CARTE：表格学习的预训练和迁移

金明俊[1]莱奥·格林斯坦[1]加埃尔·瓦罗夸克斯[1 2]

抽象的

预训练的深度学习模型是图像或文本的首选解决方案。然而，对于表格数据，标准仍然是训练基于树的模型。事实上，表格上的迁移学习面临着以下挑战：数据集成：fi找到对应关系，对应关系在条目中（实体匹配不同的词可能表示相同的实体，跨列对应关系 （模式匹配），它们的顺序、名称可能有所不同.....我们提出了一种不需要这种对应关系的神经架构。因此，我们可以在未匹配的背景数据上对其进行预训练。该架构——用于表格条目上下文感知表示的 CARTE——使用表格（或关系）数据的图形表示来处理具有不同列的表格，使用条目和列名的字符串嵌入来模拟开放词汇表，并使用图形注意力网络将条目与列名和相邻条目情境化。广泛的基准测试表明，CARTE 有助于学习，其表现优于包括最佳树型模型在内的一组坚实的基线。CARTE 还支持在具有不匹配列的表之间进行联合学习，从而用更大的表增强小表。CARTE 为表格数据的大型预训练模型打开了大门。

1. 简介

预训练模型的广泛应用极大地促进了各种数据模式的机器学习，例如图像（西蒙尼扬和齐瑟曼，2015）或文本（Devlin 等人，2019）。这些模型可以从模型中心下载，从而启动大量隐式信息和转换，即使在小型数据集上也能释放深度学习的力量。这种范式导致了基础模型的革命（Bommasani 等人，2021）例如大型语言模型（Touvron 等人，2023)。尽管表格数据对企业和机构数据具有重要意义，但这种革命尚未在表格数据上发生。一个障碍是，集成来自不同表格的数据通常很困难，有时甚至是不可能的。不同的表格可能没有任何相关数据，而当它们有相关数据时，数据集成就是整个数据库研究领域（Doan 等人，2012）。可能需要解决跨列对应关系

– 模式匹配 –或跨具有不同条目命名约定的数据源 -实体匹配。由于缺乏匹配的模式和实体，无法在野外进行跨表预训练。没有预训练，深度学习就不太实用，而基于树的方法通常是更好的选择（Grinsztajn 等人，2022）。

这里我们介绍了一种跨表学习的学习架构，无需模式和字符串匹配。关键是用图形表示表，用嵌入表示所有符号（用于列名和表条目）。该架构被称为 CARTE（表条目的上下文感知表示），在大型知识库上进行预训练，以捕获大量实体和关系的信息。然后可以针对给定的下游任务对其进行微调，即使在少数场景下也能帮助学习。它还可以用于跨多个表的联合学习，丰富具有弱相关源的目标表。CARTE 带来了相当大的性能提升，明显优于一组 42 个坚实的基线（包括最佳的基于树的方法和各种特征工程）。它特别有利于具有字符串条目的表，这些表在应用程序中很常见，但在机器学习基准中很少出现。

部分2介绍相关工作；第3节描述CARTE架构和训练程序；以及第 4 节 对许多表格数据集进行了广泛的实证研究，对单个下游表以及多个相关表的设置进行了基准测试。

2.相关工作

表格深度学习表格是许多应用程序的核心。因此，已经提出了许多针对此模式的深度学习方法（Abutbul 等人。，2020； 阿里克与菲斯特，2020；Popov 等人，2019；Gorishniy 等人， 2023b；Somepalli 等人，2021）然而，他们通常

[1]SODA 团队，Inria Saclay，法国[2]Probabl.ai，法国。联系人：Myung Jun Kim < myung.kim@inria.fr >。

会议纪要41英石国际机器学习会议，奥地利维也纳。PMLR 235，2024 年。

基于树的方法表现不佳（Grinsztajn 等人，2022； 施瓦茨-齐夫和阿蒙，2021；加德纳等人，2022）。 尽管 McElfresh 等人（2023）认为神经网络在某些类型的表格上表现良好，并且有前景的架构不断被发表（ Gorishniy 等人，2023年；陈等人，2023），与基于树的方法相比改进的难度表明深度学习必须带来更多的东西，例如背景知识。

表格数据的迁移学习迁移学习主要关注具有相同特征的数据集之间的“常规”迁移设置，IE列。 Somepalli 等人（2021）演示了在更大的未标记表格版本上进行预训练，而Levin 等人（2023) 认为，当数据点很少但相关数据集很大时，迁移学习可以弥补深度学习和基于树的模型之间的差距，就像在医疗环境中一样。它们考虑下游表中的新功能或缺失功能，但要求大多数功能之间精确匹配。

XTab（朱等人，2023) 可以使用数据特定的特征器处理具有不同列的表，这些特征器将实例映射到同一维度，然后在公共块上进行联合学习。然而，它们的表现并不优于基于树的模型（CatBoost，Dorogush 等人，2018). 转表 (王和孙，2022）还将每行表格矢量化到嵌入空间中，以跨表进行学习，证明了跨多个临床试验的数据积累优于包括 XGBoost 在内的基线（陈和格斯特林，2016）这些方法受益于子主题中的一组表，但目前尚不清楚它们是否可以适用于为广泛的应用程序构建预训练模型。

表格数据的预训练模型表 PFN (Hollmann 等人，2023) 在表格学习的预训练模型方面取得了进展：它使用在大量合成数据上预训练的转换器模型来捕捉表格数据的归纳偏差，从而在小型数据集上表现出色，尽管它没有专门处理分类列，但这是树木历来大放异彩的表格的挑战。大型语言模型 (LLM) 也可以用作表格数据的预训练模型。在 TabLLM (Hegselmann 等人，2023)，表格数据被表示为一组标记，这些标记可用于微调 LLM。然而，与树或 TabPFN 相比，LLM 中处理数值的难度使它们成为次优选择。

离散条目表格的一个挑战（数据库文献比机器学习文献更多地解决）是许多条目都是离散的，以字符串表示。塞尔达和瓦罗夸克斯（2022）创建了基于字符串的表示形式，以方便学习。 表向量器在斯克鲁布（2024）启发式地使用这些将不同类型的列转换为数值矩阵非常适合学习。KEN（Cvetkov-Iliev 等人， 2023）是另一种嵌入表实体的方法，更接近我们的预训练目标。它提供了知识图谱中所有实体的嵌入，捕获维基百科等来源中的信息。这些嵌入有助于学习，但挑战在于，列中的每个条目都必须链接到维基百科条目，实体匹配任务。

数据集成传统的统计模型需要将数据汇总到单一一致的表中，这是数据集成文献所解决的任务（Doan 等人，2012）。查找跨数据源列之间的对应关系称为模式匹配。实体匹配这是数据集成和自然语言处理 (NLP) 中常见的挑战，其中字符串必须链接到实体：一个独特的概念。例如，“Davinci”可能表示历史人物“Leonardo da Vinci”，也可能表示 OpenAI“Text-Davinci-003”GPT3 API。实体匹配必须对字符串变化具有鲁棒性，但最重要的是，它必须考虑实体出现的上下文，以消除潜在匹配的歧义。在 NLP 中，预训练的基于注意力的模型，例如 BERT（Devlin 等人，2019)，对于捕捉相应的上下文至关重要。

这些预训练语言模型在表格上也很有用，可以用很少的手动提供的示例来自动执行数据规范化和集成任务（ Narayan 等人，2022）深度学习以及最近基于注意力的模型正在构建数据库、例如，列类型，实体链接（ Hulsebos 等人，2019；邓等人，2020）。

