,每行对应1个由Criteo提供的展示广告,正(点击)和负(未点击)示例都以不同的比率进行子采样,以减少数据集大小.Criteo数据集包括label、I1—I13和C1—C26共计40列,label列的值只有0或1,其中有0为未点击,1为被点击,I1—I13为13个数值特征列,C1—C26为26个类别特征列,并且这些数值和类别特征都做了匿名处理.AutoML数据集有76个匿名特征,其中有16个特征的值完全一致,因此删除这16个特征,采用剩余60条特征进行实验.MovieLens-1M数据集包含了来自6 040名用户对3 592部电影的100万条评分数据,分为评分、用户信息和电影信息3个表.

表1 实验用数据集描述

Table 1 Description of the dataset used in the experiment

数据集	数值特征 个数	类别特征 个数	$10^{-7} \times$ 数据 实例个数
Criteo	13	26	1 000
AutoML	42	18	136
MovieLens-1M	—	—	1 000

2.2 实验环境

使用tensorflow-gpu 2.8.0框架实现所有模型,并调整超参数以记录各模型的最佳得分.实验环境配置为:GPU为Nvidia GeForce RTX 2080 Super, 16 Gbyte显存,编程语言为Python3.9.

2.3 评估指标

为验证模型性能,实验1使用曲线下面积(area

under the curve, AUC)^[29]与对数损失^[30]作为评估性能指标.AUC计算公式如式(4),取值范围为0.5~1.0,AUC值越接近1,说明检测方法真实性越高,AUC=0.5则说明模型无应用价值.

$$\text{AUC} = \frac{\sum_{\text{PCI} \in \text{PC}} r_i - \frac{P \times (P+1)}{2}}{P \times N} \quad (4)$$

其中,PC为正样本,PCI为正样本实例, $\sum_{\text{PCI} \in \text{PC}}$ 为对所有正样本形成的正负样本对(正样本预测值>负样本预测值)求和; r_i 为第 i 个样本的序号(样本概率由小到大排列); P 和 N 分别为正负样本数量.

对数损失衡量的是真实值与预测值的吻合程度,记为 $L_{\log}$,计算公式如式(5).对数损失通过惩罚错误的分类,实现对分类器准确度的量化. $L_{\log}$ 值越小越好.

$$L_{\log} = -\frac{1}{M} \sum_{i=1}^M [y_i \ln p_i + (1 - y_i) \ln (1 - p_i)] \quad (5)$$

其中, M 为样本数量; y_i 为样本的标签值, $y_i \in \{0, 1\}$; p_i 为对应标签的概率.

实验2使用以下4个指标作为评估性能指标:平均绝对误差(mean absolute error, MAE)评估预测评分与真实评分之间的偏差,计算公式如式(6).其中, r_{ui} 和 y_{ui} 分别表示用户 u 与项目 i 之间的预测评分与实际评分,MAE对异常值不敏感.

$$\text{MAE} = \frac{1}{|R|} \sum_{r_{ui} \in R} |r_{ui} - y_{ui}| \quad (6)$$

均方根误差(root mean square error, RMSE)是推荐任务中常用的精度测量方法,其计算公式如式(7).相比与MAE, RMSE预测评分与真实评分的精度差,它对异常值敏感.在实验2中,将RMSE设置为损失值.

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (r_i - y_i)^2} \quad (7)$$

准确率(precision at k , 记为 $\text{Pr}@k$)与召回率(recall at k , 记为 $\text{Re}@k$): $\text{Pr}@k$ 与 $\text{Re}@k$ 是一种来自信息检索领域的衡量推荐系统的指标.它们是决策支持指标,不考虑预测评分与真实评分的偏差,只反应推荐列表对与活跃用户的相关性.其计算公式如式(8)与式(9).其中, $\text{TP}@k$ 表示前 k 项中预测正确的数量, k 为推荐列表的长度, R_u 为实际列表的长度.

$$\text{Pr}@k = \frac{\text{TP}@k}{k} \quad (8)$$

$$\text{Re}@k = \frac{\text{TP}@k}{R_u} \quad (9)$$

3 实验结果与分析

3.1 CTR 预测模型性能对比

为验证模型的有效性, 将 SEIN 模型与主流的深度学习推荐模型 FM、NFM、AFM、DeepFM、Deep&Cross 与 xDeepFM 进行对比, 实验结果如表 2. 所有实验结果均为多次实验调参得到的最优结果.

