arXiv:2302.02241v1 [cs.IR] 4 Feb 2023

点击率预测的特征表示学习：综述与新视角

Fuyuan Lyu[1] , **Xing Tang**[2] , **Dugang Liu**[3] , **Haolun Wu**[1,4] ,**Chen Ma**[5] , **Xiuqiang He**[2] 和 **Xue Liu**[1]1麦吉尔大学计算机科学学院2腾讯 FiT3广东省人工智能与数字经济实验室（深圳）4Mila—魁北克人工智能研究所5香港城市大学计算机科学系{fuyuan.lyu, haolun.wu}@mail.mcgill.ca, {shawntang,xiuqianghe}@tencent.com, dugang.ldg@gmail.com, chenma@cityu.edu.hk,xueliu@cs.mcgill.ca

摘要

表示学习一直是机器学习中的重要课题。在点击率预测中，大多数特征都表示为嵌入向量，并与模型中的其他参数同时学习。随着点击率预测模型的发展，特征表示学习已成为热门研究主题，近年来受到工业界和学术界研究人员的广泛关注。本综述旨在从更宏观的视角总结特征表示学习，并为未来研究铺平道路。为实现这一目标，我们首先围绕两个主要问题，对当前特征表示学习研究方法进行分类：（i）表示哪些特征；（ii）如何表示这些特征。随后，我们针对这两个问题详细介绍各类方法。最后，本文讨论该领域未来的研究方向。

1 引言

机器学习的成功在很大程度上取决于数据表示 [Bengio 等人，2013]。作为机器学习在现实世界中的一种应用，点击率（CTR）预测是各种个性化在线服务的核心功能模块，包括在线广告、推荐系统和网络搜索等。 [Zhang 等人， 2021]。

一般来说，点击率预测模型的典型输入由许多类别特征组成。我们将这些类别特征的取值称为特征值，并将其组织为特征域。例如，性别 这一特征域包含三个特征值：男性、女性 和未知。这些预测模型使用嵌入表将类别特征值映射为实值稠密向量。随后，这些嵌入被输入特征交互层，例如因子分解机 [Rendle、 2010], 交叉网络 [Wang等， 2017], 以及挤压激励层 [Huang 等，2019]。最终的分类器聚合表示向量并做出预测。整个流程如图 1 所示。因此，这些点击率预测模型的性能在很大程度上取决于其所采用的特征表示 [Bengio等， 2013]。

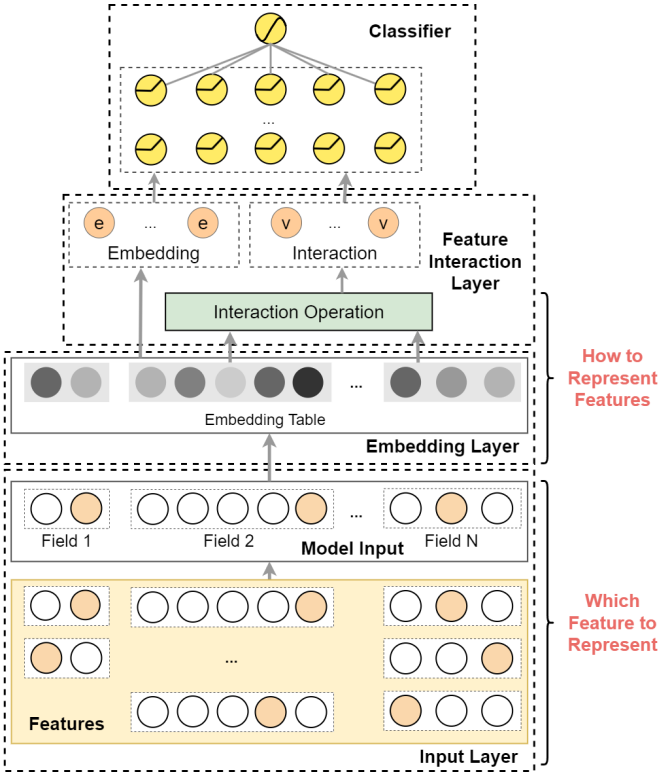


图 1：点击率预测通用框架概览。

自因子分解机 [Rendle, 2010], 引入潜在向量来建模特征以来，如何为 CTR 学习特征表示一直受到研究。然而，近期学术界和

工业界仍有许多工作关注 CTR 特征表示学习中的特定问题 [Wang 等人，2022；Lyu 等人，2022b；Liu 等人，2020b]。此外，近期也出现了一些相关综述。 [Zhang 等人， 2021] 回顾了深度 CTR 模型的发展，而 [Zheng 等人，2023] 则着重介绍自动化机器学习在推荐系统中的应用。本文首次围绕两个重要问题总结点击率预测的表示学习：i）特征表示学习应表示哪些特征；ii）应如何表示这些特征。因此，我们的讨论为这一问题提供了思路，并为 CTR 领域的特征表示带来了一些新视角。

基于上述两个问题，本文其余部分安排如下：首先，我们在第 2 节中针对这两个问题，对已有研究进行粗略的分类。第 3 节给出了特征表示学习的统一形式化。第 4 节和第 5 节分别详细介绍了同时解决这两个问题的已有研究。最后，本文在第 6 节总结了特征表示的若干潜在研究方向。

2 分类体系

我们根据在点击率预测任务中需要表示哪些特征，以及如何表示这些特征，对表示学习进行区分。图 2 概述了特征表示学习的分类体系，以及针对这两个问题所采用的方法。

考虑第一个问题：**表示哪些特征**，一些研究关注一阶特征（即原始特征）的选择，因为特征工程已被证明对准确预测至关重要。这些方法还可以根据选择的粒度进一步划分：领域或特征。除选择一阶特征外，其他研究者则关注二阶或高阶特征的选择。然而，由于搜索空间存在组合爆炸，针对二阶或高阶特征设计的方法仅限于领域级选择方法。

CTR 特征表示学习中的另一个重要问题是**如何表示这些选定的特征**。对于一阶特征，主要差异之一在于嵌入维度的选择。较高的维度可能导致表示过拟合，而较低的维度则可能导致欠拟合。关于一阶特征的另一个重要议题是特征哈希，它将多个特征映射到同一个表示中。特征哈希的设计实际上与对硬件要求较高的大规模工业模型有关。对于高阶特征，主要工作在于如何生成其表示。例如，大多数研究 [Khawar 等，2020] 将特征表示为聚合函数的输出，该函数以相关的一阶特征表示作为输入。因此，这类聚合函数的选择成为一个重要组成部分。

