计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
[doi: 10.15888/j.cnki.csa.009898] [CSTR: 32024.14.csa.009898]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

知识图谱增强的广告推荐算法①

郑翠春[1], 林欣扬[1], 骆龙泉[1], 汪璟玢[2]

[1](厦门众联世纪股份有限公司, 厦门 361008)
[2](福州大学 计算机与大数据学院, 福州 350108)
通信作者: 林欣扬, E-mail: linxy@zhonglian.com

摘　要: 随着互联网广告市场的快速增长, 精准的广告推荐变得至关重要. 如何有效学习用户特征和广告特征之间交互是点击率 (CTR) 与转化率 (CVR) 预测任务的关键. 然而, 现有的点击率与转化率预测模型存在特征依赖性偏差和广告语义信息挖掘不足的问题. 为此, 本文提出了一种知识图谱增强的广告推荐算法 (knowledge graph-enhanced advertisement recommendation algorithm, KGEARA). 该算法通过构建知识图谱将结构化数据转化为三元组的形式, 有效地整合广告特征信息并捕捉广告间的关联性. 通过知识图谱表示学习将这些特征转化为嵌入表示, 以融合广告的语义特征并捕捉交互细节. 进一步利用广告特征嵌入与其他特征嵌入结合, 通过专家网络、门控网络和任务塔预测点击率和转化率, 并引入逆向倾向评分 (IPS) 处理点击倾向不均的问题, 以纠正预测偏差. 在广告真实数据集上进行了广泛实验, 实验结果验证了模型在提升 CTR 和 CVR 预测准确性方面的有效性.

关键词: 点击率预测; 转化率预测; 知识图谱; 表示学习

引用格式: 郑翠春,林欣扬,骆龙泉,汪璟玢.知识图谱增强的广告推荐算法.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/9898.html

Knowledge Graph-enhanced Advertisement Recommendation Algorithm

ZHENG Cui-Chun[1], LIN Xin-Yang[1], LUO Long-Quan[1], WANG Jing-Bin[2]

[1](Xiamen Zhonglian Century Co. Ltd., Xiamen 361008, China)
[2](College of Computer and Data Science, Fuzhou University, Fuzhou 350108, China)

Abstract: With the rapid growth of the Internet advertising market, accurate advertisement recommendations have become crucial. Effectively capturing the interaction between user and advertisement features is key to improving the prediction of click-through rate (CTR) and conversion rate (CVR). However, existing CTR and CVR prediction models suffer from feature dependency bias and insufficient semantic information extraction from ads. To address these issues, this study proposes a knowledge graph-enhanced advertisement recommendation algorithm (KGEARA). The algorithm constructs a knowledge graph by converting structured data into triplets, effectively integrating advertisement feature information and capturing relationships between ads. Through knowledge graph representation learning, these features are transformed into embeddings that merge the semantic characteristics of ads and capture interaction details. Further, by combining advertisement feature embeddings with other feature embeddings, CTR and CVR are predicted through expert networks, gated networks, and task-specific towers. In addition, inverse propensity scoring (IPS) is introduced to address click-bias issues and correct prediction biases. Extensive experiments on real-world advertisement datasets demonstrate the effectiveness of the proposed model in improving CTR and CVR prediction accuracy.

Key words: click-through rate (CTR) prediction; conversion rate (CVR) prediction; knowledge graph; representation learning

① 基金项目: 福建省自然科学基金 (2021J01619)
收稿时间: 2024-10-08; 修改时间: 2025-01-15; 采用时间: 2025-02-11; csa 在线出版时间: 2025-04-30

随着 5G 时代到来, 以及各种手机应用的出现, 我国互联网吸引了大量用户, 据统计中国网民规模近 11 亿, 这给互联网广告带来了巨大商机, 吸引了越来越多的广告商进入互联网市场, 极大促进了个性化广告推荐技术的发展[1]. 点击率 (click-through rate, CTR) 和转化率 (conversion rate, CVR) 是衡量广告效果的重要指标, 可以帮助企业了解用户点击和购买行为的潜在逻辑. 因此准确预测 CTR 和 CVR 对于广告推荐具有重要意义.

在这一背景下, 运营商流量包广告推荐逐渐发展成为一个独特的应用场景. 传统的流量包营销模式通常依赖于静态规则和固定折扣策略, 难以满足用户日益多样化的需求. 随着移动互联网的普及和用户在线行为的复杂化, 运营商开始意识到, 仅依靠过往经验制定套餐和促销方案, 已无法满足现代用户的个性化需求和多变的流量使用习惯.

尤其在 5G 推动下, 网络流量使用模式发生了显著变化. 不同地区的用户因地域经济水平差异, 对流量资费敏感度不尽相同; 不同年龄段的用户在流量使用偏好上也呈现明显差异; 此外, 用户所处的使用场景 (如家庭场景、出差场景或旅行场景) 直接影响其对流量包的需求. 这些复杂多样的需求变化催生了精准广告投放与套餐推荐的实际需求. 在这一过程中, CTR 和 CVR 预测模型成为关键工具, 通过深入挖掘用户行为和流量包特征的深层联系, 不仅为用户提供更加贴合实际需求的流量套餐选择, 还能帮助运营商提升整体业务收益, 实现双方共赢的目标.

在广告推荐中, 如何有效学习用户特征和广告特征之间交互是 CTR、CVR 预测任务的关键挑战. 早期学者主要采用传统的逻辑回归、协同过滤等方法, 但这些方法是简单的机器学习模型, 在拟合非线性数据时存在分类能力不足的问题. 近年来, 研究者开始探索深度学习和多任务学习方法, 以提高模型预测性能. MMoE[2]模型在多任务模型中引入共享专家层, 显式地学习各子任务间的关系, 同时利用门控机制来平衡专家网络间匹配的权重. ESMM[3]模型同时对点击率和转化率进行建模, 采用共享参数的策略, 减少算法参数量, 提高了模型泛化能力. 但实际应用中发现上述模型在工业数据集中存在特征依赖性偏差, 即模型严重依赖于部分属性, 例如基本的用户特征 (如年龄、性别等), 而忽略了广告的诸多潜藏特征 (如商品地域、产品年龄定位等). 这些方法未能深入地挖掘广告数据中隐藏的语义信息.