表2 主流 CTR 预测模型的评估指标表现

Table 2 Performance of the popular CTR prediction model

模型	Criteo 数据集		AutoML 数据集	
	AUC	$L_{\log}$	AUC	$L_{\log}$
FM	0.516 1	0.707 7	0.187 9	0.642 7
NFM	0.487 2	0.756 7	0.194 3	0.724 6
AFM	0.476 2	0.767 7	0.177 8	0.731 5
DeepFM	0.481 1	0.758 8	0.176 3	0.747 4
Deep&Cross	0.482 9	0.759 5	0.195 0	0.726 1
xDeepFM	0.481 5	0.759 1	0.173 2	0.746 4
SEIN	0.463 0	0.782 6	0.169 7	0.752 1

由表 2 可见, SEIN 模型的性能较其他基准模型都有所提升. 对比基准模型 DeepFM、Deep&Cross 与 xDeepFM, SEIN 模型在 Criteo 数据集上的 AUC 分别提高了 2.38%、2.31% 和 2.35%, $L_{\log}$ 分别下降了 1.81%、1.99% 和 1.85%; 在 AutoML 数据集上的 AUC 分别提高了 1.17%、2.60% 和 0.57%, $L_{\log}$ 分别下降了 0.66%、2.53% 和 0.35%. 因此, SEIN 模型相对于其他模型, 对不同维度嵌入向量特征分离交叉得到了一定的效果, 证明了模型性能的有效性.

3.2 融合图神经网络模型对比

为进一步验证模型的有效性以及泛用性, 将本研究 SEIN 模型与近年来一些基于图神经网络模型 HeteGraph、IR-Rec、GSIRec 和 KGNCf-RNN 融合, 融合后的新模型依次命名为 HeteGraph-SEIN、IR-Rec-SEIN、GSIRec-SEIN 和 KGNCf-RNN-SEIN, 并在 MovieLens-1M 数据集上对这几种模型性能进行对比, 结果如表 3, 各模型实验结果均经过多次实验调参得到的最优结果. 由表 3 可见, 在 MAE、RMSE、Pr@5、与 Re@5 四个评估指标上, 融合 SEIN 的图神经网络模型表现的性能更好. 因此, 图神经网络模型融合 SEIN 后, 推荐性能会得到一定的提升, 进一步验证了模型的有效性和泛用性.

表3 融合图神经网络模型在 MovieLens 数据集上的评估指标表现

Table 3 Performance of the fused graph neural network model in the MovieLens dataset

模型	MAE	RMSE	Pr@5	Re@5
HeteGraph	1.355 2	1.138 2	0.180 6	0.052 7
HeteGraph-SEIN	1.292 8	1.082 6	0.193 3	0.057 6
IR-Rec	1.291 8	1.089 4	0.205 6	0.067 6
IR-Rec-SEIN	1.263 4	1.056 3	0.210 3	0.069 7
GSIRec	1.337 1	1.109 5	0.201 8	0.061 3
GSIRec-SEIN	1.279 1	1.065 5	0.206 6	0.064 8
KGNCf-RNN	1.353 3	1.125 3	0.189 7	0.058 1
KGNCf-RNN-SEIN	1.331 9	1.098 4	0.195 3	0.063 1
ITRA	1.386 2	1.172 2	0.171 5	0.050 8
ITRA-SEIN	1.329 6	1.105 6	0.197 4	0.064 4

3.3 参数分析

在 SEIN 模型中, 嵌入维度与 SEIN 层数及每层神经元数量的设置直接影响模型的性能, 因此本研究针对上述参数在 Criteo 与 AutoML 数据集上进行多次实验, 结果如图 5、图 6 和图 7.