不同的特征表示学习方法也采用了各种方法来解决这两个因素。这些方法主要可分为多类：基于强化学习的方法 [Luo 等等， 2018], 基于统计的方法 [Wold 等，1987年；Tibshirani, 1996], 边预测、基于梯度的神经架构搜索（NAS） [Liu 等， 2019], 网络剪枝 [Kusupati等， 2020]以及基于进化搜索的神经架构搜索。

3 问题表述

在我们的设定中，我们将所有可能的一阶特征表示为$\mathbf{X}^{(1)} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m\}$。这里$\mathbf{x}_i$是独热编码表示，通常具有稀疏且高维的特点。而每个领域$\mathbf{z}_i$ 包含所有可能一阶特征中的一部分，记为:

$$\mathbf{z}_i = \{\mathbf{x}_{i_p}\},\ 1 \le i_p \le m, \tag{1}$$

例如，性别 领域 $\mathbf{z}_1$ 包含三个特征 $\mathbf{z}_{1_1}$ =男性、$\mathbf{z}_{1_2}$ = 女性 和 $\mathbf{z}_{1_3}$ = 其他，它们分别经过独热编码，编码为 [1, 0, 0], [0, 1, 0] 和 [0, 0, 1] 。在本综述中，我们用 m 表示领域的数量，用 n表示一阶特征的数量。

原始特征共存形成的组合称为高阶特征（也称为特征交互）。具体而言，$\mathbf{X}^{(2)} = \{\langle \mathbf{x}_{i_p}、\mathbf{x}_{j_q} \rangle\}$ 被视为二阶特征；$\mathbf{X}^{(3)} = \{\langle \mathbf{x}_{i_p}、\mathbf{x}_{j_q}、\mathbf{x}_{k_r} \rangle\}$ 被视为三阶特征。

正如我们在第 1 节中所述，关于点击率特征表示学习，有两个关键问题：(i) 应该表示哪些特征；(ii) 应该如何表示所选特征。我们将在第 3.1 节和第 3.2 节中分别对这两个问题进行详细阐述。最后，我们将在第 3.3 节中给出整体学习准则。

3.1 哪些特征的表述

关于应表述哪些特征的首要问题，可以视为学习一个用于t阶特征的二元张量$\mathcal{G}^{(t)} \in \{0,1\}^{C_n^t}$ 。如果选择了某些特征进行表述，则张量 $\mathcal{G}_i^{(t)}$ 中的对应值为 1。相反，如果不表述某些特征，则张量 $\mathcal{G}_i^{(t)}$ 中的对应值为 0。将所有这些二元张量拼接起来，以表示各阶特征的选择情况:

$$\mathcal{G} = [\mathcal{G}^{(1)}, \mathcal{G}^{(2)}, \dots], \tag{2}$$

该结果将作为点击率预测的输入。

然而，这种二元张量 $\mathcal{G}^{(t)}$ 的任意 $t \ge 2$ 都极其庞大。为了高效地探索如此广阔的空间，以往方法将选择粒度缩小到字段级别，这可以表示为学习一个二元张量 $\hat{\mathcal{G}}^{(t)} \in \{0,1\}^{C_{m}^t}$。

3.2 如何进行表述

至于如何表述所选特征，可以将其表述为学习一个变换函数 $f^{(t)}()$，用于t阶特征。

对于一阶特征，其中 $t = 1$，将其转换为稠密向量并通过嵌入层进行处理已成为一种常见做法 [Rendle，2010；Guo 等， 2017] 。这种转换可表述如下:

$$\mathbf{e}_{i_p} = f^{(1)}(\mathbf{x}_{i_p}) = \mathbf{E} \times \mathbf{x}_{i_p}. \tag{3}$$