我们感兴趣的是另一个问题：在实体或模式层面，我们的目标是只捕获隐式数据结构和集成，以增强下游机器学习任务，而无需任何手动操作（例如查找相关来源），而不是显式匹配。这个问题很及时：有同样愿景的研究人员正在汇编大量表格（Hulsebos 等人，2023；Eggert 等人，2023）然而，由于大多数可用表格的规模较小，并且表格数据集之间的差异性较大，因此这一愿景迄今为止仍难以实现。

3. 跨表学习的 CARTE 模型

CARTE 跨表学习的能力源于两个要素的结合：一种用图形表示表格实体的新方法，以及一种可以捕获表格中上下文的深度神经网络架构。具体来说，前者可以将多个表格同步到同一个图形域，这使得对以前无法匹配的背景数据进行预训练成为可能。此外，在广泛知识范围内训练的上下文感知深度神经网络可以轻松地将背景信息传播到我们手头的下游任务中。在本节中，我们将介绍 CARTE 及其详细实现。

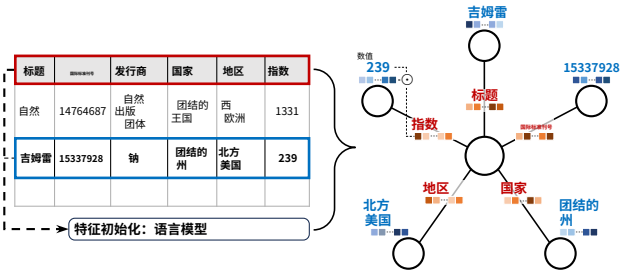

图 1.表格实体的图形表示。从表中，CARTE 将每一行表示为星形图。除缺失值外，叶节点和边缘由单元格值及其对应的列名注释。然后，CARTE 使用语言模型初始化每个节点的特征。数值节点通过元素乘积与其对应的列特征初始化。对于中心节点，它最初设置为小叶的平均值。它稍后将作为捕获图块整体信息的读数。

3.1. 表实体的图形表示

图形表示对于促进表实体的泛化至关重要。一般来说，一个图，克，由节点和边组成，前者表示实体，后者表示节点之间的关系。图对于捕获实体之间的关系信息很有用，而图深度学习对于关系数据库很有前景（Fey 等人，2023）。CARTE 将每一行视为一个小图，如图所示图 1. 从一张表钾CARTE 代表每一列我-第一个例子是图表，格我（X，E），其中组件十和埃分别表示节点和边的特征，嵌入Rd. 结构格我（X，E）是一个星型图，其中 k–p我叶节点，具有页我作为行中缺失值的列数我。在生成的图元上，每个叶节点都由单元格值及其对应的列名注释。为了使这些图元成为神经网络的可行输入，我们初始化十和埃通过使用语言模型。对于分类值和列名，CARTE 只需放置一个d-从语言模型生成的维度嵌入。对于数值，特征通过其值与相应列名的嵌入的乘积来初始化。例如，节点特征十(239)在图 1等于 239×埃（H 指数）最后，用传单的平均值初始化中心节点，稍后将作为捕获图形整体信息的读出节点。

这种设计有几个目的。首先，它代表上下文：对于表格数据，最好根据其列名来解释条目。在图 1例如，仅使用“JMLR”、“15337928”、“239’。列名“Title”、“ISSN”和“H index”表明它是期刊。CARTE 通过节点和边缘表示表中的上下文，并暴露于其神经网络架构。这种表示还可以连接具有不同列顺序或更一般的不同列的表。

其次，CARTE 对非数字条目（例如字符串、类别和名称）使用语言模型。因此，CARTE 中的图形转换不需要对离散条目进行任何干预，这与对字符串使用的典型数据预处理或清理（重复数据删除、分类编码）不同。此外，CARTE 使用的是一组开放的词汇表。对于 CARTE 来说，拼写错误或相同含义的措辞问题（例如将“North America”拼写为“Northern America”）可以轻松解决。

总之，所提出的图形表示的这些特性使跨异构表的泛化成为可能。CARTE 的图形转换将来自不同表的相同图形域实例放入其中，而无需对列进行任何架构匹配或对条目进行实体匹配。因此，学习过程可以跨多个表进行，这为预训练或迁移打开了大门。

3.2. 来自大型知识库的预训练模型

CARTE 在 YAGO3 上进行了预训练（Mahdisoltani 等人，2013YAGO 是一个由 Wikidata 和其他包含现实世界事实的来源构建的大型知识库。YAGO 将信息存储为知识图谱，它是三元组的集合（头，关系，尾）。例如，三元组（卢浮宫，位于，巴黎）从图 2是我们可以在 YAGO 中找到的一个样本。我们当前的 YAGO 版本包含18.1百万个三胞胎6.3百万个实体。

在本小节中，我们描述了 CARTE 的预训练过程，总结如下图 2。从知识图谱中，我们首先提取适合作为 CARTE 输入的实体的小图元。对于具有对比损失的自监督学习，我们将所选图元的截断版本添加到批量中。通过训练过程，CARTE 学会根据给定的上下文聚合信息。

用于预训练的 Graphlet从 YAGO 的大型知识图谱中，我们构建了可用作 CARTE 输入的实体的小图元。为了为实体构建合适的图元，我们首先在用户指定的图元中提取其子图韓-跳跃关系。为了类似于图 1虽然通过多跳获得了额外的信息，但我们设定k =2、 但限制最大数量1-跳跃和2-跳跃关系100和10表格中的图元（图 1）以行的 token 作为中心节点，而知识图谱程序可以使用实体名称（例如“Louvre”）。为了避免差异，我们使用 token 作为中心节点，并添加一个邻居，该邻居由名称作为其节点，以“有名称”作为其节点

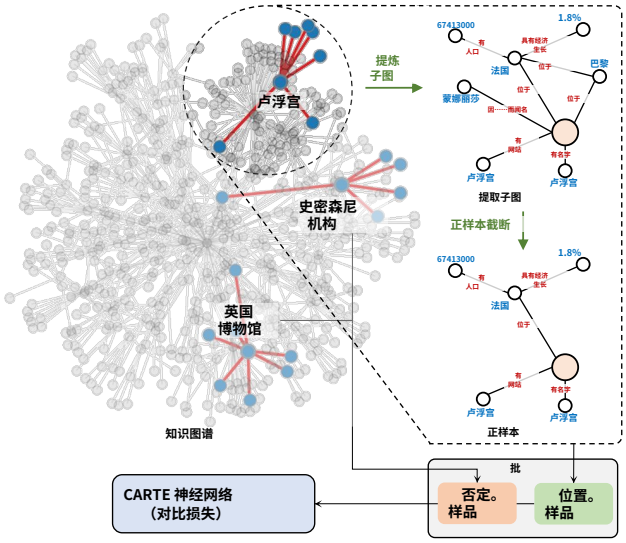

图 2.CARTE 预训练过程。从大型知识图谱开始，CARTE 首先构建图元及其正向变体。然后将提 取的样本输入 CARTE 神经网络，并采用自监督方案进行训练。神经网络学习聚合图元中的信息，这些信息反映了跨列（边）的表格条目组合。