知识图谱提供了一种有效方式来表示和利用广告之间的复杂关系, 它通过将实体和关系组织成结构化图形, 使得更为复杂的语义信息能够被计算机识别和处理. 因此, 将知识图谱与 CTR 和 CVR 预测相结合, 能够更加全面地捕捉广告之间的内在联系, 且有助于提高模型的预测能力.

为此, 本文提出了知识图谱增强的广告推荐算法 (knowledge graph-enhanced advertisement recommendation algorithm, KGEARA), 本文设计了一种广告知识图谱构建算法, 将结构化数据抽取成知识图谱的三元组形式, 以整合广告信息, 充分利用广告潜在特征. 进一步通过知识图谱表示学习将其转化为嵌入表示. 这些嵌入不仅融合了广告的潜在特征, 还捕获了广告间的交互信息. 在此基础上, 我们将预训练的广告嵌入表示与其他特征进行拼接作为初始输入, 利用专家网络、门控网络和任务塔来预测点击率和转化率. 考虑到点击率和转化率之间存在因果关联, 引入逆向倾向评分 (inverse propensity scoring, IPS) 来处理点击倾向不均的问题, 从而在计算损失时对不同样本赋予不同权重, 纠正由此产生的偏差.

相比于传统模型, KGEARA 在运营商广告对接任务具有以下创新: 本文通过将广告数据构建为知识图谱, 整合广告的结构化信息并捕捉广告间的潜在关联, 提升语义建模能力; 设计了多任务的学习框架, 通过特征共享与任务独立模块提高模型的语义捕获能力和预测协同性, 将知识图谱与多任务预测相结合; 结合广告知识图谱嵌入和用户动态行为特征, 建模广告与用户的复杂交互, 捕捉用户偏好的动态变化, 提升推荐精准性与适应性.

1 相关工作

1.1 点击率与转化率预测

点击率和转化率预测的关键任务是如何建模使得特征信息能够有效交互. 逻辑回归 (logistic regression, LR)[4]是其中应用最广泛的方法之一, 通过线性组合将特征加权映射到 0–1 之间的概率值. 近年来, 嵌入和 MLP 范式引起关注, 它们主要聚焦于特征交互. FM[5]通过建模特征间的交互关系来捕捉复杂关系. FFM[6]是 FM 的扩展, 能更好地处理“字段”结构的特征. Deep-

FM[7]在宽部分使用 FM 代替 LR 来学习二阶特征交互, DCN[8]使用 Cross Network 学习高阶表示以自动获取显式的高级交叉特征. PNN[9]将类别特征和嵌入向量乘积串联为 MLP 输入. xDeepFM[10]设计压缩交互网络来明确学习低阶和高阶特征交互. AutoInt[11]利用多头自注意机制将特征投射到多个子空间, 捕获不同特征交互. 某些模型认为行为序列有助于学习, 例如 DIN[12]、DIEN[13]和 DSIN[14]. DIN 学习用户过去行为与目标项目的注意权重. DIEN 使用带辅助损失的双层 RNNs 建模特定目标项的用户兴趣演化. DSIN 按时间分离用户行为序列, 并提取用户兴趣. 高阶注意分解机、RILKE[15]等方法也提出了创新性解决方案, 以克服数据稀疏的问题. DCIN[16]设置了一个上下文交互单元来学习决策上下文, 从而实现更准确地 CTR 预测. OptFS[17]通过联合选择特征及其交互, 从而优化特征集以提升预测性能并降低计算成本. 这些模型的不断演进推动了点击率和转化率预测领域的进步.

近年来, 单任务推荐研究已经取得诸多成果, 但在此基础上进一步提升推荐质量似乎遭遇瓶颈, 越来越多的研究者开始关注多任务领域研究. 多任务学习以集成的方式训练模型, 提高多个任务的综合性能. Shared-Bottom model[18]是一种早期方法, 将底层结构用于所有任务, 然后针对每个任务分别设立头部. MoE 模型将底层结构划分为多个专家 (每个专家是一个前向神经网络), 通过一个通用的门控机制来调整不同任务专家的权重, 实现软共享. ESMM 同时处理点击率预测和转化率预测目标, 分别使用曝光点击数据和点击转换数据进行预测. ESCM^2[19]使用反事实预测来解决 ESMM 在转化率预测上的天然高估而忽略了点击到转化的因果关系. PACC[20]通过概率分解和位置嵌入的方式, 联合消除 CTR 和 CVR 预测中的位置偏差. 以上方法通过共享模型的嵌入, 为不同任务使用不同模型, 通过概率关联多个任务, 实现多任务学习.

1.2 知识图谱

知识图谱 (knowledge graph) 是一种用于表示知识的图形化结构, 它由实体 (节点) 和它们之间的关系 (边) 组成. 它的目标是将大量的信息整合到一个结构化框架中, 使计算机能够理解和推理出各种关系.

在知识图谱表示学习方面, 翻译模型[21−23]通过距离函数度量三元组中的头尾实体语义相关性; 匹配模型引入关系参数矩阵, 强化实体与关系之间的语义匹配程度[24−26]. 近年来, 各种复杂的知识图谱表示学习模型被设计, 以学习实体和关系的分布式表示, 有效挖掘知识图谱内的关联信息. 这为利用知识图谱进行推理和完成下游任务提供了可能. 但是如何将知识图谱表示学习有效应用于广告点击率与转化率预测领域, 还有待深入探索.