在研究嵌入维度对模型性能影响的实验中, SEIN 的层数设为 1, 即包括初始交叉层, 模型共含

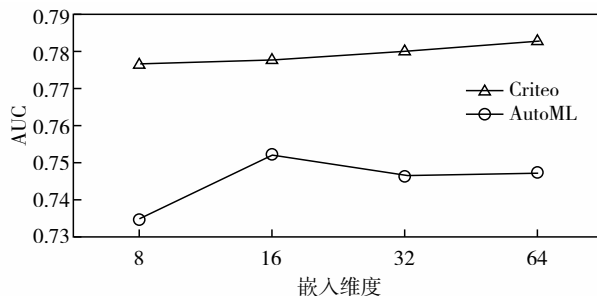


图5 不同嵌入维度下模型性能

Fig. 5 Performance of different embedding dimensions.

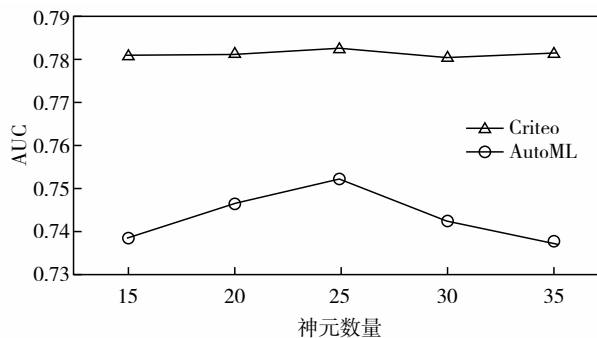


图6 不同神经元数量下模型性能

Fig. 6 Performance of different number of neurons.

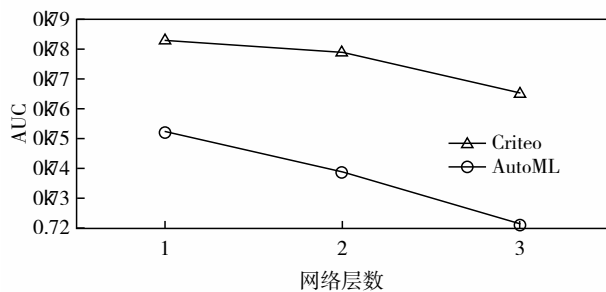


图7 不同SEIN层数下模型性能

Fig. 7 Performance of different layer number of SEIN.

有2个隐藏层, 在Criteo和AutoML数据集中神经元数量都设为25. 由图5可见, 模型随着嵌入维度的增加, 性能基本上也随之增强, 因为嵌入维度越大, 挖掘到的信息也会越多, 表达能力会变得更强大. 但当嵌入维度增加时, 性能反而可能会降低, 原因可能是不同数据集中信息量与分布等不一样. 在研究神经元数量对模型性能影响的实验中, SEIN的层数仍设为1, Criteo数据集的嵌入维度设置为64, AutoML数据集的嵌入维度设置为16. 由图6可见, 在神经元数量为25时, 模型性能处于最佳状态, 神经元数量大于或小于25时模型的性能均下降, 且在AutoML数据集上的表现尤为明显. 在研究SEIN层数对模型性能影响的实验中, 在Criteo数据集的嵌入维度设为64, AutoML数据集的嵌入维度设为16, 神经元的数量均设为25. 由图7可见, 当SEIN的层数为1时, 模型的性能处于最佳状态, 且随着层数的增加, 模型的性能呈下降趋势. 这可能是由于数据集的数量或者类别太少, 而多层SEIN的表达能力过强而导致过拟合. 因此, 针对不同的数据集需选择合适的嵌入维度、神经元数量以及SEIN层数.