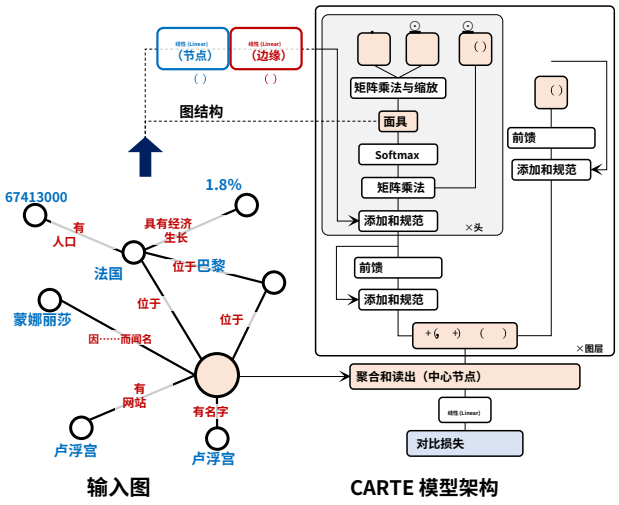

图 3.CARTE 架构CARTE 的输入是包含节点 (十）和边缘（E）特征，均用于自注意力层（显示为灰色）。注意力层使用体现边缘信息的上下文来更新节点特征；输入的图形结构由注意力掩码反映。聚合和读出层由注意力层（没有边缘更新）和中心节点的特征提取组成。然后对输出进行对比损失处理。

关系。最后，第 3.1 节，我们使用 FastText 嵌入初始化节点和边特征（Mikolov 等人， 2017) 作为语言模型。

批量样品构建大小为否b，我们首先选择要包含的 YAGO 实体，生成相应的图元。为此，我们采样0.9的 否b来自具有6或更多1-跳跃关系和其余0.1与其他子集不同。这种采样方案的主要原因是 YAGO 中的大部分实体只有一个或两个1-跳跃关系，而表格数据通常具有更多（更多列）。此外，值 6被选中，以便粗略中值为1-批次样本中的跳跃关系是15.为了实现自监督对比损失，我们加入了正样本，它们只是原始图基元的截断：删除一个随机分数（从0.3到0.7）的边缘。图 2 给出了“卢浮宫”及其正片的示例图。

模型架构图 3描述了 CARTE 的模型结构。CARTE 在其基础上采用了经典的 Transformer 编码器模型Vaswani 等人（2017)，并适应图注意力网络。CARTE 架构中的一个关键组件是自注意力层，它从节点和边缘特征计算注意力。在图模型中，注意力调节邻居对给定感兴趣节点的重要性（Velickovic 等人，2017对于表条目，它表示该条目对于给定实例的重要性，并辅以列信息作为上下文。

现在，我们详细介绍 CARTE 用于捕捉上下文和关系的注意力机制。为了保持一致的符号，我们在向量上写一个箭头答：粗体矩阵一个， 和标量一个。为了便于阅读，我们提出了一个单头注意层，但它可以轻松扩展到连接或平均注意输出的多头方案。

对于具有否节点，让 $\vec{}_{(左)}$ 我 $\in \mathbb{R}^{d}$表示特征节点数我和 $\vec{E}_{(左)伊} \in \mathbb{R}^{d}$边缘指向特征来自节点杰到我。根据设计，CARTE 的图元始终保持中心节点，我们用索引表示我 =1. 注意层的表示是查询、键和值的函数，这些是解释上下文的关键元素。查询是与我们感兴趣的值相对应的向量：节点。因此，我们采用传统方法，仅使用节点信息对其进行参数化。另一方面，键值对应该传达相邻节点可以提供的元素。因此，我们在相应的参数化中添加了边缘信息。考虑到这一点，我们将三个组件设置为1：

询问：
$$问_{我} = \vec{}_{(升我)} \cdot 西_{问} \quad (1)$$

钥匙：
$$凯尔特语_{伊} = (\vec{}_{(左)我} \odot \vec{E}^{(左)}_{伊}) \cdot 西_{钥} \quad (2)$$

价值：
$$V_{伊} = (\vec{}_{(左)我} \odot \vec{E}^{(左)}_{伊}) \cdot 西_{五} \quad (3)$$

在哪里⊙表示元素乘法和西 问，

1这里我们省略了表示层的上标一个伊，埃伊， 和投影权重问，答，为了表达清晰。

西钾， 和西五是可训练的权重，位于Rd×d。这里，元素乘积的选择是从知识图谱嵌入技术中得 到的Balazevic 等人 （2019（英文）： Cvetkov-Iliev 等人（2023这些工作表明，建模关系（IE列名）作为节点向量上的元素乘法效果最好，相比之下例如，向量加法。根据上述三个方程对点积注意力进行缩放，节点的注意力得分杰来自节点我，伊，导出如下：

$$一个_{伊}=磷\frac{指数（埃_{伊}）}{\sum_{钾\in N_{我}}指数（埃_{伊克}）}，在哪里埃_{伊}=\frac{问·凯尔特通_{电视}}{\sqrt{d}}\qquad(4)$$

其中计算一个伊仅对节点的连接邻居求和我。这对应于掩蔽步骤，它采用输入的图形结构。考虑到关系类型（IE列名称）中的注意力得分（E在等式中2和3) 对于正确地重新语境化条目（即捕捉其含义）非常重要。例如，条目“乔治布什”可能表示英石或 43路美国总统、一艘航空母舰……这种模糊性是由以下关系引起的（“乔治·布什”、“乔治·布什之子”、“乔治·布什”），然而，要完全掌握信息，确实需要了解关系的性质，因为“的父亲”会导致不同的实体解析。消融揭示了注意力的重要性（附录C.3）。

对于节点和边，注意层的输出为：

$$\text{转换条目}\quad \vec{}^{(升+1)}_{我}=\sigma+\underset{杰}{\text{X}}\ \vec{}_{一个_{伊}\cdot 五_{伊}}$$

$$\text{变换关系}\quad E^{(升+1)}_{伊}=\sigma_{埃（英语）}{}^{(}{}_{伊}^{十)}\cdot 西_{埃})$$

其中 σ 表示相应的连续操作（参见 图 3）。最后的层由没有边缘更新的注意层和提取中心节点表示的读出层组成。对于预训练，然后对输出进行对比损失处理。附录A.1详细说明模型规范和训练。

对比损失对于自监督对比损失，我们采用了以下框架陈等人（2020b)。原始图基元和一个截断被设置为正值，而批次中的其他图基元被视为负值。然后，学习损失基于网络输出的余弦相似度，输入 InfoNCE 损失（Oord 等人，2018）。

3.3. 针对下游任务进行微调

对于给定的下游任务，微调 CARTE 只需重用部分预训练架构即可（如图所示）图 3）：节点和边缘的初始层（蓝色和红色块）和“聚合和读出”层。虽然这种简化与许多微调方法不同，但它源于图神经网络的行为。事实上，下游表实体形成更简单的

比预训练期间的图更容易理解。首先，它们是星形的（图 1）。其次，与 YAGO 相比，下游表的图结构变化较少，离散变量的基数也较少。架构太深可能会冲淡输出表示中的判别特征（过度平滑问题，陈等人，2020 年； Rusch 等人，2023，见附录C.4因此，我们使用图模型中常见的惯例：将注意力层的数量设置为最大值韓-跳跃关系，这里 k =1.对于最终的分类器，我们只需附加线性层。使用基础模型进行微调，我们考虑两种不同的下游推理设置。

单表推理这是一个众所周知的设置，其中我们给出了一个包含要预测的目标变量的表。在将表实体转换为图形之前，我们使用幂变换 (杨致远，2000幂变换已在多项研究中被证实是有效的（例如，Hollmann 等人，2023； Cvetkov-Iliev 等人，2023)，同样，为 CARTE 的微调过程提供了稳定性。此外，我们采用了 bagging 策略（布雷曼，1996)，其中基于用于早期停止的不同训练-验证分割来训练不同的模型。预测输出计算为每个模型输出的平均值。