1.3 基于知识图谱增强的推荐算法

知识图谱应用于推荐系统的方法主要是通过图嵌入的方式来对实体和关系进行表示, 进而扩充原有商品和用户特征的语义信息. KSR[27]使用 TransE 生成知识图谱中的实体和商品表征, 并利用键值对记忆网络, 学习得到用户的细粒度动态特征, 从而提升广告推荐效果. RippleNet[28]从异质网络图中抽取用户节点相连的实体节点, 利用这些节点的嵌入表征更新用户表征, 从而利用用户和商品表征的点击来预测结果. MCRec[29]使用卷积神经网络对不同元路径采样得到的从用户到商品的路径进行嵌入表征, 进而构造基于元路径的偏好特征, 并结合 NeuMF[30]的算法构建推荐系统. 除此之外, MDKR[31]利用 DeepFM 提取低阶线性特征, 通过元学习增强推荐模块与知识图谱嵌入模块之间的高维特征交互, 以此解决冷启动问题. 在应用方面, 有研究者利用长文档建模技术将知识图谱增强的新闻文档, 通过捕获文档间的词级交互获得细粒度的用户表示[32]. 但是, 现有的模型还未将知识图谱与多任务预测相结合.

2 KGEARA 模型

KGEARA 模型是一种融合知识图谱的, 点击率和转化率预测模型, 总体框架如图 1 所示, 分为 4 个部分: (1) 知识图谱构造模块, 负责将结构化的广告数据转化为形如知识图谱存储常见的三元组形式. (2) 知识图谱表示学习模块, 将构造好的知识图谱利用知识图谱表示学习算法得到预训练的广告知识图谱嵌入表示. 上述两个模块构成本文的知识图谱模块. (3) 点击率转化率预测模块, 将预训练的广告嵌入与其他嵌入进行拼接, 经过专家网络, 门控网络和任务塔, 预测点击率和转化率. (4) 损失函数, 采用二元交叉熵 (binary cross entropy, BCE) 函数来约束点击率、转化率预测的结果误差. 考虑到点击率和转化率之间存在因果关系, 引入逆向倾向评分来对预测的转化率进行纠偏.

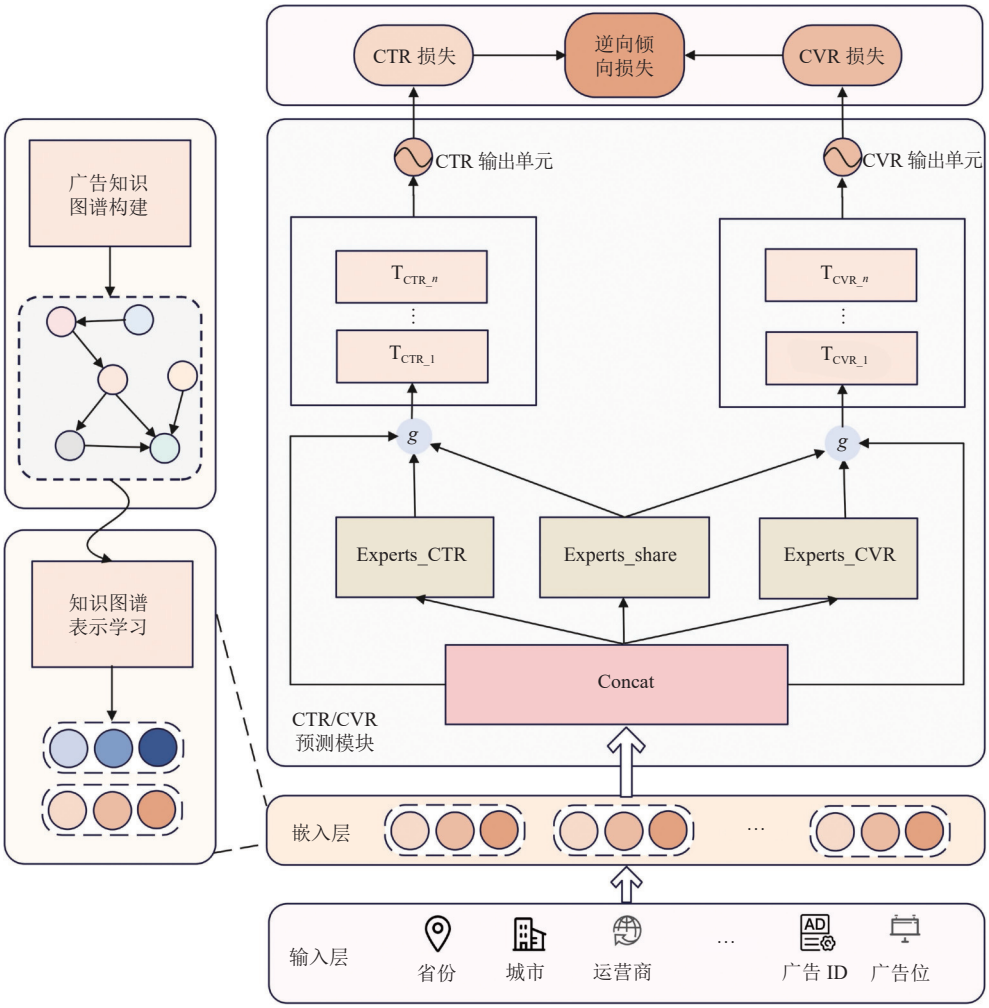

图 1 整体框架图

2.1 知识图谱模块

2.1.1 知识图谱构造模块

现有的点击率与转化率预测模型主要聚焦于用户信息与推荐给用户的广告信息, 而忽视了广告自身的语义信息及广告之间的关联性. 这种方法难以有效捕捉广告之间的相似性及用户对不同广告类型的偏好, 进而影响预测效果. 但是, 大多数广告数据存储在关系型数据库中, 尽管能够提供丰富的广告信息, 却难以捕捉广告之间的关联性.

知识图谱是一种图形化的数据表示模型, 用于描述实体之间的关系和属性. 它通常采用三元组形式进行存储, 它的格式为 (head, relation, tail), 其中 head 称为头实体, tail 称为尾实体, relation 表示头实体指向尾实体的关系. 这种结构有助于组织和表达复杂的知识关联, 使得信息更易于理解和推理.

为此, 本文针对关系型数据库存储数据, 设计了一种知识图谱构建方法, 能够将关系型数据库中的数据转化为三元组形式. 此方法不仅能整合广告信息, 还能捕捉广告之间的关联性, 从而更全面地表达广告的语义信息及相互关系.