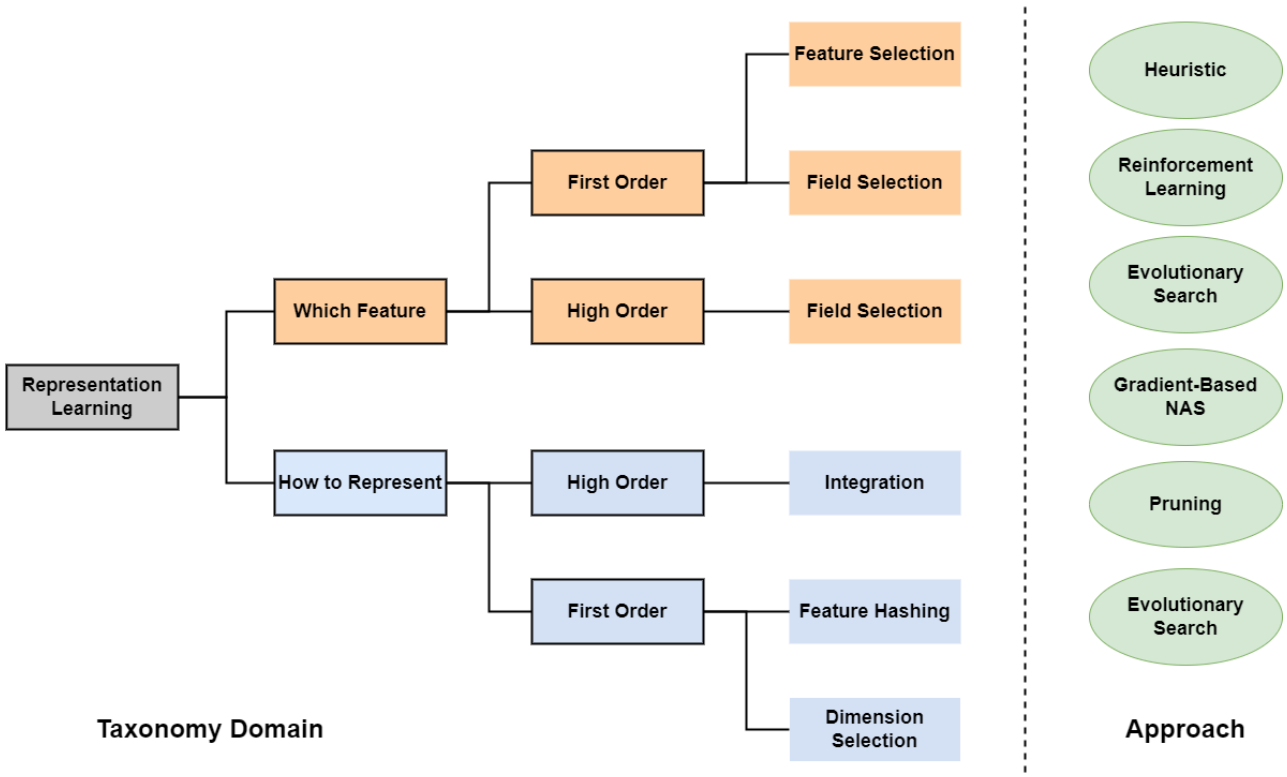


图 2：特征表示学习及相关方法的分类体系。

我们将所有输入嵌入记为 $\{\mathbf{e}_{i_p}\}$。已有多种方法 [Zhao 等，2021b；Zhao 等，2021a；Lyu 等，2022b] 被提出，用于探索嵌入层 $\mathbf{E}$ 的合适维度，而非预先指定一个数值。

对于高阶特征，其表示方法更加多样 [Lyu 等，2022a]。这里以二阶特征 $\mathbf{X}^{(2)} = \{\langle \mathbf{x}^{i_p}、\mathbf{x}_{j_q}\rangle\}$ 为例。以往的大多数工作采用特定的聚合函数来整合两个一阶特征。其形式可表示为：

$$\mathbf{e}_{\langle i_p, j_q\rangle} = f^{(2)}(\langle \mathbf{x}_{i_p}, \mathbf{x}_{j_q}\rangle) = g(\mathbf{x}_{i_p}, \mathbf{x}_{j_q}), \tag{4}$$

其中，聚合函数 $g()$ 可以是内积 [Rendle，2010]，外积 [Qu 等，2016]或交叉层 [Wang 等，2017]。

与此同时，某些方法 [Lyu 等，2022a] 将二阶特征视为交叉乘积特征，并使用额外的嵌入表 $\mathbf{E}^{(2)}$ 来表示。此方法可视为

$$\mathbf{e}_{\langle i_p, j_q\rangle} = f^{(2)}(\langle \mathbf{x}_{i_p}, \mathbf{x}_{j_q}\rangle) = \mathbf{E}^{(2)} \times \langle \mathbf{x}_{i_p}, \mathbf{x}_{j_q}\rangle. \tag{5}$$

这里将所有阶的特征表示写作

$$\mathcal{E} = [\{\mathbf{e}_{i_p}\}, \{\langle \mathbf{e}_{i_p}, \mathbf{e}_{j_q}\rangle\}, \dots]. \tag{6}$$

3.3 总体训练准则

总而言之，特征表示学习的目标可以概括如下：

$$\hat{y} = \mathcal{F}(\mathcal{G}, \mathcal{E}) \tag{7}$$

这里，分类器 $\mathcal{F}(\cdot)$ 将每个阶数的所有二元特征选择掩码和表示嵌入作为输入，并输出预测值 $\hat{y}$。

关于分类器，$\mathcal{F}(\cdot)$，近期提出了各种设计 [Wang等人，2017]。这些方法主要可分为三类：浅层 [Rendle，2010]，深层 [Qu 等人，2016] 或混合 [Guo 等人，2017；Wang 等人，2017]。然而，这些分类器的设计与特征的表示学习正交。

最终预测值根据公式 7 计算。采用交叉熵损失（即对数损失）来衡量每个样本的预测值与标签之间的差异：

$$\mathrm{CE}(y, \hat{y}) = y \log(\hat{y}) + (1 - y) \log(1 - \hat{y}), \tag{8}$$

其中，y 是用户点击的真实值。最终准确率损失总结如下：

$$\mathcal{L}_{\mathrm{CE}}(\mathcal{D}) = -\frac{1}{|\mathcal{D}|} \sum_{(\mathbf{x}, y) \in \mathcal{D}} \mathrm{CE}(y, \mathcal{F}(\mathcal{G}, \mathcal{E})), \tag{9}$$

其中，$\mathcal{D}$ 是训练数据集。

4 表示哪些特征

在预测模型中选择信息丰富的特征进行表示，对于准确预测至关重要。为有效提升性能，点击率预测领域已开发出多种方法。此外，特征共存也在点击率预测中发挥着重要作用 [Lyu等，2022a]。因此，我们将这些方法分为两类

：一类选择一阶特征，另一类选择二阶或高阶特征，下面将对此进行详细介绍。汇总表见表1。图3还提供了示意图，展示了该领域近期的进展和潜在的发展方向。

4.1 第一阶特征

一阶特征的选择可表述为学习一个二元张量 $\mathcal{G}^{(1)} \in \{0,1\}^n$，这在大规模在线点击率预测中可能会成为难题。为缩小选择空间，各种方法提出选择信息丰富的特征域，而非逐一选择每个领域。这些方法将二元选择张量 $\mathcal{G}^{(1)}$ 从 $\{0,1\}^n$ 缩减至 $\{0,1\}^m$，从而使搜索空间变得可行。

基于统计的方法 [Tibshirani, 1996年; Wold等,1987] 利用不同特征域的统计指标来进行特征域选择。主成分分析（PCA） [Wold 等，1987]将每个数据点投影到少数几个分量上，以获得低维数据，同时尽可能保留数据变化。该方法根据各分量中系数绝对值之和对特征域进行排序，并选取数值最高的少数特征域。LASSO [Tibshirani，1996] 通过限制回归系数平方和小于预设值来选择特征域，并最终将某些系数设为零，从而选出有信息量的特征域。

受近期神经架构搜索（NAS）发展的启发 [Liu 等，2019；Luo等， 2018], 研究者提出了基于神经架构搜索的方法 [Wang 等，2022；Guo 等，2022；Lin 等，2022] 以为点击率预测模型选择一阶特征。AutoField [Wang 等， 2022] 为每个特征域引入一个可学习值，用于指示保留还是丢弃该特征域。它根据基于梯度的神经架构搜索结果自动确定每个特征域的取值 [Liu 等，2019] 并交替更新可学习值和网络参数。AdaFS [Lin 等， 2022] 采用了与 AutoField [Wang 等， 2022]相同的设置。但它提出了一种新颖的控制器网络，根据输入样本决定是否选择各个特征域，因此更适用于动态推荐场景。LPFS [Guo等，2022]以平滑的L_0 优化取代基于梯度的选择器，从而能够高效地选择二阶或高阶特征域。