从一个源表传输到目标表我们还在迁移学习设置中使用 CARTE，其中我们给出了源表十年代可以帮助我们预测目标表 十电视。重要的是，源表的训练样本可能比目标表大。我们对两个表联合微调 CARTE。图形表示可以实现这种联合微调，而无需列中的对应关系；然而，我们确实需要有相似的结果是年代和是电视在两个表中。源结果是年代被转换为与目标结果的第一时刻相匹配是电视使用如上所述的幂变换（杨致远，2000，请注意，这里我们使用逆变换）。如果目标表和源表在结果的分类/回归性质上有所不同，我们会进行调整是年代如下所示：对于分类目标是电视， 我们对源表中的回归结果进行二元化，对于回归目标是电视，我们使用源表的二元分类结果，编码为{0，1}和标准缩放。然后，我们继续通过从目标表和源表中抽取具有固定比例行的批次来微调 CARTE（我们使用的批次大小为 64，其中 8 个来自目标表）。我们在目标表的验证集上使用早期停止，并且仍然依赖于在不同的随机验证集上构建多个学习器并平均预测的 bagging 策略。通常，在覆盖源的所有数据点之前，早期停止就开始了。这可以防止过度拟合源数据，这可能不如预测的目标数据那么重要是电视。我们使用在单表设置中选择的超参数。

由于我们从弱相关数据中非常松散地选择源表，如果源表没有提供足够的相关信息，则生成的成对学习器实际上可能不会比单表学习器有所改进。因此，我们将成对学习器与单表学习器结合起来，通过将它们的输出与 softmax 相结合来集成预测器。softmax 的权重是使用这些学习器的内部验证集中计算出的预测分数来计算的，但除以学习器之间的标准差来设置 softmax 的温度。

跨多表联合学习如上所述，迁移学习的关键是找到正确的源表。如果我们有来自给定域或机构的多个表，CARTE 可以调整以使用它们，找到最有用的迁移信息。在这些设置中，我们得到了一个目标表十电视和一组源表十年代,1...X山,米。我们继续构建个体学习器：首先是十电视，然后每对[2]一个源表 十,我和目标表十电视使用上面描述的成对联合学习器。同样，并非每次成对学习都能带来相同数量的有用信息。因此，为了找到数据集的最佳组合，我们使用与上述相同的策略，即集成所有成对预测器以及单表预测器。因此，如果所有源表都导致预测器也能正常工作，则它们将以相同的权重组合在一起，但如果一个源占主导地位，则预测将基于此源。

4.实验研究

4.1. 实验装置

数据集我们使用 51 个表格学习数据集，所有数据集都与一个学习任务相关 - 40 个回归和 11 个分类 - 这些数据集来自多个来源，主要来自之前的机器学习研究和 kaggle 竞赛。它们涵盖了社会和商业的各种主题：事故、选举、薪酬、食品、餐馆等。我们选择代表现代数据科学应用的数据集：具有有意义的列和离散条目的表格（表 3），与许多数据集不同美国自盟。 附录乙给出数据集的具体列表。

基线我们用以下缩写来评估不同的基线（实验设置和超参数调整的具体细节见附录A.2）：

CatBoost (Dorogush 等人，2018）梯度提升树
: 包通常用于表格学习。我们将文本特征视为分类特征，由 CatBoost 的分类编码进行编码，这是目标的改进版本编码（米奇・巴雷卡，2001）。

向量表Skrub 包中的 TableVectorizer
: （斯克鲁布，2024) 将包含字符串条目的表编码为数值数组。基数低的列（类别数）采用独热编码，基数高的列采用 skrub 中的 Gamma-Poisson 编码器进行编码，该编码器于塞尔达和瓦罗夸克斯（2022），从子字符串中提取潜在类别。对于非基于树的模型，缺失值用平均值来填补数值特征的缺失值，并作为分类特征的另一个类别处理。对于神经网络模型，应用 minmax 尺度来设置介于零和一之间的值。

XGB、HGB 和 RF基于树的模型：XGBoost（陈和格斯特林，2016)、HistGradientBoosting 和 Random-Forest（来自 scikit-learn，Pedregosa 等人，2011）。

MLP 和 ResNet经典的多层感知器
: （MLP）及其具有附加 layernorm/batchnorm 和 skip-connections 的扩展（ResNet）。

里奇和物流线性模型、Ridge 和 Logistic 重新
: 回归和分类任务的渐进性。

法学硕士受到 TabLLM 的启发（Hegselmann 等人，2023)，
: 我们研究使用大型语言模型 (LLM) 对表格的每一行进行编码。我们将每一行表示为一个句子，并使用 intfloat/e5-small-v2 (王等人，2022) 来自 HuggingFace。然而，与 TabLLM 不同的是，编码表被传递给 XGB 估计器，以便进行回归和分类的学习。对于数字条目，我们研究将它们连接为 LLM 之外的附加特征（中国), 或者将它们作为字符串传递给 LLM（英语）。

表 PFN (Hollmann 等人，2023）是一个变压器模型
: 在合成数据上进行预训练，在一次前向传递中生成对新（小）数据集的预测。我们将文本特征视为分类特征，并使用目标编码器对其进行编码（米奇・巴雷卡，2001）。

4.2. 单表结果

CARTE 在单桌学习方面的表现优于其他方法图 4比较了多种方法的预测性能，总结了不同的数据集。我们发现，无论是采用标准化分数，CARTE 还是其他方法，在不同样本量下的表现都始终优于其他方法[3]或基于 Friedman 检验后的 Conover 事后检验的临界差异图来检测成对显著性（康诺弗，1999）另一个重点是，CARTE 所采用的 bagging 策略对神经网络模型也有积极影响：这种 bagging