广告特征信息的示例如表 1 所示, 其中 ad 为广告 ID, carrier 为广告运营商, sce 为广告场景可以投放的场景版位, app 是广告所投放的应用, app_json 为该广告所支持投放的应用. 为了构建广告知识图谱, 本文将广告的特征信息分为 3 类: 定长特征、变长特征和混合型特征. 在广告知识图谱的构建中, 这 3 类特征均以广告 ID 为头实体.

- 定长特征: 指具有固定数量值的特征. 为准确区分不同特征, 本文采用表头与表中数据结合的方式作为尾实体表示, 并将表头作为关系, 形成具体的三元组.

以表 1 中的 carrier 列为例, 根据这一构造规则, 模型可以生成如下形式的三元组: (ad_4479, carrier, carrier_3)、(ad_8034, carrier, carrier_4) 等.

表 1 广告特征示例表

ad	carrier	sce	app	app_json
4479	3	[3]	2	[4]
8034	4	[1,2]	3	[5,15]
8021	2	[1,2,4]	1	[]

● 变长特征: 是指其值可以为一个列表, 且列表长度不固定. 为了更精准地构建知识图谱, 本文将列表中的每个数据与表头拼接, 形成新实体, 并将表头视为关系. 以表 1 中的 sce 列为例, 根据该构造规则, 模型可以生成如下形式的三元组: (ad_4479, sce, sce_3)、(ad_8034, sce, sce_1)、(ad_8034, sce, sce_2) 等.

● 混合型特征: 是指两个特征列之间存在约束关系. 在处理这种特征时, 本文采用一种创新的方法将两列特征联合起来, 以确保知识图谱三元组的完整性和准确性. 将其中一个特征作为约束特征, 另一个作为被约束特征, 分别按照各自的构造规则处理. 然后, 将它们按照约束特征和被约束特征的顺序拼接成新实体, 并将约束特征的表头作为关系表示. 以表 1 的 app 和 app_json 列为例, 其中 app 为约束特征, app_json 为被约束特征. 依照上述构造规则, 模型可以生成如下形式的三元组: (ad_4479, app_json, app _2_app_json_4)、(ad_8034, app_json, app_3_ app_json_5) 等.

整个过程的伪代码如算法 1 所示.

算法 1. 广告特征三元组抽取算法

输入: 广告数据集 Data, 包括广告 ID 和其特征信息.
输出: 广告三元组集合 *S*.

1. 初始化空集合 *S* 用于存储三元组
2. 针对广告特征名称, 构造关系 *R*, 形式为"*F_name*"
3. 按以下规则将广告特征分为 3 类:
   ① 定长特征 *F_fixed*
   ② 不定长特征 *F_variable*
   ③ 混合型特征 *F_mixed*
4. For each *ad* In Data:
5. 获取广告 ID, 记为 *ad_id*
6. 构建头实体 *H*, 形式为"*ad_<ad_id>*"
7. 对于定长特征 *F_fixed*, 执行以下操作:
8. 定长特征 *F_name*, 特征值为 *V*
9. 构造尾实体 *T*, 形式为"*F_name<V>*"
10. 生成三元 (*H*, *R*, *T*), 并加入集合 *S*
11. 对于不定长特征 *F_variable*, 执行以下操作:
12. 遍历 *F_variable* 特征值列表 *V_f*
13. 对列表 *V_f* 中特征值 *V* 执行以下步骤:
14. 构造尾实体 *T*, 形式为"*F_name<V>*"
15. 生成三元组 (*H*, *R*, *T*), 并加入集合 *S*
16. 对于混合特征 *F_mixed*, 执行以下操作:
17. 遍历混合特征 (*F*1, *F*2); 获取 *F*1 的特征值 *V*1; 获取 *F*2 的特征值 *V*2
18. 遍历 *F*2 的特征值列表 *V_f*2, 执行以下操作:
19. 构造尾实体 *T*, 形式"*F*1_<*V*1>*F*2_<*V*2>"
20. 生成三元组 (*H*, *R*, *T*), 并加入集合 *S*
21. 将三元组集合 *S* 保存为文件.

通过这一系列创新性的构造规则, 成功构建了广告知识图谱, 全面而精确地表达了不同类型特征之间的关系, 为推荐算法提供了更有力的支持.

2.1.2 知识图谱表示学习模块

尽管通过知识图谱构建模块可以获得广告知识图谱, 然而该图谱仅以结构化形式存在, 需要进一步转化为嵌入表示形式才能直接应用于广告的分析和预测任务. 为此, 知识图谱表示学习模块致力于通过训练构建的知识图谱, 获取具有语义信息的广告嵌入. 这些嵌入不仅融合了广告的语义特征, 还捕捉了广告间的交互信息, 为后续的 CTR 和 CVR 预测提供更为丰富的语义特征. 知识图谱表示学习的目标是将实体、关系或其他对象映射到低维向量的向量空间, 以更准确地捕捉它们之间的语义关系.

通过知识图谱构建过程可知, 不同广告对于不同广告特征有着相同的特征值 (即 N-N 问题), 为了缓解这一问题, 参考 TransH 模型, 对于给定的广告三元组 $(T_{\text{ad}}, r, T_{\text{fe}})$, 引入基于关系超平面的得分函数可以帮助区分具有不同表示的广告, 并确保只有将它们投影到特定的超平面上才能识别它们的相似性. 以下是基于超平面的得分函数的定义:

$$f_r(T_{\text{ad}}, T_{\text{fe}}) = \left\| T_{\text{ad}}^{\perp} + r - T_{\text{fe}}^{\perp} \right\| \tag{1}$$

其中, $T_{\text{ad}}^{\perp}$ 和 $T_{\text{fe}}^{\perp}$ 是广告 ID 和广告特征的投影向量, 投影过程具体公式如下所示:

$$T_{\text{ad}}^{\perp} = T_{\text{ad}} - w_r^{\text{T}} T_{\text{ad}} w_r \tag{2}$$

$$T_{\text{fe}}^{\perp} = T_{\text{fe}} - w_r^{\text{T}} T_{\text{fe}} w_r \tag{3}$$