然而，将选择粒度从特征降至领域，过于粗略，无法确定细微特征。OptFS [Lyu 等， 2023] 通过延续学习训练方案高效探索特征级选择空间$\mathcal{G}^{(1)} \in \{0,1\}^n$，并确定使用哪个特征进行表示。该方法将第一阶特征选择与高阶特征选择分解开来，以实现准确预测。

4.2 高阶特征

对 t 阶特征的选择可以表述为学习二元张量 $\mathcal{G}^{(t)} \in \{0,1\}^{C_n^t}$。待探索的选择空间可能极其庞大。此前的所有方法 [Liu等，2020b；Su 等， 2021] 均采用字段级选择近似，将选择空间从 $\{0,1\}^{C_n^t}$ 缩减至 $\{0,1\}^{C_m^t}$。

高阶特征的选择始于针对特定模型的算法。BP-FIS [Chen 等人，2019；

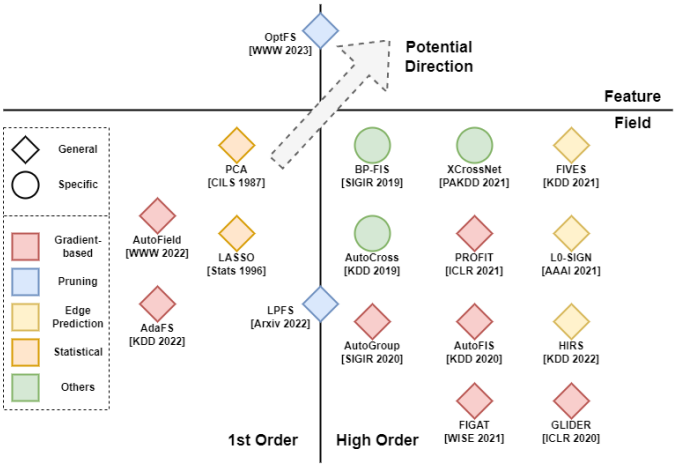


图3：关于选择哪些特征进行表示的总结与潜在方向。此处，不同颜色表示相应方法所采用的不同方法。不同形状表示这些方法所应用的骨干模型。

Chen 等人， 2022] 针对因子分解机 [Rendle，2010] 将二阶特征的选择视为学习个性化稀疏因子分解机。他们提出了贝叶斯变量选择方法，以减少二阶特征的数量。AutoCross [Luo 等人， 2019] 明确针对高阶交叉特征 [Wang 等人， 2017] 并通过高效的束搜索方法，在基于树的搜索空间中搜索有用的特征。

对于通用模型，AutoFIS [Liu 等，2020b]提出为每种可能的高阶特征添加一个注意力门来学习高阶特征，而不是枚举所有可能的组合。在此基础上，AutoGroup [Liu 等，2020a] 将高阶特征的选择视为结构优化问题，并通过基于梯度的神经架构搜索交替更新结构参数和网络参数（例如嵌入层和预测器） [Liu 等， 2019]。遵循 AutoFIS [Liu 等，2020b], 的设定，PROFIT [Gao 等， 2021] 提出通过对称分解来分解 k 阶特征域的选择过程。与 AutoGroup [Liu 等，2020a]类似，它也采用基于梯度的神经架构搜索来解决这一问题。GLIDER [Tsang 等， 2020] 提出从源模型中解释特征交互，并在目标模型中显式编码这些交互，而不假设模型结构。FIGAT [Long 等， 2021] 提出了一种名为门控注意力变换器的新型模型，它能够高效利用变换器的强大能力，同时去除冗余的高阶特征。与自注意力相比，原始注意力机制兼具可解释性和效率。XCrossNet [Yu 等， 2021]提出了一种高效且可解释的方法，以显式方式学习包含稠密和稀疏一阶特征的高阶交叉特征。

将每个一阶特征构建为一个节点时，高阶特征的选择也可以视为一个边预测问题。L0-SIGN [Su 等， 2021] 首次提出了一种选择信息丰富的二阶特征的方法，并

方法	顺序	粒度	骨干模型	方法
主成分分析 [1987]	第一	领域	通用	统计
LASSO [1996]	第一	领域	通用	统计
LPFS [2022]	首个且最高	领域	通用	剪枝
AutoField [2022]	第一	领域	通用	基于梯度的
AdaFS [2022]	第一	样本	通用	基于梯度的
OptFS [2023]	首个且最高	特征	通用	剪枝
BP-FIS [2019; 2022]	第一&第二	领域	因子分解机 [2010]	贝叶斯变量选择
AutoCross [2019]	high	领域	Cross [2017]	束搜索
GLIDER [2020]	high	领域	通用	基于梯度的
AutoFIS [2020b]	high	领域	通用	基于梯度的
FIGAT [2021]	high	领域	通用	基于梯度的
AutoGroup [2020a]	high	领域	通用	基于梯度的
PROFIT [2021]	high	领域	通用	基于梯度的
XCrossNet [2021]	high	领域	Cross [2017]	基于梯度的
L0-SIGN [2021]	第二	领域	通用	边预测
HIRS [2022]	high	领域	通用	边预测
FIVES [2021]	high	领域	通用	边预测

表1：特征表示方式汇总。通用 表示骨干模型可以是任意预测模型。

遵循信息瓶颈原理和统计交互理论解决该问题。作者进一步将边（连接两个节点）的定义扩展为超边（连接多个节点），并通过边预测选择高阶特征 [Su 等， 2022]。同样旨在图设置下选择高阶特征的 FIVES [Xie等， 2021] 采用了图神经网络中的层这一思想。一旦选出一个（$t-1$）阶特征，其信息就会聚合到 GNN 中的某个节点。通过预测该节点与表示一阶特征的其他节点之间的边，可以选出一个 A t阶特征。

5 如何表示选定的特征

如何表示选定的特征也是点击率表示学习中的一个重要问题。

5.1 一阶特征

以往关于一阶特征表示的研究 [Zhao 等，2021b；Ginart等，2021；Yan 等， 2021] 可分为两类：特征哈希和表示维度选择。前者倾向于对满足硬件要求的特征进行哈希，而后者旨在为每个特征选择合适的表示维度。