[2]为了限制计算成本，我们不探索源表的完整组合。

[3]归一化分数的计算改编自 Grinsztajn 等人（2022)，其中最低分数固定为 ρ 值：ρ =0对于回归和 ρ =0.5进行分类。

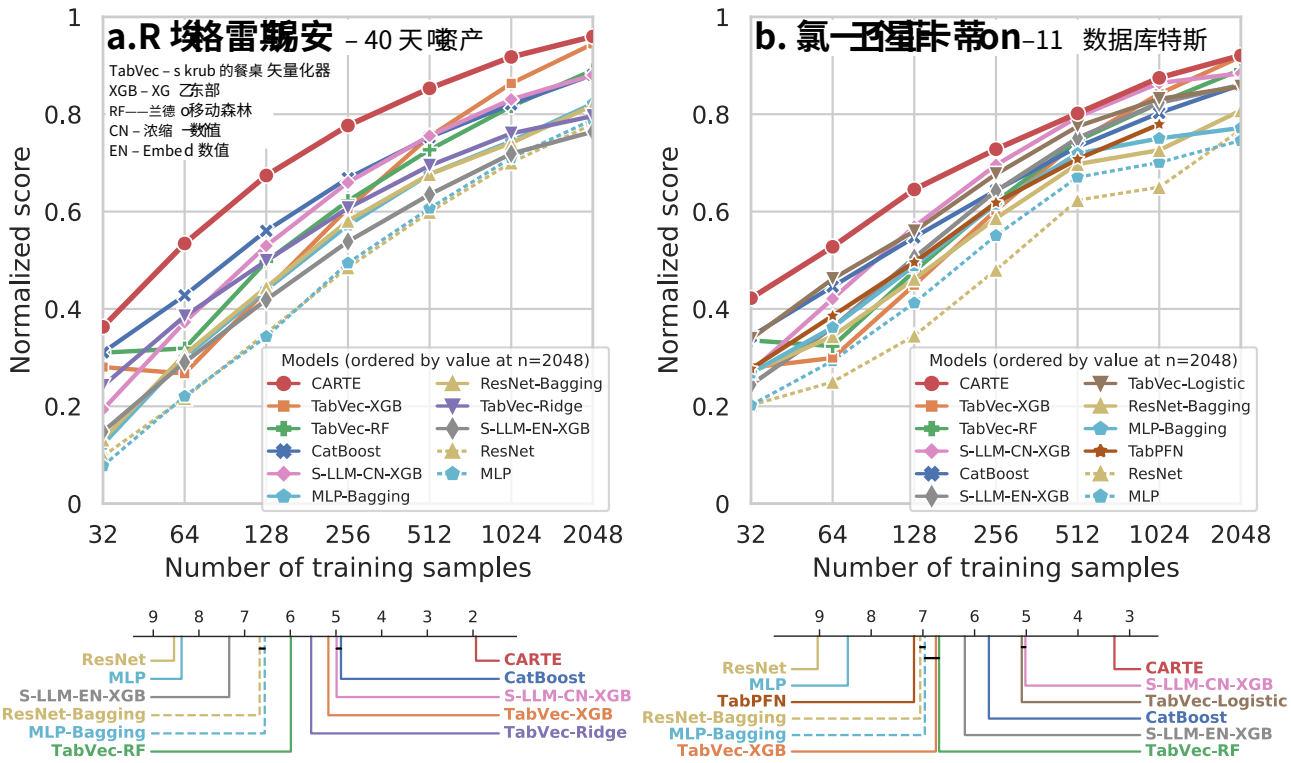

图 4.CARTE 在单桌上学习时表现最佳。(a) 回归和 (b) 分类任务的学习曲线。顶部：标准化分数（1 表示在所有方法和数据集的训练规模中表现最佳，0 表示最差），跨数据集取平均值。底部：关键差异图（特尔皮洛夫斯基，2019) 适用于所有型号的列车。图 9给出所有研究方法的临界差异图。

为提前停止而采用不同的训练/验证分割可能对深度学习总体上是有益的。附录C.1 给出了 CARTE 和 42 条基线的综合结果。

CARTE 对于缺失值具有很强的鲁棒性。处理缺失值时，CARTE 会在图形构建步骤中丢弃具有缺失值的列。例如，在具有 10 列的表中，一个具有一个缺失值的数据点在图形构建步骤之后将具有 9 个叶节点。表 1比较绩效下降百分比及其标准化分数（与图 4) 和几个决策树基线，这些基线本身可以处理缺失值。在实验中，我们随机删除每个样本（包括训练/测试）的一定比例的特征。删除列的比例设置为 0.1（删除 10% 的特征）和 0.3。表格显示，CARTE 的表现继续优于基线，缺失值造成的性能下降较小。

表 1.CARTE 对于缺失值具有很强的鲁棒性。性能下降百分比及其缺失值的归一化分数，其中一定比例（0.1 或 0.3）的特征被随机删除。

缺失值造成的百分比减少

方法	列车规模（缺失分数）			
	64（0.1）	64（0.3）	512（0.1）	512（0.3）
菜單	13.28%	38.35%	10.19%	24.42%
CatBoost	21.70%	53.32%	12.23%	29.70%
TabVec-XGB	15.11%	51.27%	12.61%	30.35%
TabVec-RF	7.68%	44.43%	12.77%	29.79%

归一化绝对得分（如同图 4）

菜單	0.44(0.20)	0.29(0.18)	0.75(0.12)	0.61(0.14)
CatBoost	0.31(0.22)	0.17(0.17)	0.65(0.15)	0.50(0.15)
TabVec-XGB	0.19(0.20)	0.11(0.15)	0.65(0.17)	0.50(0.17)
TabVec-RF	0.23(0.21)	0.14(0.16)	0.63(0.15)	0.49(0.15)

CARTE 的计算时间权衡表 2显示了前四条基线的平均计算时间（以秒为单位）。CARTE 的强大预测性能是以计算时间为代价的，并且差距随着训练规模的扩大而增大（n).这个成本需要进一步优化。

CARTE 无需实体匹配CARTE 之所以表现良好，其原因在于它已在 YAGO 知识库上进行了预训练，因此集成了许多实体的信息。然而，在给定的下游表中，实体可能以

表 2.前四条基线的计算时间（以秒为单位） 对 51 个数据集进行预处理、训练和测试。

方法	预处理	学习 n =64	学习 n =512
菜單	50.20±63.68	85.43±60.30	315.49±119.84
CatBoost	-	0.98±1.19	1.05±1.06
TabVec-XGB	64.72±139.23	0.40±0.21	1.19±0.94
法学硕士	207.87±361.56	0.87±0.71	3.49±1.79

不同的方式：例如用“Londres”代替“London”。这就引出了一个问题：如果实体的字符串表示不同，CARTE 是否能很好地传输有用的信息。例如，当使用条目的特征向量嵌入（如 KEN 嵌入）时，字符串匹配是必要的（ Cvetkov-Iliev 等人，2023）。

在我们的四个数据集（公司员工、电影、美国事故和美国大选）中，我们对条目与其对应的 YAGO 实体进行了手动实体匹配。 图 5表明虽然使用 KEN 需要实体匹配才能获得良好的性能，但 CARTE 中的字符串级建模使其性能对实体变体具有鲁棒性：使用手动匹配的实体或原始条目。消融证实了字符串级表示的重要性，它也能捕获语义相似性（附录图 10）。

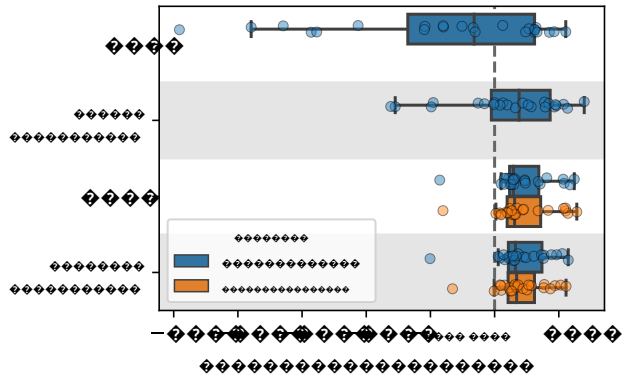

图 5.CARTE 不需要实体匹配，并且下游实体不需要在 YAGO 中。我们在完整数据集或与 YAGO 中存在的实体相对应的数据集的简化版本上评估 CARTE 和 KEN。此外，当实体存在于 YAGO 中时，我们要么将它们与 YAGO 中的规范名称匹配（蓝色），要么保留原始名称（橙色）。当使用 KEN 来丰富数据集时，CatBoost 用作估计器，未匹配的实体将替换为缺失值。图上的每个点都对应于没有任何丰富的 Catboost 的性能提升。KEN 为 YAGO 实体上的 CatBoost 带来了性能提升，证实了背景信息的附加价值。附录

C.5给出了详细的结果。

表 3.本研究的基准和 TabLLM 数据集之间的差异。我们的基准数据集包含更多分类列，特别是基数较高的列（|的 C|）。

特征	本研究的基准	法学硕士
数值列的分数。	0.194	0.613
带有 | 的列的分数頁面| >10	0.625	0.043
数据大小上的基数	0.263	0.001

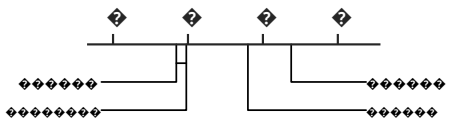

图 6.与 TabLLM 中的三个基线进行比较（Hegselmann 等人，2023）。数据集主要包含数值特征或低基数分类列。在这样的设置下，TabPFN 表现最佳，其次是 CARTE、XGBoost 和 TabLLM。