其中, w_r 是对应的关系的映射矩阵. 为了区分正负三元组, 使用了基于 Margin 的损失函数:

$$\mathcal{L} = \sum_{(T_{\text{ad}}, r, T_{\text{fe}}) \in \Delta} \sum_{\left(T'_{\text{ad}}, r', T'_{\text{fe}}\right) \in \Delta_{\left(T'_{\text{ad}}, T_{\text{fe}}\right)}} \left[f_r(T_{\text{ad}}, T_{\text{fe}}) + r - f_{r'}\left(T'_{\text{ad}}, T'_{\text{fe}}\right) \right] \tag{4}$$

为减少假负例的生成, 采用 $\frac{T_{\mathrm{fe}_p}T_{\mathrm{ad}}}{T_{\mathrm{fe}_p}T_{\mathrm{ad}}+T_{\mathrm{ad}_p}T_{\mathrm{fe}}}$ 概率对头实体进行替换. 其中, $T_{\mathrm{fe}_p}T_{\mathrm{ad}}$ 表示该特征下的广告 ID 所对应的平均广告特征数目, $T_{\mathrm{ad}_p}T_{\mathrm{fe}}$ 表示每个特征下的特征值所连接的平均广告数目. 经过不断迭代优化, 得到带有广告语义信息的广告 ID 的嵌入表示 T_{ad}.

2.2 点击率转化率预测模块

点击率转化率预测模块由任务感知专家网络、门控网络和任务塔组成. 点击率和转化率预测可以看作两个任务. 对于每个任务 i, 输入广告 ID 和其他特征, 广告 ID 的嵌入 T_{ad} 来自知识图谱表示学习模型, 其他特征的嵌入 T_i 来自随机初始化. 这些任务的特征参数是共享的, 将这些特征的嵌入拼接作为初始输入:

$$x=[T_{\mathrm{ad}};T_1;\cdots;T_m] \tag{5}$$

其中, ;表示拼接操作, m 表示其他特征的个数.

专家网络又被区分为任务独有的专家网络以及任务间共享的专家网络. 任务独有的专家网络关注特定任务的局部知识, 任务共享的专家网络考虑任务共享的全局特征. 由于任务的不同, 任务共享的专家网络和任务独享的专家网络针对不同的任务做出不同的贡献. 因此, 采用门控网络来实现不同的加权机制. 具体来说, 每个专家模块通过门控网络来调节其贡献. 专家网络的输出通过线性变化与 *Softmax* 函数确定任务独享专家和任务共享专家对特定任务的贡献.

$$w_i^{\mathrm{task}}=Softmax(W_i^{\mathrm{task}}x+b_i^{\mathrm{task}}) \tag{6}$$

其中, w_i^{task} 表示任务 i 对应的权重, W_i^{task} 和 b_i^{task} 分别表示线性变化的权重矩阵和偏置项. 任务共享专家网络权重 w_i^{share} 采用相同计算方式. 任务独享和任务共享的专家网络结果根据权重进行合并. 任务 i 的输出 o_i 表示为:

$$o_i=w_i^{\mathrm{task}}E_i(x)+w_i^{\mathrm{share}}E_s(x) \tag{7}$$

其中, $E_i(x)$ 表示任务 i 的任务独享专家网络, $E_s(x)$ 表示任务共享专家网络.

点击率预测和转化率预测都在专属的任务塔分支中进行. 每个塔分支由一个两层全连接神经网络构成, 并包含 *ReLU* 和 *Sigmoid* 层以实现非线性转化.

$$\hat{y}_i=Sigmoid(W_{i2}(ReLU(W_{i1}o_i))) \tag{8}$$

2.3 损失函数

采用二元交叉熵损失来度量每个任务预测结果的误差, 每个任务的损失求和作为推荐任务误差 L_{RS}:

$$L_{\mathrm{RS}}=\sum_{i=1}^{n}BCELoss(\hat{y}_i,y_i) \tag{9}$$

其中, $\hat{y}_i$ 表示任务 i 的预测结果, y_i 表示任务 i 的真实结果.

在实际情况下, 转化行为依赖于用户是否已经点击广告或推荐内容, 二者之间存在因果关系. 仅使用二元交叉熵作为损失函数, 没有考虑这种因果关系, 而是将点击和转化视为独立事件, 可能导致 CVR 预测结果有偏差. 为解决这个问题, 引入逆向倾向评分 (inverse propensity scoring, IPS) 来纠偏:

$$L_{\mathrm{IPS}}=\frac{1}{|D|}\sum_{j\in D}\frac{y_j^{\mathrm{ctr}}BCELoss\left(y_j^{\mathrm{cvr}},\hat{y}_j^{\mathrm{cvr}}\right)}{\hat{y}_j^{\mathrm{ctr}}} \tag{10}$$

其中, D 表示样本空间, $|D|$ 是其大小, y_j^{ctr} 表示第 j 个广告的点击情况, $\hat{y}_j^{\mathrm{ctr}}$ 是其预测的点击率, y_j^{cvr} 表示第 j 个广告的转化情况, $\hat{y}_j^{\mathrm{cvr}}$ 表示预测的转化率. IPS 考虑了每个样本的点击倾向, 从而在计算损失时给予不同样本不同的权重. 这种方法可以有效纠正因点击倾向不均匀而引入的偏差, 确保模型在预测转化率时更为准确. 最终的损失函数为:

$$L=L_{\mathrm{RS}}+\lambda L_{\mathrm{IPS}} \tag{11}$$

3 实验分析

本节将详细介绍实验验证分析, 首先介绍所需数据集和实验设置, 然后对模型进行实验验证和性能评估. 在真实广告数据集上, 本文模型在 CTR 和 CVR 预估任务优于基线模型.

