T2G-FORMER 论文翻译与解析

摘要

翻译： 深度神经网络(DNNs)在表格学习中的最新发展很大程度上得益于 DNNs 自动特征交互的能力。然而，表格特征的异质性使得这些特征相对独立，开发有效的方法来促进表格特征交互仍然是一个开放问题。在本文中，我们提出了一种新颖的图估计器(Graph Estimator)，它能自动估计表格特征之间的关系，并通过在相关特征之间分配边来构建图。这种关系图将独立的表格特征组织成一种图数据，使得节点(表格特征)之间的交互可以有序地进行。基于我们提出的图估计器，我们提出了一个为表格学习量身定制的 Transformer 网络，称为 T2G-FORMER，它通过执行由关系图引导的表格特征交互来处理表格数据。一个特定的跨层读出模块收集 T2G-FORMER 各层预测的显著特征，并获取最终预测的全局语义。全面的实验表明，我们的 T2G-FORMER 在 DNNs 中实现了优越的性能，并且与非深度梯度提升决策树模型相比具有竞争力。代码和模型可在 https://github.com/jyansir/t2g-former 获取。

解析与重点：

- 研究问题：表格数据特征的异质性使得特征交互困难，需要新方法促进表格特征之间的有效交互
- 主要贡献：提出了图估计器(Graph Estimator)，可以自动估计表格特征之间的关系并构建关系图
- 技术亮点：将独立表格特征组织成图数据结构，实现有序的特征交互
- 提出的模型：T2G-FORMER，一个为表格学习定制的 Transformer 网络
- 创新点：使用跨层读出(Cross-level Readout)模块收集各层预测的显著特征
- 实验结果：在表格学习任务上超越了现有的深度神经网络，与 GBDT 模型相比具有竞争力

1. 引言

翻译： 表格结构形式的数据在许多领域普遍存在，例如医疗记录和点击率(CTR)预测。观察到，梯度提升决策树(GBDT)是机器学习和工业应用中表格数据任务的主导模型。由于深度神经网络(DNNs)在各个领域的巨大成功，越来越多

的专门针对表格数据学习的 DNNs 得到了发展。这些研究要么利用神经网络的集成来构建可微分树模型，要么探索多样化的交互方法，通过融合不同的表格特征来学习综合特征。

然而，与图像和文本不同，对于基于融合的模型来说，由于特征异质性问题，处理表格特征交互是具有挑战性的。DANets 建议使用"选择与抽象"原则，通过首先选择然后交互所选特征来处理表格数据。已知的神经特征选择方案可以分为软选择和硬选择两种版本。软选择本质上是在特征之间施加全连接交互，如乘性交互、特征交叉和基于注意力的交互。然而，表格特征在本质上是异质的，全连接交互是次优选择，因为它盲目地将所有特征融合在一起。DANets 通过对相关特征进行分组，然后约束分组特征之间的交互来执行硬选择。尽管 DANets 取得了令人期待的结果，但其特征选择操作无法彻底解决组内交互问题，因此分配在同一组中的特征被不加区别地融合，使模型表达能力较差。

