基于深度学习的推荐算法研究综述

叶卫根

（赣东学院信息工程学院，江西 抚州 344100）

摘 要 随着深度学习在人工智能各领域取得的突破性成果，其在个性化推荐领域的应用也吸引了学界的广泛关注，并成为研究热点。本文概述了传统推荐系统分类，讨论了深度学习与协同过滤的融合（如基于受限玻尔兹曼机、自编码器的协同过滤及神经矩阵分解模型等）、深度学习与特征交互的推荐模型（如基于注意力网络的因子分解机、乘积神经网络、Wide & Deep模型等），并分析了推荐系统未来的研究趋势，旨在为构建更高效、公平的智能推荐系统提供理论参考。

关键词 个性化推荐；推荐系统；深度学习；特征交互；协同过滤

基金项目：江西省教育省厅科学技术青年项目（项目编号：GJJ2203707）；赣东学院院长基金（项目编号：YZJJ202206）。

中图分类号：TP181；TP391 **文献标志码**：A **DOI**：10.3969/j.issn.2097-3365.2025.30.001

0 引言

推荐系统已成为互联网企业的核心工具，广泛应用于电商、社交媒体、在线教育等领域，既能提升客户满意度，又能拉动营收增长。相较传统用户主动搜索，个性化推荐可过滤低关联信息、减少筛选成本，还能突破“用户需明确表达需求”限制，其能依据购买记录、点击频率等行为轨迹，从海量信息中挖掘契合需求的商品或内容，既缓解用户决策疲劳，又助力企业激活长尾商品曝光。随着深度学习在多领域取得成就，其在个性化推荐领域的应用成为研究热点。深度学习可捕捉用户与物品的非线性复杂关联，还能从上下文、文本、社交网络等数据源中挖掘深层关联。

1 推荐系统概述

1.1 基于内容的推荐

基于内容的推荐系统通过解析物品与用户的内容特征，挖掘并推荐符合用户兴趣偏好的物品。其推荐逻辑可拆解为三个核心步骤。一是提取物品属性标签、文本描述等特征，勾勒“本质属性”以明确物品“是什么”；二是分析用户个人资料、兴趣标签等，捕捉“偏好倾向”构建用户画像，理解用户“喜欢什么”；三是匹配两者属性，量化潜在兴趣后推荐物品。

1.2 基于协同过滤的推荐

基于协同过滤的推荐技术是推荐系统经典方案，核心依托群体智能，通过分析用户—物品历史交互记录（评分、点击等）挖掘相似性以实现推荐，成熟方法包括最近邻居、矩阵分解及深度学习三类。其中，矩阵分解法在Netflix Prize竞赛中表现突出，其对高维稀疏的用户—物品评分矩阵降维，得到低维用户、物品特征矩阵，分别表征用户偏好与物品属性。深度学习推荐法本质仍靠群体智能，不少早期模型是传统协同过滤的神经网络升级，如NeuMF模型[1]，将经典矩阵分解的向量点积评分计算，替换为神经网络多层感知机计算。

1.3 混合推荐

传统的混合推荐系统其基本思路是将基于内容的方法与基于协同过滤的方法进行融合，以得到更精准的推荐效果。混合推荐系统通常会通过加权式、切换式和特征增强式等方法将多种推荐技术进行综合。当然，另一种研究思路将混合定义为两种或多种深度神经网络（DNNs）的融合。

2 基于深度学习的推荐

2.1 基于受限玻尔兹曼机的协同过滤

受限玻尔兹曼机（RBM）是无向浅层神经网络，由可见层、隐藏层及模型参数构成，神经元仅跨层连接、同层无连接。Ruslan Salakhutdinov、Andriy Mnih与Geoffrey Hinton首次将其应用于推荐领域[2]。

该模型中神经元可以跨层相互连接，但同一层内的神经元之间不允许有连接可见层为输入层，其输入的是单个用户已观测到的数据，输入层中每个节点为独热编码结构表示物品，其由用户对该物品的评分转换得到。可见层为输入层，输入单个用户已观测数据，

节点以独热编码表示物品，未评分物品为缺失值，不参与训练。RBM无传统输出层，需通过“输入用户评分—模型训练—生成推荐”三步实现预测：输入时将评分整理为二元形式（如0表不喜欢、1表喜欢，未评分用-1表示），确定可见层神经元状态；训练用对比散度算法，正向计算隐藏层激活概率、反向重建可见层数据，调整权重以最小化重构误差，让模型通过隐藏层学习潜在特征；生成推荐时，依据隐藏层激活状态反向计算可见层未评分项的激活概率，概率高的即为推荐结果。