与 TabLLM 基线的比较我们将 CARTE 与 TabLLM 中提出的九个数据集上的三个基线进行比较（Hegselmann 等人，2023）与 图 4，TabLLM 中的数据集包含更高比例的数值特征，而分类列的基数较少（参见表 3）。图 6给出了基线的临界差异图。TabPFN 在这样的设置中表现出色。然而，尽管 CARTE 适合处理数值和字符串，但它可以获得具有竞争力的性能。详细结果分析见附录C.2。

4.3. 跨表学习

我们研究跨多个表的学习，而没有列之间的明确对应关系。我们使用“野生”表格：在我们的 51 个数据集中，我们发现涵盖相同一般主题（自行车价格、餐厅评级）的组，尽管它们来自不同的来源（附录B.3)。在这种情况下，我们可以轻松使用 CARTE 和 S-LLM 方法，因为它们使用列的开放词汇表示来嵌入条目（但只有 S-LLM 的 EN（嵌入数值）版本，因为 CN（连接数值）需要数值列的对应关系）。由于 CatBoost 本身处理缺失值，我们通过连接数据集并为不匹配的列添加缺失值来使用它。我们还研究了列的手动匹配。

CARTE 不需要架构匹配图 7 显示了仅在两个表中进行的迁移学习的结果。我们看到，对于所有方法，迁移学习都可以提供帮助（虚线表示仅在目标表上的学习，如下所示），但只有 CARTE 提供了一致的改进，而无需手动匹配列。

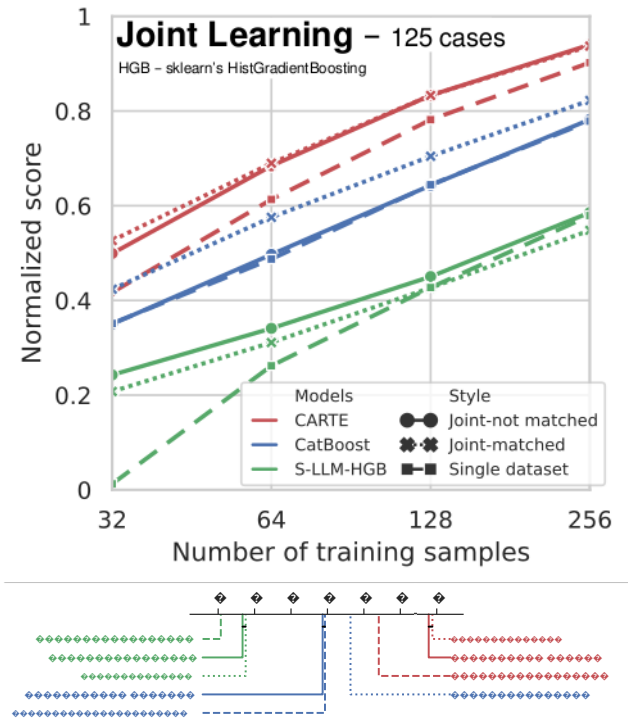

图 7.CARTE 不需要模式匹配，通过联合学习不断改进。我们比较了三种场景（风格）——单一（虚线）：只有目标表；联合（实线）：没有任何手动操作的自动迁移学习；匹配（虚线）：手动匹配列后的迁移学习。

对于 CatBoost，匹配（虚线）至关重要，而其他方法则不然。对于 S-LLM，迁移的好处会随着训练样本数量的增加而迅速下降。结果表明，CARTE 不需要模式匹配，并且它在目标表中通过迁移提供了一致的改进。模式匹配的扩展结果可以在附录中找到C.6。

多表联合学习迁移学习的难点可能在于找到好的源表。在图 8，我们研究引入更多源表，最多可达 4 个表（1 个目标表和 3 个源表，总共 245 个案例）。添加源表对 CARTE 有好处：不仅中位数性能得到改善，而且变异性的下限也得到改善。换句话说，源表越多，找到有益表的机会就越大，因此最坏的情况会变得更好。

5.讨论和结论

表格中的字符串和数字我们的研究涉及表格数据中字符串的重要性。它们在表格机器学习中经常被忽视（表 3展示了如何

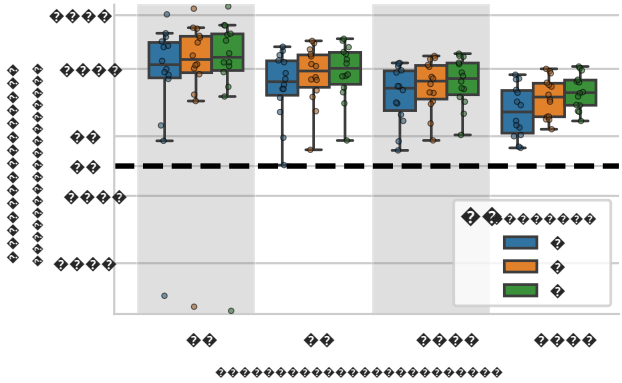

图 8.CARTE 从附加源表中进一步受益。

数据集大多是数字或低基数类别），但对数据库研究至关重要，数据库研究侧重于离散条目。与大多数表格学习模型（无论是基于树的模型，还是 TabPFN 等神经网络）相比，CARTE 的优势在于字符串。skrub 的 TableVectorizer 中的字符串预处理也提升了基线（图 9)。相反，语言模型（如 LLM）专注于字符串，可以在大型语料库上进行预训练。它们对字符串进行了很好的预处理，但必须与基于树的方法相结合才能处理数字（我们的基准测试中为 S-LLM-CN-XGB）。CARTE 是针对字符串和数字量身定制的。

提升性能的架构通过使用对表条目进行建模的体系结构，而不是我⊟CARTE 不是数据矩阵中的特征，而是其上下文（列名和相邻条目）以及字符串的开放词汇嵌入的函数，因此能够对非常不同的表格进行一致表示。这为跨背景表进行预训练以及在没有匹配实体或模式的情况下对下游任务进行微调打开了大门。结果表明，在大型知识库上进行预训练后，生成的模型为下游分析任务带来了显着的好处，无论是在单个表上学习还是在具有不完美对应关系的表上进行迁移学习，其表现都始终优于广泛的基线。它能够从野外的表格中进行迁移，这是迄今为止从未研究过的设置。

走向表格基础模型预训练是深度学习在图像和文本上广泛应用的关键。我们希望 CARTE 背后的理念能够为表格学习带来这些好处，从而产生表格基础模型。这将需要进一步改进架构：针对更大的训练规模进行优化；使用以下理念改进数值表示：Gorishniy 等人（2022）；像大型语言模型一样利用更多的训练信息和表达注意力；与 TabPFN 的互补元学习思想相结合（Hollmann 等人，2023) 在包含大量数字的表格上表现亮眼。

影响声明

本文介绍了旨在推动机器学习领域发展的工作。我们的工作有许多潜在的社会影响，我们认为没有必要在这里特别强调。一种方 法的社会影响取决于如何使用它。然而，我们确实注意到，表格数据对于医疗保健等领域至关重要，这些领域使用大量代码和或多或少标准化的实体（例如，ICD10 疾病代码，医学信息学作为一个领域在数据集成方面投入了巨额资金）。因此，我们希望对健康数据进行表格模型的预训练可以为该领域提供价值，进而产生积极的社会影响。

在另一个话题上，我们注意到，与基线相比，我们的模型 CARTE 需要额外的计算成本。我们确实希望进一步的研究和工程能够降低这些成本。然而，我们的工作为表格数据的预训练模型打开了大门，也许有一天会成为基础模型。这些模型导致了一场规模不断扩大的竞赛，这在能源和财务成本方面带来了可怕的后果，并延伸到生态和权力集中。