结 语

提出基于分离嵌入交叉网络的推荐模型, 通过对不同嵌入向量分别进行训练, 在进行特征交叉时避免了不同维度嵌入向量之间的影响, 提升了模型的性能. CTR预测实验与top-k推荐的实验结果表明, SEIN模型在Criteo与AutoML数据集上的性能优于其他基准模型, 并且引入基于图神经网络的推荐系统中, 在电影推荐场景中也能提高模型精度. 目前, 深度学习推荐模型迅速发展, 数据正在海量增加, 若能将许多以前不能利用的信息或方法都结合起来, 或许能达到更好的效果. 自然语言处理领

域也是当前一个研究热点, 下一步可尝试将用户行为序列、内容与自然语言处理领域的一些方法结合, 利用更多有效的信息, 对模型进行优化.

基金项目: 上海高校应用型本科试点建设资助项目(Z32004-17-84); 上海市教育委员会一流本科专业建设专项资助项目(JYLB202002)

作者简介: 封书蕾(1996—), 上海海洋大学硕士研究生. 研究方向: 推荐系统. E-mail: shalei1207@126.com

引 文: 封书蕾, 蒋中云. 基于分离嵌入交叉网络的推荐模型[J]. 深圳大学学报理工版, 2023, 40(4): 513-520.

参考文献 / References:

- [1] NAKAMURA A, ABE N. Collaborative filtering using weighted majority prediction algorithms [C]// Proceedings of the Fifteenth International Conference on Machine Learning. San Francisco, USA: Morgan Kaufmann Publishers Inc., 1998: 395-403.
- [2] RICH E. User modeling via stereotypes [J]. Cognitive Science, 1979, 3(4): 329-354.
- [3] PAZZANI M J, BILLSUS D. Content-based recommendation systems [M]// BRUSILOVSKY P, KOBAS A, NEJDL W. The Adaptive Web: Methods and Strategies of Web Personalization. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007: 325-341.
- [4] BALABANOVIĆ M, SHOHAM Y. Fab: content-based, collaborative recommendation [J]. Communications of the ACM, 1997, 40(3): 66-72.
- [5] SHAN Ying, HOENS T R, JIAO Jian, et al. Deep crossing: web-scale modeling without manually crafted combinatorial features [C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: Association for Computing Machinery, 2016: 255-262.
- [6] QU Yanru, CAI Han, REN Kan, et al. Product-based neural networks for user response prediction [C]// IEEE 16th International Conference on Data Mining (ICDM). Piscataway, USA: IEEE, 2016: 1149-1154.
- [7] CHENG H T, KOC L, HARMSEN J, et al. Wide & deep learning for recommender systems [C]// Proceedings of the 1st Workshop on Deep Learning for Recommender Systems. New York, USA: Association for Computing Machinery, 2016: 7-10.
- [8] HE Xiangnan, LIAO Lizi, ZHANG Hanwang, et al. Neural collaborative filtering [C]// Proceedings of the 26th International Conference on World Wide Web. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2017: 173-182.
- [9] GUO Huifeng, TANG Ruiming, YE Yunming, et al. DeepFM: a factorization-machine based neural network for CTR prediction [C]// Proceedings of the 26th International Joint Conference on Artificial Intelligence. Palo Alto,