特征哈希

哈希一直是提升模型效率的经典方法。在点击率预测中，特征哈希 [Guo等，2021；Shi 等，2020；Kang 等，2021] 已被用于表示一阶特征。AutoDis [Guo 等， 2021] 专注于表示数值特征。它为每个数值字段引入元嵌入，以较少的参数表示该字段的全局知识。通过采用软离散化来捕捉元嵌入与数值特征之间的相关性，并通过聚合函数学习表示。

其他工作 [Shi 等，2020；Kang等， 2021] 则专注于对类别特征进行哈希。在组合嵌入 [Shi等， 2020], 中，每个特征的表示由互补划分所确定的多个嵌入组合而成。因此，系统存储多个较小的嵌入表，而不是一个大型嵌入表，在保留每个特征唯一性的同时提升效率。DHE [Kang 等， 2021] 首先通过多个哈希函数和变换，将特征值编码为唯一的标识符向量，然后应用深度神经网络将标识符向量转换为表示。编码模块具有确定性、不可学习且无需存储，而嵌入网络则在训练过程中更新，以学习表示生成过程。

表示维度选择

选择合适的维度来表示一阶特征已吸引了许多研究。已有研究指出，为特征基数较小的特征设置过多参数可能导致过拟合，而特征基数较大的特征则需要更大的维度来传达丰富的信息 [Lyu 等，2022b]。

最直接的方法是以启发式方式融入专家在这一主题上的经验。MDE [Ginart等，2021] 引入了混合维度嵌入层，而不是统一维度，并根据特征的频率以启发式方式分配每个表示的维度。EMDE [Beloborodov 等， 2022]借鉴了 MDE [Ginart 等， 2021] 的公式，并针对矩阵分解提出了两种混合嵌入层的变体。

除了启发式方法外，其他方法 [Liu等人，2020c；Zhao 等人，2021b；Joglekar 等人， 2020] 将维度选择建模为神经架构搜索问题。ESPAN [Liu 等人，2020c] 通过自动化强化学习智能体，根据流行度动态搜索用户和物品的表示维度。

方法	粒度	灵活	连续	可移除	方法
MDE [2021]	领域	✗	✓	✗	启发式
EMDE [2022]	领域	✗	✓	✗	启发式
ESAPN [2020c]	用户与物品	✗	✓	✗	强化学习
NIS [2020]	特征组	✗	✓	✗	强化学习
DNIS [2020]	特征组	✓	✓	✗	基于梯度的
AutoEmb [2021a]	用户与物品	✗	✓	✗	基于梯度的
AutoDim [2021b]	领域	✗	✓	✗	基于梯度的
ATML [2021]	特征组	✓	✓	✗	剪枝
RULE [2021]	项组	✗	✗	✗	进化搜索
PEP [2021]	特征	✓	✗	✓	剪枝
Autosrh [2022]	特征	✓	✗	✓	剪枝与基于梯度的
i-Razor [2022]	领域	✗	✓	✓	基于梯度的
OptEmbed [2022b]	特征	✓	✓	✓	剪枝与进化搜索

表 2：维度搜索方法总结。其中，*方法* 表示解决该问题的方案。*灵活* 表示无需预定义掩码。*连续* 表示所选表示是连续的。*可移除* 表示某些表示可以被完全移除。

然而，这种方法要求预先定义用户和物品的候选维度集合。类似地，NIS [Joglekar 等人，2020] 提出通过基于强化学习的神经架构搜索 [Pham 等人， 2018]来搜索合适的维度和词汇表大小。首先，它根据频率对所有特征进行排序，并将相邻特征归为一个特征组，作为基本搜索单元。然后，它从若干预定义维度中进行选择。DNIS [Cheng 等人， 2020]沿用了 NIS [Joglekar 等人，2020] 相同的特征分组机制，但通过连续松弛和可微优化，在更灵活的空间中放宽了维度选择。AutoEmb [Zhao 等人，2021a] 提出使用控制器网络，根据推荐系统设置下的频率，动态选择用户和物品的维度。AutoDim [Zhao 等人，2021b]将 AutoEmb [Zhao 等人，2021a]的方案扩展到任意领域。此外，Gumbel-Softmax [Jang 等人， 2017]采用这一技巧而非控制器网络来确定表示的维度。

除了将维度选择视为神经架构搜索问题外，已有多项研究 [Xiao 等，2022；Liu 等， 2021] 还将其表述为剪枝问题。HAM [Xiao 等， 2022] 提出通过字段级别的硬辅助掩码，对表示的维度进行结构化剪枝。该方法首先在嵌入表上引入正交正则化，以减少嵌入列之间的相关性并增强表示能力。受连续稀疏化 [Kusupati等， 2020], 的进展启发，PEP [Liu 等， 2021] 提出对特征表示中的冗余参数进行逐一剪枝。Autosrh [Kong 等， 2022] 同时采用剪枝和基于梯度的神经架构搜索方法。该方法类似于 DNIS [Cheng 等， 2020] 通过基于梯度的神经架构搜索放宽维度选择，并通过剪枝参数实现特定的压缩率。

如果将表示维度等于零视为可能的维度之一，那么第一阶特征的选择就可以看作维度选择问题的一个特例。i-Razor [Yao 等， 2022] 将维度等于零引入搜索候选集合，并提出一种可微神经架构搜索机制，以同时考虑第一阶特征的“如何选择”和“选择哪些”。OptEmbed [Lyu 等，2022b] 提出对这两个问题进行统一的联合处理。该方法同时利用具有信息量的第一阶特征和嵌入参数，高效地训练一个超网。随后，基于该超网执行单次进化搜索方法，以生成最优的嵌入表。

针对一阶特征的维度选择问题，也可能与其他特定问题相关，而不只是提升模型性能。ATML [Yan 等人，2021] 面向工业场景中的基于热启动的应用，并提出了一种置于嵌入层之后的自适应掩码双胞胎层，以连续的方式屏蔽每个特征表示中不需要的维度。然而，与以往的工作类似，它也根据频率对特征进行分组 [Joglekar 等人, 2020年; Cheng 等人，2020]. 另一项工作 RULE [Chen 等人，2021], 关注如何将训练好的模型轻松部署到具有任意设备特定内存限制的设备上，而无需重新训练。该工作提出了一种弹性嵌入，将一组嵌入块拼接起来。它还设计了一个基于估计器的进化搜索函数，用于在弹性嵌入中进行选择。