致谢

作者感谢法国国家研究署 ANR-20-CHIA-0026 (LearnI) 拨款的部分支持。

我们还要感谢 ICML 的审稿人，他们以良好的方式向我们提出挑战，从而改进了 CARTE 的实证研究，并因此撰写出了更好的手稿。

参考

Abutbul, A.、Elidan, G.、Katzir, L. 和 El-Yaniv, R. DNF-Net：表格数据的神经架构，2020 年 6 月。

Arik, SO 和 Pfister, T. TabNet：Attentive Interpretable 表格学习，2020 年 12 月。

Balazevic, I.、Allen, C. 和 Hospedales, T. 多关系庞加莱图嵌入。神经信息处理系统的进展，32，2019。

Bommasani, R.、Hudson, DA、Adeli, E.、Altman, R.、Arora, S., von Arx, S., Bernstein, MS, Bohg, J., Bosselut, A., Brunskill, E., et al. 论基础模式的机遇与风险。arXiv 预印本 arXiv:2108.07258， 2021 年。

Breiman，L. Bagging 预测器。机器学习，24：123–140，1996年。

Cerda, P. 和 Varoquaux, G. 编码高基数字符串分类变量。IEEE 知识与数据工程学报，34(3):1164–1176, 3月 2022.ISSN 1041-4347、1558-2191、2326-3865。 doi: 10.1109/TKDE.2020.2992529。

陈 D.、林 Y.、李 W.、李 P.、周 J. 和孙 X. 从拓扑角度测量和缓解图神经网络的过度平滑问题。AAAI 人工智能会议论文集， 第 34 卷，第 3438-3445 页，2020a。

Chen, J.、Yan, J.、Chen, DZ 和 Wu, J. Excel前任：在表格数据上超越 GBDT 的神经网络，2023 年 1 月。

Chen, T. 和 Guestrin, C. Xgboost：可扩展的树提升系统。在第 22 届 acm sigkdd 知识发现和数据挖掘国际会议论文集，第 785–794 页，2016 年。

Chen, T.、Kornblith, S.、Norouzi, M. 和 Hinton, G. A 视觉表征对比学习的简单框架。在国际机器学习会议，第 1597-1607 页。PMLR，2020b。

康诺弗，WJ实用非参数统计，体积350.约翰威利父子公司，1999年。

Crossley, S.、Heintz, A.、Choi, JS、Batchelor, J.、Karimi, M. 和 Malatinszky，A. 用于评估文本可读性的大规模语料库。行为研究方法，55(2): 491–507, 2023。

Cvetkov-Iliev, A.、Allauzen, A. 和 Varoquaux, G. Rela-使用背景信息来丰富特征的数据嵌入。机器学习，112(2):687–720, 2023.

Deng, X.、Sun, H.、Lees, A.、Wu, Y. 和 Yu, C. TURL：通过表征学习实现表格理解，2020 年 12 月。

Devlin, J.、Chang, M.-W.、Lee, K. 和 Toutanova, K. BERT：用于语言理解的深度双向变压器的预训练，2019 年 5 月。

Doan, A.、Halevy, A. 和 Ives, Z.数据集成原理授予。爱思唯尔，2012 年。

Dorogush, AV、Ershov, V. 和 Gulin, A. CatBoost：Gra-通过分类特征支持来增强饮食。arXiv 预印本 arXiv:1810.11363，2018 年。

Eggert, G.、Huo, K.、Biven, M. 和 Waugh, J. TabLib：A 包含上下文的 627M 表格数据集，2023 年 10 月。

Fey, M.、Hu, W.、Huang, K.、Lensen, JE、Ranjan, R.、Robinson, J.、Ying, R.、You, J. 和 Leskovec, J. 关系深度学习：关系数据库上的图形表示学习。arXiv 预印本 arXiv:2312.04615， 2023 年。

Gardner, J.、Popovic, Z. 和 Schmidt, L. 子群稳健能力生长在树上：一项实证基线调查。神经信息处理系统的进展， 35：9939–9954，2022 年 12 月。

Gorishniy, Y., Rubachev, I., 和 Babenko, A. 关于嵌入表格深度学习中的数值特征。神经信息处理系统的进展，35: 24991–25004, 2022。

Gorishniy，Y.，Rubachev，I.，Kartashev，N.，Shlenskii，D.，Kotelnikov, A. 和 Babenko, A. TabR：表格深度学习将于 2023 年与最近邻相遇，2023 年 10 月 a。

Gorishniy, Y.、Rubachev, I.、Khrulkov, V. 和 Babenko，A.重新审视表格数据的深度学习模型，2023 年 10 月 b。

Grinsztajn, L., Oyallon, E., 和 Varoquaux, G. 为什么树基于模型的表格数据表现仍然优于深度学习？，2022 年 7 月。

Hegselmann, S.、Buendia, A.、Lang, H.、Agrawal, M.、Jiang, X. 和 Sontag, D. TabLLM：使用大型语言模型对表格数据进行小样本分类，2023 年 3 月。

Hollmann, N.、Müller, S.、Eggensperger, K. 和 Hutter，F.TabPFN：一种可在一秒钟内解决小型表格分类问题的 Transformer，2023 年 9 月。

Hulsebos, M.、Hu, K.、Bakker, M.、Zgraggen, E.、Satyanarayan, A.、Kraska, T.、Demiralp, Ç. 和 Hidalgo, C. Sherlock：一种用于语义数据类型检测的深度学习方法。第 25 届 ACM SIGKDD 知识发现与数据挖掘国际会议论文集，第 1500–1508 页，2019 年。

Hulsebos, M.、Demiralp, Ç. 和 Groth, P. GitTables：A 大规模关系表语料库。ACM 数据管理会议纪要，1(1):1–17，2023 年 5 月。ISSN 2836-6573。doi：10.1145/3588710。

Levin, R.、Cherepanova, V.、Schwarzschild, A.、Bansal, A.、Bruss, CB、Goldstein, T.、Wilson, AG 和 Goldblum, M.《使用深度表格模型进行迁移学习》，2023 年 8 月。

Mahdisoltani, F.、Biega, J. 和 Suchanek, FM Yago3：A 来自多语言维基百科的知识库。CIDR， 2013年。

McElfresh, D.、Khandagale, S.、Valverde, J.、C、VP、Feuer, B.、Hegde、C.、Ramakrishnan、G.、Goldblum、M. 和 White、C. 神经网络在何时能在表格数据上胜过提升树？，2023 年 10 月。

Micci-Barreca, D. 一种高预处理方案分类和预测问题中的基数分类属性。ACM SIGKDD 探索简讯，3(1):27–32, 2001。

Mikolov, T.、Grave, E.、Bojanowski, P.、Puhrsch, C. 和 Joulin，A.预训练分布式词汇表征的进展。arXiv 预印本 arXiv:1712.09405，2017年。

Narayan, A.、Chami, I.、Orr, L. 和 Ré, C. Can 基金会模型会操纵你的数据吗？VLDB 捐赠基金论文集，16(4):738–746, 2022.