有许多日常应用展示了异质表格特征选择性交互的重要性。图 1 左部给出了医疗数据表的一个例子。使用潜在的医学知识，可以形成静态图来表示合理特征对之间的关系。例如，身高和体重的关系给出一个代表高级语义体格的概率。同样，体重和血压(BP)之间的关系可能表示心血管健康的语义。此外，可能有一些"惰性特征"与任何其他特征无关，例如表示 HIV 抗体水平(HIV-Ab)的特征。在右部，图 1(a)显示了原始表格特征，其关系未指定，如果不确定特征关系，则无法直接获得更高级别的语义。图 1(b)说明了软选择的全连接交互，这可能在特征融合中引入一些噪声关系(例如，"惰性特征"与其他特征连接)。具有分组操作的硬选择(例如在 DANets 中使用)通过对相关特征进行分组来实现部分选择性交互(见图 1(c))，但仍可能包含噪声交互。它只能对相关特征进行分组，但无法处理同一组内的特征关系。在图 1(c)中，分组设计只能将身高、体重和 BP 特征放在一起进行相互交互，但不能排除无意义的身高-BP 对。很直观，精确的健康状况评估可以基于特定记录值(例如图 1 中身高为 173.6 厘米)和由关系图边缘表示的潜在知识进行。对于医疗表中的第一个样本(ID=1)，共同考虑身高和体重值可以提示超重症状。同样，结合体重和 BP 值表明心血管问题的风险。第二个样本(ID=2)仅基于 HIV-Ab 特征直接表明 HIV 感染的风险。因此，我们认为处理此类复杂决策过程的理想方式是建立一个具有自适应边权的图。边权(在图 1(d)中用不同颜色和宽度表示)基于特定特征值指示关系强度，而静态图拓扑表示约束有意义关系的潜在知识。

受上述观察启发，在本文中，我们提出构建表格特征图来指导特征交互。我们

开发了一个新颖的图估计器(GE)，用于将独立表格特征组织成特征关系图(FR-Graph)。此外，我们提出了一个为表格学习量身定制的 Transformer 网络，称为 T2G-FORMER，通过堆叠包含 GE 的块来进行选择性特征交互。GE 通过组装(i)描述任务潜在知识的静态图拓扑和(ii)图边的数据自适应边权来建模 FR-Graph。静态图描述了潜在知识(特征对的关系)，而数据自适应边权表示基于特定特征值的关系强度。使用 FR-Graph，我们可以有效地捕获更微妙的交互，这些交互可能会被分组策略误处理(如图 1(c)所示)。在我们提出的 T2G-FORMER 中，每一层使用 FR-Graph 将层输入特征转换为图数据，并基于图边的规范以有序方式执行异质特征交互。此外，特殊的跨层读出模块从每个层收集显著特征，并获取最终预测的全局表格语义。

T2G-FORMER 的工作流程如下。一个 FR-Graph，其边表示具有由 GE 模块预测的数据自适应权重的静态特征关系，指导表格特征交互过程以预测更高级别的特征。然后，为更高级别的表格特征构建另一个 FR-Graph 来组织特征交互，过程继续。T2G-FORMER 可以通过重复上述过程输出来自不同特征级别的综合语义。共享的跨层读出用于聚合来自不同特征级别的语义，并在最终预测中考虑所有这些特征。

总体而言，我们的工作的主要贡献如下：

- 我们首次利用特征关系图处理表格数据的异质特征交互，并提出了一种新颖的 GE 模块用于特征关系组织。
- 我们将特征关系图适配到 Transformer 架构中，并构建了一个专门用于表格分类和回归的表格学习 Transformer T2G-FORMER。
- 全面的实验表明，T2G-FORMER 在许多数据集上始终优于最先进的表格 DNNs，并且与 GBDTs 相比具有竞争力。

解析与重点：

- 研究背景：表格数据广泛存在于医疗、电子商务等领域，GBDT 传统上是主导模型
- 问题挑战：表格特征的异质性使得特征交互困难，现有方法有局限性
- 两种特征选择方法的比较：
    - 软选择：全连接交互，但可能引入噪声

    - 硬选择：分组特征，但无法处理组内细节交互
- 提出的创新：图估计器(GE)和特征关系图(FR-Graph)
    - 静态图拓扑：代表潜在领域知识
    - 数据自适应边权：根据具体数据反映关系强度
- 实际应用场景：基于医疗数据的例子展示了方法的必要性和有效性
- 技术架构：T2G-FORMER 将表格特征组织为图结构，进行有序特征交互
- 主要流程：通过图引导特征交互，跨层读出收集关键信息
- 三大贡献：特征关系图应用、Transformer 架构设计、实验验证