何登平等人指出上述模型只能处理离散值，为此提出了IRC-RBM模型[3]。该模型将可见层单元改为实值表示，直接输入用户对项目的原始评分，这简化了数据预处理流程，减少了信息损失，提升了对原始评分数据的建模能力。

2.2 基于自编码器的协同过滤

Suvash Sedhain等人[4]结合自编码器(AutoEncoder,简称AE)提出了基于自编码器的协同过滤AutoRec算法。基于用户的AutoRec模型，其主要由输入层、隐藏层与输出层构成。其中，最下面的一层为输入层，表示一个用户对所有物品的评分：r(i)=(Ri1,Ri2,,Ri3, ...,Rim)；类似的可以将输入层替换为所有用户对某一个物品的评分，这样便得到了基于项目的AutoRec模型。中间一层为隐藏层，且只有一层隐藏层，输入层的高维向量r(i)经过隐藏层被压缩为一个低维向量，即信息的编码过程，该向量可以理解为用户在隐藏状态空间的兴趣偏好表示。最上面一层为重构后的评分向量，用于预测目标用户对项目的评分，即信息的解码过程。由于从输入层到隐藏层与从隐藏层到输出层的全连接操作，都可通过对应的非线性激活函数进行编码与解码，该模型能更好地完成非线性建模，捕获复杂的用户项目交互特征。

虹霞黄等人基于AutoRec模型实现就业推荐，通过AutoRec模型对用户—岗位特征数据进行低维特征提取与重构[5]。

2.3 神经矩阵分解

传统的基于矩阵分解的协同过滤算法，在建模用户与物品的交互关系时，仅通过用户隐向量与物品隐向量的点积运算来捕获二者关联，这种方式对交互模式的刻画较为简单。为此，何向阳等人[6]基于深度学习框架提出了一种创新的矩阵分解范式，即神经矩阵分解（NeuMF）。

该模型突破了传统点积操作的局限，通过引入多层感知机来提升非线性建模与泛化能力，其由左右两个分支构成：左边的分支是广义矩阵分解（GMF）层，它是对传统矩阵分解的扩展，通过用户隐向量与物品隐向量的逐元素乘积捕捉线性交互；右边的分支是多层感知机（MLP）层，用于建模用户与物品间的非线性交互。GMF层与MLP层的输出向量会进行拼接，随后通过全连接层和sigmoid激活函数，最终输出目标用户u对物品i产生交互的概率。

NeuMF模型融合了矩阵分解的线性建模与神经网络的非线性建模能力，适用于处理隐式反馈数据（如点击、购买、收藏等）。这类数据能在一定程度上反映用户对物品的偏好倾向，因此模型在训练时采用交叉熵损失函数，以预测用户对物品产生交互行为的概率。需要特别强调的是，在NeuMF模型中，输入的用户与物品独热编码会先通过嵌入层转换为低维稠密的隐向量；而GMF层和MLP层并不共享这些用户与物品的隐向量。这样的处理能保证模型的灵活度：由于GMF层为线性操作，MLP层为非线性操作，两者对应的用户与物品嵌入向量的维度通常不一致，独立参数可更好地适配各自的建模需求。Liu等人[7]受NeuMF模型启发提出了一种基于显式—隐式反馈的神经矩阵分解算法，该算法为用户—物品矩阵学习线性和非线性的显式—隐式反馈特征，取得了较多的推荐效果。

3 基于深度学习处理特征交互

3.1 基于注意力网络的因子分解机

Jun Xiao等人指出，传统因子分解机（FM）模型虽能覆盖所有二阶特征组合，却无法区分不同特征交互组合的权重，限制了模型表达与泛化能力。为此，他们提出基于注意力网络的因子分解机（AFM）[8]。AFM模型含稀疏输入、嵌入表示层、二阶特征交互层、基于注意力的池化层、预测值输出五层。其中，二阶特征交互层每个节点输出k维向量，借鉴FM思想但做改进：FM用两特征嵌入向量内积建模交互，此层改为元素积，保留交互信息维度特征，为后续区分重要性提供细粒度基础。

基于注意力的池化层，针对二阶特征交互层输出的k维交互向量，通过注意力网络学习各向量权重系数，再聚合加权后的交互向量。该过程既动态区分特征交互重要性，又压缩高维信息为综合特征，有效提升模型对关键交互模式的捕捉能力。崔少国等人[9]受AFM模型启法，提出多注意力机制融合低高阶特征的神经推荐算法（DeepNRM），该算法通过因子分解机（FM）提取低阶组合特征、多层前馈神经网络提取高阶组合特征，再借助注意力网络和多头自注意力机制筛选关键特征并按重要性融合。