5.2 高阶特征

对于高阶特征，关键问题在于如何生成其表示。

大多数以往的研究 [Luo 等人，2019；Khawar等人，2020; Meng 等人， 2021]将高阶特征表示为聚合函数或网络的输出，该函数或网络以

方法	聚合函数候选项	方法
AutoCTR [2020]	{单层感知机, 因子分解机 [2010], 内积}	进化搜索
AutoRec [Wang等, 2020]	{单层感知机、拼接、自注意力、内积、随机搜索}	{随机搜索、网格搜索、贝叶斯优化}
AutoFeature [2020]	{加法、外积、拼接、广义乘积、空}	朴素贝叶斯
AutoPI [2021]	{跳跃连接、SE层 [2019],自注意力、单层感知机、一维卷积}	基于梯度的
NAS-CTR [2022]	{多层感知机、Cross [2017], 因子分解机 [2010], 逐元素}	近端算法
AutoIAS [2021]	{OP、拼接、加法、最大池化}	强化学习
AIM [2021]	{IP、OP、KP}	剪枝
OptInter [2022a]	{嵌入、外积、空}	基于梯度的

表 3：高阶特征表示方法总结。其中，*方法* 表示用于解决相应公式的方法。*内积*、*外积*、*单层感知机*、*多层感知机*、*核乘积*、*广义乘积* 和*跳跃连接* 分别表示内积、外积、单层感知机、多层感知机、核乘积、广义乘积和跳跃连接。*逐元素* 表示多种逐元素交互，包括加法、平均、内积、最大池化和最小值。*嵌入* 表示使用显式嵌入表来表示特征。*拼接* 表示拼接。*空* 表示该表示为零向量。*网格搜索* 表示网格搜索。*贝叶斯优化* 表示贝叶斯优化。

多个特征为输入。AutoCTR [Song 等人， 2020] 提出了一种由学习排序引导的进化搜索流程，并从三种不同的集成函数中选取特定组合。AutoRec [Wang 等人， 2020] AutoFeature [Khawar 等人，2020] 将搜索空间限制在简单函数中，例如内积，并设计了一种名为 NBTree 的进化搜索算法，以确定高阶特征的聚合函数。与 AutoFeature [Khawar 等人相比，2020], AutoPI [Meng 等人， 2021] 采用了更通用的搜索空间，其候选函数取自具有代表性的手工设计方法。NAS-CTR [Zhu 等人，2022] 将架构搜索表述为带有离散约束的联合优化问题，并提出了一种基于近端迭代的可微搜索算法。AutoIAS [Wei 等人， 2021] 提出了一个用于建模高阶特征表示的集成搜索空间。此外，在通过策略网络探索搜索空间时，采用经知识蒸馏训练的超网，对每种候选架构的性能进行一致预测。

不同于以往将高阶特征表示生成为某个函数输出的工作，OptInter [Lyu 等，2022a] 首次提出通过嵌入表显式表示高阶特征。为了减少显式建模带来的巨大存储需求，该方法利用基于梯度的神经架构搜索算法，自动为每个领域选择合适的表示。如何表示高阶特征表示可能与如何表示第一阶特征密切相关，因为前者通常基于后者计算。AIM [Zhu 等， 2021]也在统一框架中纳入了这两个问题，并利用基于梯度的方法技术生成稀疏结果。

6 总结与未来展望

过去几年中，特征表示学习一直是点击率预测领域中一个富有启发性且令人兴奋的研究主题。工业界和学术界的研究人员都在这一领域开展了大量研究。在本文综述中，我们提出了如图 2 所示的分类体系，按照两个主要问题对以往工作进行划分：（i）表示哪些特征；（ii）如何表示所选特征。

这两个问题也在第 3 节中统一为相同的形式。第 4 节和第 5 节分别详细解释了不同方法如何处理这两个问题。尽管近年来特征表示学习取得了显著进展，但我们注意到，该领域仍然存在重大挑战和未解决的问题。我们将潜在的发展方向总结如下：

细粒度特征选择 目前，特征选择方法不得不在庞大的可能搜索空间中做出取舍，并将粒度降低至字段级别，从而大幅减少了过程中可能的组合数量。在建模高阶特征时，这一现象尤为严重。图3也总结了该领域的主要进展。展望不久的将来，我们期待看到更多研究以更细粒度的方式探索高阶特征选择。

不同顺序特征的综合考量以往的大多数研究只关注选择某一特定顺序下有意义的特征。它们要么选择一阶特征并保留所有高阶特征，要么在固定一阶特征的同时选择高阶特征。前者无法筛除无用的高阶特征，导致计算成本增加、模型性能下降；后者则从所有可用的一阶特征中识别有用的高阶特征，造成一阶特征冗余。联合考虑不同顺序特征的选择，可能是一个很有前景的研究方向。

整合两类问题大多数研究似乎都专注于回答一个问题：要么是“选什么”，要么是“如何表示”。不过，一些研究者尝试在不同层面上统一这两类问题。Autosrh [Kong 等人、 2022], i-Razor [Yao 等人、 2022], PEP [Liu 等人， 2021] 以及 OptEmbed [Lyu 等人，2022b] 共同考虑了一阶特征的“选什么”和“如何表示”。AutoFeature [Khawar 等人， 2020] 以及 Opt-Inter [Lyu 等人，2022a] 将空表示作为表示高阶特征的一种可能方式，相当于舍弃该特征。“选什么”和“如何表示”密切相关，因为特征的表示方式可能受到所选特征的影响，反之亦然。整合这两类问题具有很大的潜力。