2. 相关工作

2.1 表格学习的深度神经网络

翻译： 表格学习指的是基于分类或连续特征进行预测的表格数据机器学习应用。传统的非深度方法在此类任务中是普遍选择，特别是决策树的集成方法，如 GBDT、XGBoost、LightBGM 和 CatBoost。

与其浅层对应物相比，深度神经网络具有强大的自动特征学习能力，因此提供了挖掘隐藏特征的良好潜力。近来，越来越多的研究将 DNNs 应用于表格数据，这些研究可以大致分为可微分树模型和基于融合的模型两类。

可微分树模型：此类 DNNs 受到集成树框架成功的启发。NODE 结合了可微分的不透明决策树与多层次层级表示，并获得了与 GBDT 相竞争的性能。TabNet 采用注意力机制来顺序选择树状决策的显著特征。Net-DNF 引入了析取范式的偏置来选择和聚合每个块中的特征子集。NODE 和 Net-DNF 很大程度上受益于模型集成，但没有利用 DNNs 的特征表示能力。TabNet 设计了非交互式 transformer 块用于特征表示和选择，但没有特征融合。所有这些 DNNs 都作为特征选择器和分割器，但忽略了表格特征之间的潜在交互。

基于融合的模型：融合基于模型利用 DNNs 通过特征交互融合更高级别的特征。DeepFM 对编码特征执行乘性交互，用于 CTR 预测。DCN 将 DNNs 与交叉组件结合，学习具有高阶交互的复杂特征。最近，由于其交互偏置和显著性能，注意力模块成为了一种流行选择。AutoInt 使用多头自注意力交互低维嵌入特征。TabTransformer 直接将 Transformer 块转移到表格数据，但忽略了分类特

征和连续特征之间的交互。FT-Transformer 通过对这两种类型的特征进行标记化并平等处理它们来解决这个问题。DANets 选择相关的表格特征，并将所选特征注意力融合为更高级别的特征。

2.2 表格特征交互

翻译： 之前基于融合的工作大多只是简单地将成功的神经架构(如 MLP、自注意力和 Transformer)转移到表格数据中，并通过软选择交互特征。然而，特征异质性导致了归纳偏置的差距，使这些模型(为同质数据如图像和文本设计的)次优。DANets 首次通过硬选择调整了选择性特征交互，限制在特征组内的交互，并取得了令人期待的结果；但是，组内特征的关系仍然没有得到很好的管理。因此，本文提出了特征关系图并将其适配到定制的 Transformer 网络中。

解析与重点：

- 表格学习概述：处理分类或连续特征的表格数据预测任务
- 两大类深度学习方法：
    1. 可微分树模型：
        - NODE：结合可微分决策树和层级表示
        - TabNet：使用注意力机制顺序选择特征
        - Net-DNF：使用析取范式进行特征选择和聚合
        - 共同缺点：忽略了特征之间的潜在交互
    2. 基于融合的模型：
        - DeepFM：乘性特征交互
        - DCN：结合 DNNs 和交叉组件
        - AutoInt：多头自注意力
        - TabTransformer：直接应用 Transformer 架构
        - FT-Transformer：处理分类和连续特征
        - DANets：选择相关特征并融合
- 现有方法的主要局限：

    - 简单转移架构而不考虑表格数据的特殊性
    - 特征异质性问题未得到充分解决
    - 组内特征关系未得到良好管理
- 本文的独特贡献：提出特征关系图和定制 Transformer 解决上述问题

3. 图估计器

翻译： 我们提出图估计器(GE)(图 2(b))用于自动构建特征关系图(FR-Graphs)，它将表格特征视为图中的节点，并估计特征关系作为边。GE 设计的灵感来自知识图谱补全(KGC)，后者可能使用两个实体的语义相似性来估计它们关系的合理性。测量语义相似性的基本形式是：

fr(h, t) = h^T Mr t, (1)

其中 h, t ∈ R^n 是编码的头实体节点和尾实体节点，可学习矩阵 Mr ∈ R^n×n 表示知识图(KG)中的关系 r。随后的各种方法遵循这一思路，它们之间的区别仅在于关系嵌入和评分函数。