3.2 基于乘积的神经网络

Yanru Qu等人[10]针对多字段分类数据的用户响应预测任务，为更高效地建模特征间交互关系、同时降低多层感知机（MLP）处理高维稀疏数据时的计算成本，提出了基于乘积的神经网络（PNN）。PNN包含嵌入层、乘积层、全连接隐藏层及输出层，形成从稀疏特征降维、交互捕获到高阶模式挖掘的完整链路。

这里着重介绍一下二阶特征交互层，其由Z和P部分组成，其中Z由特征嵌入层所有特征向量构成，Z与相应的权重矩阵相乘后得到线性信号lz；P由所有特征向量两两乘积（可为内积或外积）操作的结果组成，P与相应的权重矩阵相乘后得到二次信号lp；最终生成的lz、lp与偏置向量b1将作为输入传到第一层的全连接隐藏层。该模型的核心创新在于，在特征嵌入层与MLP层之间增设了二阶特征交互层（乘积层），通过内积或外积运算直接捕获跨字段特征的成对交互信息，避免了传统MLP仅依赖“加法”操作难以适配分类数据交互逻辑的问题。

3.3 Wide & Deep模型

在推荐系统中，模型需同时兼顾记忆性与泛化性，才能实现精准且有多样性的推荐。为此Google公司[11]提出了经典的双塔模型Wide & Deep，其由左侧的宽度模块与右侧的深度模块构成。Wide & Deep模型中左侧的宽度模块负责记忆性，通过线性模型处理输入的原始特征和若干交叉特征；右侧的深度模块负责泛化，其输入层为各特征域构成的独热编码，然后通过特征嵌入查找得到对应的特征隐向量，特征隐向量会输入上层的深度神经网络模块完成特征的交互学习；最后宽度模块与深度模块的输出会拼接在一起进行联合学习。

Ruoxi Wang等人[12]指出Wide & Deep模型的宽度模块部分依赖人工设计交叉特征，需要先验知识和大量人力且难以实现高阶特征交叉，为此提出了DCN V2模型，该模型将Deep & cross交叉网络的权重向量升级为2D全矩阵，实现了对Deep & cross的多维度优化，能更好地适应Web推荐任务。

4 结束语

本文梳理个性化推荐系统技术演进与核心模型，首先从传统推荐系统分类切入，概括了基于内容、协同过滤及混合推荐的核心原理与差异；然后着重介绍深度学习与协同过滤融合的经典算法，深入剖析了这些算法如何突破传统协同过滤浅层线性局限，捕捉用户—物品复杂非线性关联；接着梳理了基于深度学习处理特征交互完成推荐的经典模型，详细地介绍了如何解决传统精排人工设计特征成本高、低阶交互建模不足的问题。

当前推荐系统虽进入“多技术融合”阶段，但仍面临冷启动、模型可解释性与准确性平衡等挑战。未来研究将拓展“技术深度”与“应用广度”，如大语言模型融合实现“意图驱动”升级，联邦推荐、多模态推荐等推动系统向更智能、适配复杂场景发展，最终实现兼具多维度优势的个性化服务。

参考文献：

[1] He X, Liao L, Zhang H, et al. Neural collaborative filtering[C]//Proceedings of the 26th international conference on world wide web, 2017.

[2] Salakhutdinov R, Mnih A, Hinton G. Restricted Boltzmann machines for collaborative filtering[C]//Proceedings of the 24th international conference on Machine learning, 2007.

[3] 何登平,张为易,黄浩.基于多源信息聚类和IRC-RBM的混合推荐算法[J].计算机工程与科学,2020,42(06):1089-1095.

[4] 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, 2015.

[5] 虹霞黄,军曾,历曾.多算法融合的就业推荐算法研究[J].教育理论与研究,2025,03(20):31-34.

[6] 同[1].

[7] Liu H, Wang W, Zhang Y, et al. Neural matrix factorization recommendation for user preference prediction based on explicit and implicit feedback[J]. Computational Intelligence and Neuroscience, 2022, 2022(01): 9593957.

[8] Xiao J, Ye H, He X, et al. Attentional factorization machines: Learning the weight of feature interactions via attention networks[J]. arXiv preprint arXiv:1708.04617, 2017.

[9] 崔少国,独潇,杨泽田.多注意力机制融合低高阶特征的神经推荐算法[J].计算机工程与应用,2023,59(08):192-199.

[10] Qu Y, Cai H, Ren K, et al. Product-based neural networks for user response prediction[C]//2016 IEEE 16th international conference on data mining (ICDM). IEEE, 2016.

[11] Cheng, Heng-Tze, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson et al. Wide & deep learning for recommender systems[C]//In Proceedings of the 1st workshop on deep learning for recommender systems, 2016.

[12] Wang R, Shivanna R, Cheng D, et al. Dcn v2: Improved deep & cross network and practical lessons for web-scale learning to rank systems[C]//Proceedings of the web conference 2021, 2021.