参考文献

[Beloborodov 等人， 2022] Dmitrii Beloborodov、Andrei Zimovnov、Petr Molodyk 和 Dmitrii Kirillov。用于矩阵分解的高效混合维度嵌入。CoRR，2022年。[Bengio 等人， 2013] Yoshua Bengio、Aaron C. Courville 和 Pascal Vincent。表示学习：综述与新视角。TPAMI，2013年。[Chen 等人， 2019] Yifan Chen、Pengjie Ren、Yang Wang 和 Maarten de Rijke。因子分解机的贝叶斯个性化特征交互选择。发表于 SIGIR，2019年。[Chen 等人， 2021] Tong Chen、Hongzhi Yin、Yujia Zheng、Zi Huang、Yang Wang 和 Meng Wang。学习弹性嵌入以定制端侧推荐系统。发表于 KDD，2021年。[Chen 等人， 2022] Yifan Chen、Yang Wang、Pengjie Ren、Meng Wang 和 Maarten de Rijke。因子分解机的贝叶斯特征交互选择。《人工智能》，2022年。[Cheng 等人， 2020] Weiyu Cheng、Yanyan Shen 和 Linpeng Huang。面向推荐系统的可微神经输入搜索。CoRR，2020年。[Gao 等人， 2021] Chen Gao、Yinfeng Li、Quanming Yao、Depeng Jin 和 Yong Li。深度稀疏网络的渐进式特征交互搜索。发表于 NeurIPS，2021年。[Ginart 等人， 2021] Antonio A. Ginart、Maxim Naumov、Dheevatsa Mudigere、Jiyan Yang 和 James Zou。混合维度嵌入及其在内存高效推荐系统中的应用。发表于 ISIT，2021年。[Guo 等人， 2017] Huifeng Guo、Ruiming Tang、Yunming Ye、Zhenguo Li 和 Xiuqiang He。DeepFM：一种用于点击率预测的基于因子分解机的神经网络。发表于 IJCAI，2017年。[Guo 等人， 2021] Huifeng Guo、Bo Chen、Ruiming Tang、Weinan Zhang、Zhenguo Li 和 Xiuqiang He。用于点击率预测中数值特征的嵌入学习框架。发表于 KDD，2021年。[Guo 等人， 2022] Yi Guo、Zhaocheng Liu、Jianchao Tan、Chao Liao、Daqing Chang、Qiang Liu、Sen Yang、Ji Liu、Dongying Kong、Zhi Chen 和 Chengru Song。LPFS：面向点击率预测的可学习极化特征选择。CoRR，2022年。[Huang 等人， 2019] Tongwen Huang、Zhiqi Zhang 和 Junlin Zhang。FiBiNET：结合特征重要性与双线性特征交互进行点击率预测。发表于 RecSys，2019年。[Jang 等人， 2017] Eric Jang、Shixiang Gu 和 Ben Poole。使用 Gumbel-Softmax 的类别重参数化。发表于 ICLR，2017年。[Joglekar 等人， 2020] Manas R. Joglekar、Cong Li、Mei Chen、Taibai Xu、Xiaoming Wang、Jay K. Adams、Pranav Khaitan、Jiahui Liu 和 Quoc V. Le。面向大规模推荐模型的神经输入搜索。发表于 KDD，2020年。

[Kang 等人， 2021] Wang-Cheng Kang、Derek Zhiyuan Cheng、Tiansheng Yao、Xinyang Yi、Ting Chen、Lichan Hong 和 Ed H. Chi。无需嵌入表即可学习推荐系统的类别特征嵌入。发表于 KDD，2021年。[Khawar 等人， 2020] Farhan Khawar、Xu Hang、Ruiming Tang、Bin Liu、Zhenguo Li 和 Xiuqiang He。AutoFeature：为点击率预测搜索特征交互及其架构。发表于 CIKM，2020年。[Kong 等人， 2022] Shuming Kong、Weiyu Cheng、Yanyan Shen 和 Linpeng Huang。Autosrh：一种用于表格数据预测的嵌入维度搜索框架。TKDE，2022年。[Kusupati 等人， 2020] Aditya Kusupati、Vivek Ramanujan、Raghav Somani、Mitchell Wortsman、Prateek Jain、Sham M. Kakade 和 Ali Farhadi。用于可学习稀疏性的软阈值权重重参数化。发表于 ICML，2020年。[Lin 等人， 2022] Weilin Lin、Xiangyu Zhao、Yejing Wang、Tong Xu 和 Xian Wu。Adafs：深度推荐系统中的自适应特征选择。发表于 KDD，2022年。[Liu 等人， 2019] Hanxiao Liu、Karen Simonyan 和 Yiming Yang。DARTS：可微架构搜索。发表于 ICLR，2019年。[Liu 等，2020a] Bin Liu、Niannan Xue、Huifeng Guo、Ruiming Tang、Stefanos Zafeiriou、Xiuqiang He 和 Zhenguo Li。AutoGroup：用于在点击率预测中建模显式高阶特征交互的自动特征分组。发表于 SIGIR，2020年。[Liu 等，2020b] Bin Liu、Chenxu Zhu、Guilin Li、Weinan Zhang、Jincai Lai、Ruiming Tang、Xiuqiang He、Zhenguo Li 和 Yong Yu。AutoFIS：用于点击率预测的分解模型中的自动特征交互选择。发表于 KDD，2020年。[Liu 等人，2020c] Haochen Liu、Xiangyu Zhao、Chong Wang、Xiaobing Liu 和 Jiliang Tang。深度推荐系统中的自动嵌入大小搜索。发表于 SIGIR，2020年。[Liu 等人， 2021] Siyi Liu、Chen Gao、Yihong Chen、Depeng Jin 和 Yong Li。推荐系统的可学习嵌入大小。发表于 ICLR，2021年。[Long 等人， 2021] Chao Long、Yanmin Zhu、Haobing Liu 和 Jiadi Yu。利用门控注意力变换器高效学习特征交互。发表于 WISE，2021年。[Luo 等， 2018] Renqian Luo、Fei Tian、Tao Qin、Enhong Chen 和 Tie-Yan Liu。神经架构优化。发表于 NeurIPS，2018年。[Luo 等， 2019] Yuanfei Luo、Mengshuo Wang、Hao Zhou、Quanming Yao、Wei-Wei Tu、Yuqiang Chen、Wenyuan Dai 和 Qiang Yang。AutoCross：面向现实世界应用的表格数据自动特征交叉。发表于 KDD，2019年。[Lyu 等，2022a] Fuyuan Lyu、Xing Tang、Huifeng Guo、Ruiming Tang、Xiuqiang He、Rui Zhang 和 Xue Liu。