与 KGC 模型仅计算实体的静态关系合理性不同，GE 通过具有数据自适应边权的静态底层图拓扑估计特征关系。我们将每个表格特征作为节点，首先执行语义匹配以估计表格特征之间成对交互的软合理性，这在本节中被称为数据自适应边权。其次，根据表格列语义学习静态知识拓扑，以保留显著特征对的交互。最后，边权与知识拓扑组装形成 FR-Graph。

3.1 FR-Graph 结构组件

翻译： 为了挖掘表格特征之间的关系，我们通过将表格特征视为图节点候选并预测它们之间的边来构建 FR-Graph。边从两个角度产生：表示数据特定信息的自适应边权，和表示所有数据的底层知识的静态边拓扑。请注意，如果没有其他节点与它们连接，一些特征会与 FR-Graph 隔离。

自适应边权：给定两个表格特征嵌入向量 xi, xj ∈ R^n (i, j ∈ {1, 2, ..., N})，其中 N 是输入特征(表列)的数量，我们使用以下成对评分函数评估它们的交互合理性：

Gw[i, j] = gw(f^h_i, f^t_j) = f^h_i^T diag(r) f^t_j, (2)

f^h_i = W^h xi, f^t_i = W^t xi,

W^h ≡ W^t 如果对称, W^h ≠ W^t 如果不对称, (3)

其中两个可学习参数 W^h, W^t ∈ R^m×n 表示头特征和尾特征的转换，diag(r) ∈ R^n×n 是由可学习关系向量 r ∈ R^n 参数化的对角矩阵，语义上表示特征交互关系。这里 W^h 和 W^t 在成对特征边权对称(即 Gw[i, j] ≡ Gw[j, i])时共享参数，在不对称情况(即 Gw[i, j] ≠ Gw[j, i])下则是参数独立的。为简洁起见，省略了所有偏置向量。因此，所有特征对的自适应权重分数 gw 构成了一个完全连接的加权关系图 Gw。注意，当 r 填充标量值 1(diag(r)成为单位矩阵)时，边权分数退化为注意力分数，因此它能够测量加权特征相似性。

静态知识拓扑：虽然我们为所有特征对引入了软边权，但全局考虑表格数据的底层知识也很重要。因此，我们使用一系列列嵌入来表示表格特征的语义，静态关系拓扑分数可以如下计算：

Gt[i, j] = gt(e^h_i, e^t_j) = (e^h_i^T e^t_j) / (||e^h_i||2 ||e^t_j||2), (4)

e^h_i = E^h[:, i], e^t_i = E^t[:, i],

其中 E ∈ {E^h, E^t}是可学习的列嵌入，分为头视图或尾视图，E = (e1, e2, ..., eN) ∈ R^d×N，d 是嵌入维度。类似地，关系拓扑分数 gt 具有对称和不对称对应物，在对称关系拓扑(即 Gt[i, j] ≡ Gt[j, i])中 E^h 和 E^t 共享参数，但在不对称情况(即 Gt[i, j] ≠ Gt[j, i])下是参数独立的。我们在 gt 分数函数中使用 L2 归一化，将嵌入转换为类似尺度并提高训练稳定性。

我们基于等式(4)中的 Gt 分数生成静态关系拓扑，如下：

A = ftop(Gt) = 1[σ1(Gt + b) > T], (5)

其中 σ1 是由可学习偏置 b 参数化的元素级激活(类似 PReLU 中的操作)，Gt 是由关系拓扑分数 gt 组成的邻接矩阵分数，T 是用于信号剪裁的常数阈值，1 表示指示函数。通过这种方式，我们获得一个全局图拓扑(邻接矩阵 A)来约束特征交互，这个拓扑可以被视为整个任务的静态知识。

3.2 关系图组装

翻译： 由于我们从数据视角获得了"软"自适应边权，并从知识视角获得了"硬"静态关系图拓扑，我们将它们组合生成 FR-Graph，遵循"基于特定数据和底层知识的决策"理念。具体来说，我们将两个组件组装如下：