3.1 数据集

实验是在真实广告数据集上进行, 该数据来源于公司真实数据, 为运营商流量包广告. 数据集包含广告属性 (广告 ID, 运营商, 流量包名称, 价格等) 和用户属性 (用户 ID, 所在城市, 年龄等) 以及对应的点击和转化情况. 经过脱敏处理以确保隐私安全, 数据集涵盖为期 30 天的时间段, 总计包含 19 962 455 条记录. 每天部分数据情况如表 2 所示. 为确保算法的训练和泛化效果, 我们将 2024 年 6 月 1 日–2024 年 6 月 24 日的数据作为训练集, 2024 年 6 月 25 日–2024 年 6 月 27 日的数据作为验证集, 2024 年 6 月 28 日–2024 年 6 月 30 日的数据为测试集. 由于现有的公开数据集缺乏可供构建知识图谱的广告和商品数据, 本研究未在公开数据集上进行实验, 而是专注于利用真实广告数据集验

证 KGEARA 模型的有效性.

表 2 数据统计表

数据集	数据量	点击数	转化数
训练集	16252934	10564921	882635
验证集	1825256	1403291	109652
测试集	1884265	1485859	104564

3.2 实验设置

本文的实验使用 PyTorch 框架实现. 真实广告数据集的批量为 1024, 嵌入层维度为 32, 学习率设置为 0.001. Dropout 率设置为 0.5. 使用 Ubuntu 20.04 操作系统, 编程语言为 Python 3.7, GPU 为 NVIDIA RTX 3090, 所有实验都在同一实验环境下进行.

3.3 评价指标

实验中, 本文选择 CTR 和 CVR 预测常用的 *AUC* 和 LogLoss 作为评估指标.

AUC 表示 ROC (receiver operating characteristic) 曲线下的面积, 用于衡量模型在不同阈值下真正例率和假正例率之间的性能, *AUC* 提供了对模型整体性能的综合评估. 计算公式如下所示:

$$AUC = \sum \frac{I(P_{正样本}, P_{负样本})}{M \times N} \tag{12}$$

$$I(P_{正样本}, P_{负样本}) = \begin{cases} 1, & P_{正样本} > P_{负样本} \\ 0, & P_{正样本} = P_{负样本} \\ 0, & P_{正样本} < P_{负样本} \end{cases} \tag{13}$$

其中, I表示统计函数, M 为正样本数, N 为负样本数.

LogLoss 即对数损失, 是二元分类中广泛使用的指标, 用于测量两个分布之间的距离. LogLoss 的下限为 0, 表示两个分布完美匹配, 值越小表示性能越好.

3.4 实验结果

本文提出的算法与其他基线模型的对比结果如表 3 所示, 通过分析实验结果可以得到以下结论.

表 3 实验结果对比

模型	CTR预测		CVR预测	
	AUC	LogLoss	*AUC*	LogLoss
LR[4]	0.8616	0.3496	0.8416	0.1661
SharedBottom[18]	0.8821	0.3387	0.8594	0.1588
MMoE[2]	0.9085	0.3353	0.8689	0.1574
PLE[33]	0.9098	0.3359	0.8644	0.1594
ESMM[3]	0.9151	0.3351	0.8728	0.1571
MRDR[34]	0.9196	0.3329	0.8804	0.1556
ESCM^2[19]	0.9214	0.3317	0.8869	0.1527
AECM[35]	0.9250	0.3304	0.8927	0.1519
Ours	**0.9403**	**0.3223**	**0.9134**	**0.1467**

本文提出的 KGEARA 在真实场景的广告数据集上表现优于所有基线模型. 在 *AUC* 指标上, 相较于最优基线分别取得了 1.6% 和 2.3% 的提升, 在 LogLoss 指标上分别取得了 2.6% 和 3.4% 的提升, 表明了其对点击率和转化率预测有更强的建模能力. 现有的模型仅考虑推送给用户的广告信息和用户信息, 忽略了广告自身的语义信息和广告之间的联系. 为此本文引入知识图谱的预训练嵌入, 将预训练向量与原始特征向量通过设计的多特征融合模块, 将它们联合作为模型输入, 实现了对广告语义信息的利用以及对用户-广告关系的建模, 使得模型有效地刻画了用户与广告之间的复杂关系. 本文提出的 KGEARA 模型优于其他基线, 这验证了本文提出的融合知识图谱的预训练嵌入能够融合广告的语义信息和广告间的交互信息.

3.5 消融实验

本实验旨在探究模型各组件对点击率 (CTR) 和转化率 (CVR) 预测性能的影响 通过去除或替换模型中的关键部分, 以评估这些部分对模型整体性能的贡献. 其中, –KG 表示去除知识图谱模块; –PS 表示去除参数共享, 让 CTR 和 CVR 预测的参数相互独立; –Loss 表示替换现有的损失, 改用 *BCELoss* 损失. 具体实验结果如表 4 所示.

表 4 消融实验结果

模型	CTR预测		CVR预测	
	AUC	LogLoss	*AUC*	LogLoss
–KG	0.9218	0.3348	0.8918	0.1553
–PS	0.9289	0.3315	0.8952	0.1519
–Loss	0.9311	0.3294	0.8994	0.1506
Ours	**0.9403**	**0.3223**	**0.9134**	**0.1467**

针对知识图谱模块, 首先尝试移除 KG 模块, 发现 CTR 预测的 *AUC* 值从完整模型的 0.9403 下降至 0.9218, 降低了 2%, LogLoss 从 0.3223 上升至 0.3348, 增加了 3.9%. 对于 CVR 预测, *AUC* 值从 0.9134 下降至 0.8918, 降低了 2.4%, LogLoss 从 0.1467 上升至 0.1553, 增加了 5.7%. 这表明知识图谱模块中附带的知识图谱信息对于提高 CTR 和 CVR 预测的准确性和降低损失具有显著作用. 而针对点击率转化率预测模块的融合策略, 在删除共享参数时, CTR 和 CVR 预测的 *AUC* 值相比完整模型分别下降了 1.2% 和 2.0%, LogLoss 分别增加了 2.6% 和 3.5%. 将损失函数替换为 *BCELoss* 后, CTR 和 CVR 预测的 *AUC* 值相比完整模型分别下降

了 1.0% 和 1.5%, LogLoss 分别增加了 2.2% 和 2.6%. 这表明引入逆向倾向评分能够缓解由于点击和转化存在因果关系而导致的转化率预测偏差.