- USA: AAAI Press, 2017: 1725-1731.
- [10] HE Xiangnan, CHUA T S. Neural factorization machines for sparse predictive analytics [C]// Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval. New York, USA: Association for Computing Machinery, 2017: 355-364.
 - [11] XIAO Jun, YE Hao, HE Xiangnan, et al. Attentional factorization machines: learning the weight of feature interactions via attention networks [C]// Proceedings of the 26th International Joint Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2017: 3119-3125.
 - [12] LIAN Jianxun, ZHOU Xiaohuan, ZHANG Fuzheng, et al. XDeepFM: combining explicit and implicit feature interactions for recommender systems [C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, USA: Association for Computing Machinery, 2018: 1754-1763.
 - [13] ZHOU Guorui, ZHU Xiaoqiang, SONG Chenru, et al. Deep interest network for click-through rate prediction [C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, USA: Association for Computing Machinery, 2018: 1059-1068.
 - [14] ZHOU Guorui, MOU Na, FAN Ying, et al. Deep interest evolution network for click-through rate prediction [C]// Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2019: 5941-5948.
 - [15] RENDLE S, GANTNER Z, FREUDENTHALER C, et al. Fast context-aware recommendations with factorization machines [C]// Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: Association for Computing Machinery, 2011: 635-644.
 - [16] SEDHAIN S, MENON A K, SANNER S, et al. AutoRec: autoencoders meet collaborative filtering [C]// Proceedings of the 24th International Conference on World Wide Web. New York, USA: Association for Computing Machinery, 2015: 111-112.
 - [17] HUANG Jin, ZHAO W X, DOU Hongjian, et al. Improving sequential recommendation with knowledge-enhanced memory networks [C]// The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. New York, USA: Association for Computing Machinery, 2018: 505-514.
 - [18] SHA Xiao, SUN Zhu, ZHANG Jie. Hierarchical attentive knowledge graph embedding for personalized recommendation [J]. Electronic Commerce Research and Applications, 2021, 48: 101071.
 - [19] TRAN D H, SHENG Q Z, ZHANG W E, et al. HeteGraph: graph learning in recommender systems via graph convolutional networks [J/OL]. Neural Computing and Applications, 2021. <https://doi.org/10.1007/s00521-020-05667-z>.
 - [20] CHEN Jiaying, YU Jiong, LU Wenjie, et al. IR-Rec: an interpretive rules-guided recommendation over knowledge graph [J]. Information Sciences, 2021, 563: 326-341.
 - [21] LI Anchen, YANG Bo. GSIRec: learning with graph side information for recommendation [J]. World Wide Web, 2021, 24(5): 1411-1437.
 - [22] SANG Lei, XU Min, QIAN Shengsheng, et al. Knowledge graph enhanced neural collaborative filtering with residual recurrent network [J]. Neurocomputing, 2021, 454: 417-429.
 - [23] ZHENG Qiqi, LIU Guanfeng, LIU An, et al. Implicit relation-aware social recommendation with variational auto-encoder [J]. World Wide Web, 2021, 24(5): 1395-1410.
 - [24] BORDES A, USUNIER N, GARCIA-DURÁN A, et al. Translating embeddings for modeling multi-relational data [C]// Proceedings of the 26th International Conference on Neural Information Processing Systems: Volume 2. Red Hook, USA: Curran Associates Inc., 2013: 2787-2795.
 - [25] ZHANG Jiani, SHI Xingjian, KING I, et al. Dynamic key-value memory networks for knowledge tracing [C]// Proceedings of the 26th international conference on World Wide Web. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2017: 765-774.
 - [26] SUN Ke, XIAO Bin, LIU Dong, et al. Deep high-resolution representation learning for human pose estimation [C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, USA: IEEE, 2019: 5686-5696.
 - [27] DING Haisong, CHEN Kai, HUO Qiang. Compressing CNN-DBLSTM models for OCR with teacher-student learning and Tucker decomposition [J]. Pattern Recognition, 2019, 96: 106957.
 - [28] HU Han, GU Jiayuan, ZHANG Zheng, et al. Relation networks for object detection [C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2018: 3588-3597.
 - [29] WALTER S D. The partial area under the summary ROC curve [J]. Statistics in Medicine, 2005, 24(13): 2025-2040.
 - [30] VOVK V. The fundamental nature of the log loss function [M]// BEKLEMISHEV L D, BLASS A, DERSHOWITZ N, et al. Fields of Logic and Computation II: Essays Dedicated to Yuri Gurevich on the Occasion of his 75th Birthday. Cham, Switzerland: Springer International Publishing, 2015: 307-318.

【中文责编：英子；英文责编：木柯】