G = σ2(fnsi(A) ⊙ Gw), (6)

其中 σ2 是竞争性激活(例如，Lp 归一化、softmax、entmax、sparsemax)，用

于限制每个"特征节点"的入度，⊙表示 Hadamard 乘积。结果关系图 G 是基于自适应特征匹配和静态知识拓扑的加权图。为了帮助 FR-Graph 专注于学习特征之间有意义的交互，执行"无自交互"函数 fnsi 以显式排除 G 中的自环。我们使用 FR-Graph 指导后续特征交互。由于边权和知识拓扑都有对称和不对称版本，因此有四种 FR-Graph 组合，覆盖完整的关系图。在实验中，我们将进一步讨论 FR-Graph 类型的影响。

解析与重点：

- 图估计器概念：将表格特征视为图节点，估计特征间关系作为边
- 灵感来源：知识图谱补全(KGC)中的语义相似性评估
- FR-Graph 的两个关键组件：
    1. 自适应边权：
        - 基于特定数据值评估特征对的交互合理性
        - 使用可学习参数 W^h, W^t 转换特征
        - 可设计为对称或不对称版本
        - 当 r 为标量 1 时，相当于注意力机制
    2. 静态知识拓扑：
        - 使用列嵌入表示表格特征的全局语义
        - 通过余弦相似度计算特征对关系
        - 应用阈值剪裁获得二元图拓扑
        - 代表整个任务领域的静态知识
- 关系图组装策略：
    - 结合自适应边权和静态拓扑
    - 使用 Hadamard 乘积融合两部分信息
    - 应用竞争性激活控制节点入度
    - 专门排除自环，关注不同特征间交互

    - 四种可能的组合类型(对称/不对称)

- 技术创新点：
    - 同时考虑数据特定模式和领域知识
    - 能处理特征异质性问题
    - 支持精细化的特征关系建模

4. T2G-FORMER

翻译： 我们将 GE 整合到类似注意力的基本块中，并通过堆叠多个块构建 T2G-FORMER 来进行选择性表格特征交互(见图 2)。T2G-FORMER 使用估计的 FR-Graphs 交互特征，并逐层获得更高级别的特征。跨层读出按顺序转换到每个层的特征空间，并选择性地收集显著特征用于最终预测。添加了一个快捷路径以保留来自前面层的信息，使得在不同特征级别的门控融合促进了模型能力。

4.1 基本块

翻译： 一个单一块配备了 GE 用于选择性特征交互(见图 2(b))。给定输入到第 l 层的特征 $X^l \in R^{n \times N}$，我们获得更高级别的特征 $X^{(l+1)}$如下：

$$G^l = GE(X^l),\ V^l = W_v X^l, \quad (7)$$

$$H^l = G^l V^l + g(X^l),\ X^{(l+1)} = FFN(H^l) + g(H^l), \quad (8)$$

其中 $W_v \in R^{m \times n}$ 是特征转换的可学习参数，V^l 是转换后的输入特征。FFN 表示前馈网络。由于在 G^l 中排除了自交互(见等式(6))，因此添加了快捷路径 g 以保护来自前一层的信息，这在实验中是一个简单的 dropout 层。值得注意的是，我们生成并使用 FR-Graph 进行特征交互，并不影响由快捷路径进行的特征内部更新。在第一层，我们将 X^0 设置为由简单特征标记器编码的输入表格数据。通过这种方式，可以使用生成的 FR-Graphs 和选择性交互迭代获得更高级别的特征。在实现中，执行层归一化(见图 2(b))以稳定训练。

4.2 跨层读出

翻译： 我们设计了一个全局读出节点，选择性地收集每一层的显著特征，并获取最终预测的综合语义。具体来说，我们注意力融合当前层的选定特征，并通过快捷路径将它们与前面层的低级特征组合。给定当前读出状态 $z^l \in R^n$，第 l 层的收集过程定义为：