3.6 知识图谱模块迁移实验

为了验证知识图谱模块在其他模型中的有效性, 在实验中将知识图谱模型迁移到 MMOE、ESMM 和 ESCM^2 模型上, 将广告嵌入同其他特征嵌入表示相融合. 本实验所用数据集使用与第 3.1 节相同的真实广告数据集, 实验结果如表 5 所示.

表 5 KG 模块迁移实验结果

模型	CTR_*AUC*	CVR_*AUC*
MMoE	0.9085	0.8689
MMoE+KG	0.9174	0.8804
ESMM	0.9151	0.8728
ESMM+KG	0.9262	0.8843
ESCM^2	0.9214	0.8869
ESCM^2+KG	**0.9317**	**0.8982**

实验结果表明, 知识图谱模块能够有效地迁移到 MMoE、ESMM 和 ESCM^2 模型上, 并提升模型的点击率和转化率预测性能. 在 MMoE 模型中, 引入知识图谱模块后 CTR 和 CVR 预测的 *AUC* 值分别提升了 1.0% 和 1.3%. 在 ESMM 模型中, 引入知识图谱模块后 CTR 和 CVR 预测的 *AUC* 值分别提升了 1.2% 和 1.3%. 在 ESCM^2 模型中, 引入知识图谱模块后, CTR 和 CVR 预测的 *AUC* 值分别提升了 1.1% 和 1.3%. 这表明知识图谱模块能够有效提升 MMoE、ESMM 和 ESCM^2 的预测性能, 这主要是因为知识图谱模块能够捕捉广告间的语义关联性和相似性, 从而更好地理解用户对不同类型广告的偏好.

3.7 参数实验

嵌入维度对模型的性能有着直接影响. 本节测试了 4、8、16、32 及 64 这几个嵌入维度的设置. 实验结果如图 2 所示, 较低的 4 和 8 嵌入维度可能因为特征表示的维度过小, 未能有效表达特征, 从而影响模型的性能. 随着嵌入维度的增加, 模型具有更强的建模能力. 然而, 当嵌入维度过高时 (如 64), 可能会引入原始特征中的噪声, 导致模型过拟合. 实验结果表明, 将嵌入层维度设置为 32 时可获得最佳的预测性能.

3.8 灰度测试

我们进行了线上灰度测试, 进一步展示了本文模型相对于原基线的优势. 具体来说我们使用 PyTorch 深度学习框架对模型进行实现. 广告有多种的场景版位, 我们通过灰度测试的结果评估模型在不同场景版位的性能, 评估指标是单用户价值、单点击价值及相对于原基线的广告消耗差值.

- 短信失败: 指用户在办理业务时因未接收到短信等原因导致办理失败. 我们在该场景中部署了 20 天, 覆盖了大约 3.4 万的独立访客和 4.7 万的页面浏览. 总体而言 KGEARA 模型的点击率提高了 7%, 转化率提高了 6.6%, 单用户价值提高了 4.3%, 单点击价值提高了 3.4%, 整体的广告消耗提高了 7.9%.
- 办理失败: 指用户在办理业务时业务不支持该用户办理等原因导致办理失败. 我们在该场景中部署了 20 天, 覆盖了大约 4 万的独立访客和 4.8 万的页面浏览. 总体而言 KGEARA 模型的点击率提高了 4.6%, 转化率提高了 5.2%, 单用户价值提高了 13.7%, 单点击价值提高了 13.7%, 整体的广告消耗提高了 15.5%.

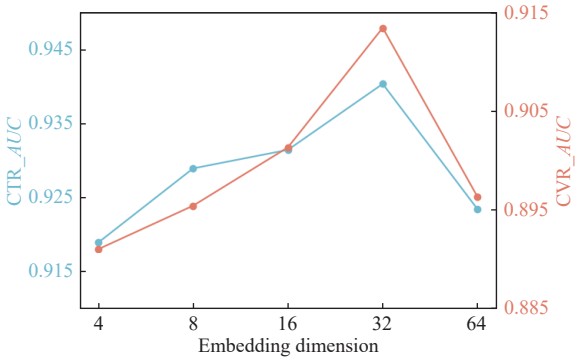

图 2 嵌入维度实验结果

4 结论与展望

本文提出了一个知识图谱增强的广告推荐算法 (KGEARA). 实验结果表明, 该模型提出了一种从关系型数据库构建知识图谱的算法, 通过将结构化数据转换为三元组形式, 旨在整合广告特征并捕捉广告之间的关系. 在知识图谱表示学习模块, 利用 TransH 模型, 将这些特征转化为嵌入表示, 这些嵌入表示融合广告的语义特征并有效地捕获广告之间的交互信息. 将这些嵌入与其他特征嵌入结合, 通过专家网络、门控网络和任务塔来预测点击率和转化率. 同时, 引入逆向倾向评分来处理点击倾向不均的问题, 从而修正预测中的偏差. 与其他模型相比, KGEARA 模型在两种转化率指标的预测效果上都获得显著提升. 本研究证明了知识图谱表示学习与深度学习相结合在广告领域的有

效性, 为知识驱动的推荐系统方法提供了新的思路, 也为提升数字广告的转化效果带来新的可能.

参考文献

1 Jefferson M, 骆佳. “危机”解除, 数字广告引领乐观未来. 国际品牌观察, 2022(4): 34–36.

2 Ma J, Zhao Z, Yi X, *et al*. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London: ACM, 2018. 1930–1939.

3 Ma X, Zhao LQ, Huang G, *et al*. Entire space multi-task model: An effective approach for estimating post-click conversion rate. Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. Ann Arbor: ACM, 2018. 1137–1140.

4 Kumar R, Naik SM, Naik VD, *et al*. Predicting clicks: CTR estimation of advertisements using logistic regression classifier. Proceedings of the 2015 IEEE International Advance Computing Conference (IACC). Banglore: IEEE, 2015. 1134–1138.

5 Rendle, S. Factorization machines. Proceedings of the 2010 IEEE International Conference on Data Mining (ICDM’10). Sydney: IEEE, 2010. 995–1000.