记忆、分解，还是采用朴素方法：学习点击率预测的最优特征交互方法。发表于 ICDE，2022年。[Lyu等，2022b] Fuyuan Lyu、Xing Tang、Hong Zhu、Huifeng Guo、Yingxue Zhang、Ruiming Tang 和 Xue Liu。OptEmbed：学习用于点击率预测的最优嵌入表。发表于 CIKM，2022年。[Lyu 等人， 2023] Fuyuan Lyu、Xing Tang、Dugang Liu、Liang Chen、Xiuqiang He 和 Xue Liu。优化点击率预测的特征集。CoRR，2023年。[Meng 等人， 2021] Ze Meng、Jinnian Zhang、Yumeng Li、Jiancheng Li、Tanchao Zhu 和 Lifeng Sun。一种用于自动发现点击率预测中强交互的通用方法。发表于 SIGIR，2021年。[Pham 等人，2018] Hieu Pham、Melody Y. Guan、Barret Zoph、Quoc V. Le 和 Jeff Dean。通过参数共享实现高效的神经架构搜索。CoRR，2018年。[Qu 等人， 2016] Yanru Qu、Han Cai、Kan Ren、Weinan Zhang、Yong Yu、Ying Wen 和 Jun Wang。用于用户响应预测的基于乘积的神经网络。发表于 ICDM，2016年。[Rendle， 2010] Steffen Rendle。因子分解机。发表于 ICDM，2010年。[Shi 等人， 2020] Hao-Jun Michael Shi、Dheevatsa Mudigere、Maxim Naumov 和 Jiyan Yang。利用互补划分构建组合嵌入，实现内存高效的推荐系统。发表于 KDD，2020年。[Song 等人， 2020] Qingquan Song、Dehua Cheng、Hanning Zhou、Jiyan Yang、Yuandong Tian 和 Xia Hu。面向点击率预测的自动化神经交互发现。发表于 KDD，2020年。[Su 等人， 2021] Yixin Su、Rui Zhang、Sarah M. Erfani 和 Zhenghua Xu。检测推荐系统中的有益特征交互。发表于 AAAI，2021年。[Su 等人， 2022] Yixin Su、Yunxiang Zhao、Sarah M. Erfani、Junhao Gan 和 Rui Zhang。检测推荐系统中任意阶的有益特征交互。发表于 KDD，2022年。[Tibshirani， 1996] Robert Tibshirani。通过 LASSO 进行回归收缩与变量选择。《英国皇家统计学会杂志：B辑（方法论）》，1996年。[Tsang 等人， 2020] Michael Tsang、Dehua Cheng、Han-peng Liu、Xue Feng、Eric Zhou 和 Yan Liu。特征交互的可解释性：通过神经交互检测解释广告推荐系统的案例。发表于 ICLR，2020年。[Wang 等人， 2017] Ruoxi Wang、Bin Fu、Gang Fu 和 Mingliang Wang。用于广告点击预测的深度交叉网络。发表于 ADKDD，2017年。[Wang 等人， 2020] Ting-Hsiang Wang、Xia Hu、Haifeng Jin、Qingquan Song、Xiaotian Han 和 Zirui Liu。AutoRec：一种自动化推荐系统。发表于 RecSys，2020年。[Wang 等人， 2022] Yejing Wang、Xiangyu Zhao、Tong Xu 和 Xian Wu。AutoField：自动化深度推荐系统中的特征选择。发表于 WWW，2022年。[Wei 等人， 2021] Zhikun Wei、Xin Wang 和 Wenwu Zhu。AutoIAS：用于点击率预测的自动集成架构搜索器。发表于 CIKM，2021。[Wold 等人， 1987] Svante Wold、Kim Esbensen 和 Paul Geladi。主成分分析。《化学计量学与智能实验室系统》，1987年。[Xiao 等人， 2022] Tesi Xiao、Xia Xiao、Ming Chen 和 Youlong Chen。通过结构化硬辅助掩码剪枝进行字段级嵌入维度搜索，以实现点击率预测。CoRR，2022。[Xie 等人，2021] Yuexiang Xie、Zhen Wang、Yaliang Li、Bolin Ding、Nezihe Merve G¨urel、Ce Zhang、Minlie Huang、Wei Lin 和 Jingren Zhou。FIVES：通过边搜索实现大规模表格数据的特征交互。发表于 KDD，2021。[Yan 等人，2021] Bencheng Yan、Pengjie Wang、Kai Zhang、Wei Lin、Kuang-Chih Lee、Jian Xu 和 Bo Zheng。通过自适应掩码双分支层学习高效且有效的嵌入。发表于 CIKM，2021。[Yao 等人， 2022] Yao Yao、Bin Liu、Haoxun He、Dakui Sheng、Ke Wang、Li Xiao 和 Huanhuan Cao。i-Razor：用于大规模推荐系统中特征选择和维度搜索的神经输入剪刀。CoRR，2022。[Yu 等人， 2021] Runlong Yu、Yuyang Ye、Qi Liu、Zihan Wang、Chunfeng Yang、Yucheng Hu 和 Enhong Chen。XCrossNet：面向特征结构的点击率预测学习。发表于 PAKDD，2021。[Zhang 等人，2021] Weinan Zhang、Jiarui Qin、Wei Guo、Ruiming Tang 和 Xiuqiang He。用于点击率估算的深度学习。发表于 IJCAI，2021。[Zhao等，2021a] Xiangyu Zhao、Haochen Liu、Wenqi Fan、Hui Liu、Jiliang Tang、Chong Wang、Ming Chen、Xudong Zheng、Xiaobing Liu 和 Xiwang Yang。AutoEmb：流式推荐中的自动嵌入维度搜索。发表于 ICDM，2021。[Zhao 等人，2021b] Xiangyu Zhao、Haochen Liu、Hui Liu、Jiliang Tang、Weiwei Guo、Jun Shi、Sida Wang、Huiji Gao 和 Bo Long。AutoDim：推荐系统中的字段感知嵌入维度搜索。发表于 WWW，2021。[Zheng 等人， 2023] Ruiqi Zheng、Liang Qu、Bin Cui、Yuhui Shi 和 Hongzhi Yin。深度推荐系统的自动机器学习：综述。《ACM 信息系统汇刊》，2023。[Zhu 等人， 2021] Chenxu Zhu、Bo Chen、Weinan Zhang、Jincai Lai、Ruiming Tang、Xiuqiang He、Zhenguo Li 和 Yong Yu。AIM：用于点击率预测的自动交互机器。TKDE，2021。[Zhu 等人， 2022] Guanghui Zhu、Feng Cheng、Defu Lian、Chunfeng Yuan 和 Yihua Huang。NAS-CTR：用于点击率预测的高效神经架构搜索。发表于 SIGIR，2022。