α^l_i = g_w(h^l, f^t_i) · f_top(g_t(e^l, e^t_i)), h^l = W_h z^l, (9) r^l = softmax(α^l)^T V^l + z^l, (10) z^(l+1) = FFN(r^l) + r^l, (11)

其中 α^l_i 表示第 i 个特征的权重，构成权重向量 α^l ∈ R^N，e^l ∈ R^d 是表示第 l 层读出节点语义的可学习向量，f^t_i 是每层的编码特征(等式(3))，e^t_i 是层级列嵌入(等式(4))。V^l 是转换后的输入特征(等式(7))。这里我们将 z^l 送入相同的 FFN 转换，以将当前读出转换到(l+1)层的特征空间，用于下一轮收集。快捷路径直接添加，无信息丢失。这个收集过程从输入特征到最高级别特征重复进行，因此鼓励跨级别特征之间的交互。

4.3 整体架构与训练

翻译： 在 T2G-FORMER 中堆叠基本块(图 2(a))。如果没有特殊说明，在实验中我们默认在每个块中使用 8 头 GE(图 2(b))。基于处理最终层 L 后的读出状态进行预测，如下：

ŷ = FC(ReLU(LN(z^L))),

其中 LN 和 FC 表示层归一化和全连接层。对于优化，我们使用交叉熵损失用于分类，均方误差损失用于回归，类似于之前的 DNNs。我们测试了各种任务，并观察到在整个训练阶段继续优化等式(5)中的静态图拓扑 A 可能会导致在一些简单任务(如二进制分类、小数据集或少量输入特征)上的性能不稳定。因此，我们在收敛后冻结它，以固定拓扑方式进一步训练。

注意，我们引入了额外的超参数 d(等式(4))和 T(等式(5))。在实验中，我们自适应设置 $d = 2\lceil \log_2 N \rceil$，这是为了呈现具有 N^2二进制元素的邻接矩阵所需的最小信息量，并在所有数据集中保持 T = 0.5。我们选择 sigmoid 作为 σ_1，softmax 作为 σ_2。使用直通技巧解决等式(5)中指示函数

5. 实验

翻译： 在本节中，我们展示了广泛的实验结果，并与各种最先进的表格学习 DNNs 和 GBDT 进行了比较。此外，我们进行了实证实验，以检验 T2G-FORMER 的一些关键组件的影响，包括比较特征关系图(FR-Graph)类型、自交互的消融研究以及 GE 的效果。另外，我们通过在两个语义丰富的数据集上可视化 FR-Graphs 和读出选择来探索模型的可解释性。

5.1 实验设置

翻译： 数据集：我们使用了十二个开源表格数据集。手势阶段预测(GE)、客户流失建模(CH)、眼动(EY)、加州住房(CA)、House 16H(HO)、成人(AD)、Helena(HE)、Jannis(JA)、Otto Group 产品分类(OT)、Higgs Small(HI)、Facebook 评论(FB)和 Year(YE)。对于每个数据集，数据预处理和训练-验证-测试分割根据(Gorishniy 等人，2021; Gorishniy 等人，2022)固定。表 1 给出了数据集统计信息，更多细节见附录 A。

实现细节：我们使用 PyTorch 在 Python 3.8 上实现了 T2G-FORMER 模型。所有实验都在 NVIDIA RTX 3090 上运行。在训练中，如果没有特殊说明，我们在 GE 中使用具有对称边权和非对称图拓扑的 FR-Graphs。优化器是 AdamW，除学习率和权重衰减率外，使用默认配置。对于 DANet-28，我们遵循其 QHAdam 优化器和(Chen 等人，2022)中给出的预设超参数，不进行调整。对于 NODE，我们使用网格搜索。对于其他 DNNs 和 XGBoost，我们遵循(Gorishniy 等人，2021)中提供的设置(包括优化器和超参数空间)，并使用 Optuna 库和网格搜索(仅针对 NODE)进行超参数调整。更详细的超参数信息见附录 B。