6 Juan Y, Lefortier D, Chapelle O. Field-aware factorization machines in a real-world online advertising system. Proceedings of the 26th International Conference on World Wide Web Companion. Perth: IW3C2, 2017. 680–688.

7 Guo HF, Tang RM, Ye YM, *et al*. DeepFM: A factorization-machine based neural network for CTR prediction. Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne: AAAI, 2017. 1725–1731.

8 Wang RX, Fu B, Fu G, *et al*. Deep & cross network for ad click predictions. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Halifax: ACM, 2017. 12.

9 Qu YR, Cai H, Ren K, *et al*. Product-based neural networks for user response prediction. Proceedings of the 16th IEEE International Conference on Data Mining (ICDM). Barcelona: IEEE, 2016. 1149–1154.

10 Lian JX, Zhou XH, Zhang FZ, *et al*. xDeepFM: Combining explicit and implicit feature interactions for recommender systems. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London: ACM, 2018. 1754–1763.

11 Song WP, Shi CC, Xiao ZP, *et al*. AutoInt: Automatic feature interaction learning via self-attentive neural networks. Proceedings of the 28th ACM International Conference on Information and Knowledge Management. Beijing: ACM, 2019. 1161–1170.

12 Zhou GR, Zhu XQ, Song CR, *et al*. Deep interest network for click-through rate prediction. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London: ACM, 2018. 1059–1068.

13 Zhou GR, Mou N, Fan Y, *et al*. Deep interest evolution network for click-through rate prediction. Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu: AAAI Press, 2019. 5941–5948.

14 Feng YF, Lv FY, Shen WC, *et al*. Deep session interest network for click-through rate prediction. Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao: IJCAI, 2019. 2301–2307.

15 Xie Y, Jiang D, Wang XM, *et al*. Robust transfer integrated locally kernel embedding for click-through rate prediction. Information Sciences, 2019, 491: 190–203. [doi: 10.1016/j.ins.2019.04.006]

16 Li X, Chen SW, Dong J, *et al*. Decision-making context interaction network for click-through rate prediction. Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington: AAAI Press, 2023. 5195–5202.

17 Lyu FY, Tang X, Liu DG, *et al*. Optimizing feature set for click-through rate prediction. Proceedings of the 2023 ACM Web Conference. Austin: ACM, 2023. 3386–3395.

18 Caruana R. Multitask learning: A knowledge-based source of inductive bias1. Proceedings of the 10th International Conference on Machine Learning. Amherst: Morgan Kaufmann Publishers Inc., 1993. 41–48.

19 Wang H, Chang TW, Liu TQ, *et al*. ESCM^2: Entire space counterfactual multi-task model for post-click conversion rate estimation. Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. Madrid: ACM, 2022. 363–372.

20 Wang YB, Xue YB, Liu B, *et al*. Click-conversion multi-task model with position bias mitigation for sponsored search in eCommerce. Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. Taipei: ACM, 2023. 1884–1888.

21 Bordes A, Usunier N, Garcia-Duran A, *et al*. Translating embeddings for modeling multi-relational data. Proceedings of the 27th International Conference on Neural Information Processing Systems. Lake Tahoe: Curran Associates Inc., 2013. 2787–2795.

22 Wang Z, Zhang JW, Feng JL, *et al*. Knowledge graph embedding by translating on hyperplanes. Proceedings of the 28th AAAI Conference on Artificial Intelligence. Québec City: AAAI Press, 2014. 1112–1119.

23 Lin YK, Liu ZY, Sun MS, *et al*. Learning entity and relation embeddings for knowledge graph completion. Proceedings of the 29th AAAI Conference on Artificial Intelligence. Austin: AAAI Press, 2015. 2181–2187.

24 Nickel M, Tresp V, Kriegel HP. A three-way model for collective learning on multi-relational data. Proceedings of the 28th International Conference on Machine Learning. Bellevue: Omnipress, 2011. 809–816.

25 Yang BS, Yih WT, He XD, *et al*. Embedding entities and relations for learning and inference in knowledge bases. arXiv:1412.6575, 2015.

26 Trouillon T, Welbl J, Riedel S, *et al*. Complex embeddings for simple link prediction. Proceedings of the 33rd International Conference on Machine Learning. New York: JMLR.org, 2016. 2071–2080.

27 Huang J, Zhao WX, Dou HJ, *et al*. Improving sequential recommendation with knowledge-enhanced memory networks. Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. Ann Arbor: ACM, 2018. 505–514.

28 Wang HW, Zhang FZ, Wang JL, *et al*. RippleNet: Propagating user preferences on the knowledge graph for recommender systems. Proceedings of the 27th ACM International Conference on Information and Knowledge Management. Torino: ACM, 2018. 417–426.

29 Hu BB, Shi C, Zhao WX, *et al*. Leveraging meta-path based context for top-N recommendation with a neural co-attention model. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London: ACM, 2018. 1531–1540.

30 He XN, Liao LZ, Zhang HW, *et al*. Neural collaborative filtering. arXiv:1708.05031v2, 2017.

31 李灏然, 张超群, 汤卫东, 等. 基于元学习与 DeepFM 的知识图谱特征增强推荐模型. 信息与电脑 (理论版), 2024, 36(20): 51–53.

32 熊晓波, 方文涛. 基于知识增强的细粒度个性化新闻推荐用户建模. 计算技术与自动化, 2024, 43(4): 161–166. [doi: 10.16339/j.cnki.jsjsyzdh.202404026]

33 Tang H, Liu J, Zhao M, *et al*. Progressive layered extraction (PLE): A novel multi-task learning (MTL) model for personalized recommendations. Proceedings of the 14th ACM Conference on Recommender Systems. ACM, 2020. 269–278.

34 Guo S, Zou L, Liu Y, *et al*. Enhanced doubly robust learning for debiasing post-click conversion rate estimation. Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2021. 275–284.

35 Zhang X, Huang C, Zheng K, *et al*. Adversarial-enhanced causal multi-task framework for debiasing post-click conversion rate estimation. Proceedings of the 2024 ACM Web Conference. OpenReview.net, 2024. 3287–3296.

(校对责编: 王欣欣)