比较方法：在我们的实验中，我们将 T2G-FORMER 与代表性的非深度方法 XGBoost 和已知的 DNNs 进行比较，包括 NODE、AutoInt、TabNet、DCNv2、FT-Transformer 和 DANets。一些其他常见的 DNNs，如 MLP 和 SNN(具有 SELU 激活的 MLP 网络)也被纳入比较。

5.2 主要结果和分析

翻译： 性能比较：表 2 报告了 DNNs 和非深度模型的性能。T2G-FORMER 在八个数据集上优于这些 DNNs，并在大多数情况下与 XGBoost 相当。所有模型都通过选择 Optuna 驱动调整的最佳验证结果进行超参数调整。

FR-Graph 类型的影响：我们比较了 GE 中四种类型的 FR-Graphs。表 3 报告了结果，可以看出选择对称边权和非对称知识拓扑通常更好。这表明两个表格特征之间的相互交互可能是相同的，而非对称拓扑提供了更大的语义探索空间，更有可能产生有用的特征。其他数据集的结果见附录 C。

自交互的影响：我们在 GE 中的一个关键设计是"无自交互函数"，它明确排除了 FR-Graphs 中的自环。表 4 报告了几个数据集上没有自环的 FR-Graphs(我们的方法)与有自环的 FR-Graphs 的比较结果。结果表明，在大多数情况下，移除自环并专注于与其他特征的交互略微有利于分类和回归的性能。这可能是因为特征自交互影响了与其他特征交互的概率(因为我们在等式(6)中使用竞争性激

活)，而我们的快捷路径已经保留了自信息。

GEs 的影响：我们探索了在 T2G-FORMER 的不同层包含 GEs 的影响。表 5 报告了仅在模型位置和 GEs 数量上不同的不同模型版本的性能。对于没有 GE 的层，我们使用普通注意力分数进行替代。总体而言，GEs 的位置对回归任务的影响比对分类任务的影响更大。如下所示，在回归任务中，当 GEs 配备在更高层时，模型性能下降更大，而在分类中，与 GE 位置相关的下降似乎不那么显著。此外，仅配备注意力分数的模型比在高层(而非第一层)配备单个 GE 的模型在回归任务中表现更好，但在分类任务中始终是次优的。一个可能的解释是回归需要比分类更平滑的优化空间，因此全连接注意力分数提供了应对连续特征值的交互类型，而高层中的单个 GE 难以捕捉以全连接方式融合的特征之间的明确关系。因此，对于回归来说，完全使用注意力分数比在高层使用单个 GE 更好。第一层中的单个 GE 在回归和分类中都表现出最小的性能下降，这可以通过 GE 在捕捉具有明确语义的表格特征之间的底层关系方面的强度来解释。

总结来说，在任何层中移除 GE 都可能导致性能下降，而通过对所有层应用 GE 获得最佳结果。

拓扑学习方法的比较：除了第 3.1 节中提出的列嵌入方法外，还有一些其他直观的获取 RF-Graph 知识拓扑的直接方法，例如，直接对自适应边权执行阈值剪裁(我们称之为"自适应拓扑")或学习一个 N 乘 N 的邻接矩阵(我们称之为"自由拓扑")。具体来说，对于学习自适应拓扑，我们用等式(2)中的 Gw 替换等式(5)中的 Gt。对于学习自由拓扑，我们直接用一个 N 乘 N 矩阵表示 Gt。表 6 报告了这些拓扑学习策略的比较结果。可以看出，整个数据集共享的静态知识拓扑(我们的方法)比自适应拓扑获得更优越的性能，这表明我们在第 1 节中提到的底层知识假设的合理性。此外，完全自由拓扑也取得了较差的性能，这可能是因为给可学习矩阵过多的自由度。

与 DANet 分组交互的比较：如图 1 所示，DANets(Chen 等人，2022)在由"entmax"操作确定的组中交互表格特征。在这里，我们将我们的基于图的交互与该基于组的交互进行比较，以检验 FR-Graph 的好处。具体来说，我们用 DANet 分组选择掩码替换等式(5)中的知识拓扑 A。表 7 中的结果表明，将表格特征组织成图更有益，因为图拓扑能够捕捉关系边并提供比组结构更微妙的交互。

5.3 可解释性

翻译： 在图 3 中，我们可视化了第一层 FR-Graph 和输入特征(即来自特征标记器的特征；见图 2(a))上的读出收集策略。在 CA 数据集上，我们发现区块组内居民的中位收入(MedInc, MI)与家庭成员的平均数量(AveOccup, AvOcc)相关，而 AveOccup 可以影响卧室的平均数量(AveBedrooms, AvB)是合理的。此外，似乎还存在一些关系，如经度(Long)-房龄(HoA)、经度-平均房间数(AvR)和经度-人口(Pop)，这些可能来自数据集偏差。至于读出，可以看到仅收集了 HouseAge，这在房屋价格预测中是一个有意义的特征。在 CH 数据集上，客户的银行余额(Bal)和估计薪资(EstSal)之间，以及客户年龄(Age)和估计薪资之间存在合理的关系。此外，客户的信用评分(CreditScore, CrSc)与客户的年龄和余额高度相关也是可以理解的。读出仅在当前级别收集 Age 来预测客户是否会离开银行，这也很直观。

6. 结论

翻译： 在本文中，我们提出了 T2G-FORMER，一个新的为表格学习量身定制的 Transformer 模型，它具有一个用于基于估计关系图促进异质特征交互的新模块图估计器(GE)。我们以注意力类似的方式将特征关系图适配到 T2G-FORMER 的基本块中，以简化和适用性。在广泛的公共数据集上的实验表明，T2G-FORMER 比各种 DNNs 取得了更好的性能，并且与 XGboost 相当。我们期望我们的 T2G-FORMER 将作为表格学习研究中的强基线，并增强对处理表格数据的特征异质性的研究兴趣。

解析与重点：

- 实验设计全面：
    - 12 个开源表格数据集，包括分类和回归任务
    - 与多种最先进方法比较，包括深度模型和 GBDT
    - 固定的数据预处理和分割方案确保公平比较
- 主要实验发现：

1. 性能优势：T2G-FORMER 在 8 个数据集上优于现有 DNNs，与 XGBoost 相当

2. FR-Graph 类型：对称边权+非对称知识拓扑组合效果最佳
    - 解释：特征间交互可能是对称的，而非对称拓扑提供更大

语义探索空间

3. 自交互效果：移除自环有助于性能提升
    - 解释：排除自环使模型更关注特征间交互，快捷路径已保留自信息
4. GE 的影响：
    - 回归任务对 GE 位置更敏感
    - 第一层中的 GE 最关键，能捕获基础特征关系
    - 所有层都使用 GE 效果最佳
5. 拓扑学习方法：
    - 静态共享知识拓扑(本文方法)优于自适应或自由拓扑
    - 验证了底层知识假设的合理性
6. 与分组策略比较：
    - 图结构优于 DANet 的分组方法
    - 图拓扑能捕获更细微的关系

- 可解释性分析：
    - 可视化 FR-Graph 展示了合理的特征关联
    - 在 CA 数据集上发现的关系(如收入与家庭规模)符合常识
    - 读出模块选择的关键特征(如房龄、客户年龄)对预测任务具有实际意义
- 研究贡献总结：

1. 创新的图估计器(GE)模块解决表格数据特征异质性问题
2. 将关系图与 Transformer 架构有效结合
3. 实验验证了方法的有效性和适用性
4. 为表格学习研究提供了新的方向和强基线

- 局限与未来方向：

- 可能需要进一步改进拓扑学习方法
- 针对不同类型任务的 FR-Graph 构建策略可进一步优化
- 对更大规模数据的适应性有待探索