Oord, A. vd、Li, Y. 和 Vinyals, O. 表示学习与对比预测编码相结合。arXiv 预印本 arXiv:1807.03748，2018 年。

佩德雷戈萨 (F. Pedregosa)、瓦罗夸克斯 (G. Varoquaux)、格拉姆福特 (A. Gramfort)、米歇尔 (V. Michel) Thirion, B.、Grisel, O.、Blondel, M.、Prettenhofer, P.、Weiss, R.、Dubourg, V. 等人。Scikit-learn：使用 Python 进行机器学习。《机器学习研究杂志》，12：2825–2830，2011。

Popov, S.、Morozov, S. 和 Babenko, A. Neural Oblivious 用于表格数据深度学习的决策集成，2019 年 9 月。

Rusch, TK、Bronstein, MM 和 Mishra, S. 一项关于图神经网络中的过度平滑。arXiv 预印本 arXiv:2303.10993，2023 年。

Sanjib, D.、AnHai, D.、Suganthan, P.、Chaitanya, G.、Pradap, K.、Yash、G. 和 Derek、P.《麦哲伦数据存储库》，2023 年。

Shwartz-Ziv, R. 和 Armon, A. 表格数据：深度学习 ing 并不是您所需要的全部，2021 年 11 月。

Simonyan, K. 和 Zisserman, A. 非常深的卷积网络进行大规模图像识别。在国际学习表征会议（ICLR 2015），2015 年。

Skrub。Skrub，为机器学习准备表格。https://skrub-data.org，2024年。

Somepalli，G.、Goldblum，M.、Schwarzschild，A.、Bruss，CB 和 Goldstein、T. SAINT：通过行注意和对比预训练改进表格数据的神经网络，2021 年 6 月。

Terpilowski，M. scikit-posthocs：成对多重比较 son 在 python 中进行测试。《开源软件杂志》，4(36):1169, 2019. doi: 10.21105/joss.01169。

Touvron, H.、Lavril, T.、Izacard, G.、Martinet, X.、Lachaux, M.-A.、Lacroix、T.、Rozière、B.、Goyal、N.、Hambro、E.、Azhar、F. 等。Llama：开放且高效的基础语言模型。arXiv 预印本 arXiv:2302.13971，2023 年。

UCI。加州大学欧文分校机器学习库。https：//archive.ics.uci.edu。

Vaswani, A.、Shazeer, N.、Parmar, N.、Uszkoreit, J.、Jones, L.、Gomez、AN、Kaiser、Ł. 和 Polosukhin、I. 您所需要的只是注意力。神经信息处理系统的进展，30，2017。

Velickovic，P.，Cucurull，G.，卡萨诺瓦，A.，罗梅罗，A.，Lio，P., Bengio, Y. 等人。图注意力网络。统计，1050 (20): 10–48550, 2017年。

王丽、杨娜、黄晓、焦斌、杨丽、Jiang, D.、Majumder, R. 和 Wei, F. 通过弱监督对比预训练实现文本嵌入。arXiv 预印本 arXiv:2212.03533，2022 年。

Wang, Z. 和 Sun, J. Transtab：学习可转移表格跨表的变压器。神经信息处理系统的进展，35：2902–2915，2022年。

Yeo, I.-K. 和 Johnson, RA 一个新权力家族变换来改善正常性或对称性。 生物识别技术，87(4):954–959, 2000年。

Zhu, B.、Shi, X.、Erickson, N.、Li, M.、Karypis, G. 和 Shoaran，M. Xtab：表格转换器的跨表预训练。arXiv 预印本 arXiv:2305.06090，2023 年。

A. 培训详细信息

A.1. CARTE 的预训练模型

模型规范和训练细节模型规范和训练细节主要参考了（Devlin 等人，2019）我们设定12注意层，每个层由以下部分组成：12多头注意力，隐藏维度固定为与输入相同的大小（300）。最终模型包含9.3百万个参数。为了运行预训练，我们选择了128实体加上一个额外的正值，导致批量大小为256.训练总步数为1,000,000,大约涵盖40针对 YAGO 实体的 epoch。我们使用 AdamW 优化器和余弦调度器，左心室$_{分钟}$=5×10^{-6}，左心室$_{最大限度}$=1×10^{-4}以及第一场热身赛10,000步骤。辍学率固定为0.1并使用了gelu激活函数。

A.2. 下游任务实验设置详情

单个表为了评估单个表上基线的性能，我们重点关注了每个表的训练大小有限的设置，范围从32、64、128、256、512、1、024、和2，048；其余数据设置为测试集。为了找到基线的最优超参数，5-折叠交叉验证100除了 CARTE 和 TabPFN 之外，所有比较方法都进行了随机搜索迭代。对于 CARTE，5-折叠交叉验证，但只对学习率进行了网格搜索。对于 TabPFN，我们使用了本文建议的默认值。有关每种方法的超参数空间的详细信息，请参阅下面的超参数调整段落。性能记录在10不同的训练/测试分割，将性能指标设置为R^2回归分数和分类任务的受试者工作曲线下面积（AUROC）。

跨多表联合学习联合学习的实验设置与单表设置几乎相同，除了一些小细节。目标上的训练集数量随32、64、128、256、而为了使结果具有可比性，我们设置了与单表情况相同的分割。在超参数优化方面，CARTE 仅采用从单表情况中获得的最佳值（第3节）。对于其他基线，也进行了相同方案的超参数搜索。

超参数空间超参数调整是使用网格搜索完成的，因为我们只调整学习率，并使用随机搜索完成基线，因为它们带有两个以上的超参数需要调整。XGBoost、HistGradientBoosting、RandomForest、Resnet 和 MLP 基线的超参数空间基于Grinsztajn 等人（2022）；对于 CatBoost，我们遵循 CatBoost 论文中使用的方法（Dorogush 等人，2018）。对于跨多个表的联合学习的基线，我们采用了一个额外的超参数“分数源”，它表示用于训练的源数据分数。表4下面总结了每个估计器的超参数空间。

A.3. 硬件规格

CARTE 的预训练模型是在 GPU 上训练的。对于我们其余的实验，它们在 32 个 CPU 核心上运行，并根据可用性选择硬件。

GPU：NVIDIA V100（32GB 显存）

CPU：AMD EPYC 7742 64 核处理器、AMD EPYC 7702 64 核处理器 (512GB RAM)、Intel(R) Xeon(R) CPU E5-2660 v2，英特尔(R) 至强(R) 金牌 6226R CPU（256GB 内存）

A.4. CARTE 的实施

CARTE 的实施将在https://github.com/soda-inria/carte。

B. 数据描述

B.1. 数据预处理

数据准备过程中仅进行了最少的数据预处理。对于所有数据集，我们排除了仅包含一个唯一值或缺失值超过数据集大小一半的列。

表 4.CARTE 和基线估计量的超参数空间。

方法	参数	网格
菜單	学习率	[2.5, 5, 7.5]× [1埃-4，1埃-3]
CatBoost	最大深度 学习率 装袋温度 升2-leaf 正则化 一个 热最大尺寸 迭代	统一整数 [2, 10] 对数均 匀分布 [1埃-5，1] 制服 [0，1] 对数均匀分布 [1, 10] 统 一整数 [2, 25] 统一整数 [400，1000]
XGBoost	估计量数量 最大深度 学习率 儿童最低体重 子采样 按级别进行抽样 按树进行采样 伽玛 拉姆达 阿尔法	统一整数 [50，1000] 统一 整数 [2, 10] 对数均匀分布 [1埃-5，1] 对数均匀分布 [1，100] 制服 [0.5, 1] 制服 [0.5, 1] 制服 [0.5, 1] 对数 均匀分布 [1埃-8，7] 对数 均匀分布 [1, 4] 对数均匀 分布 [1埃-8，100]
HistGradientBoosting	学习率 最大深度 最大叶节点数 最小样本叶 升2-正则化	对数均匀分布 [1埃-2，10] [没 有任何，2、3、4] 普通整数 [31, 5] 普通整数 [20, 2] 对数均匀分布 [1埃-6 ，1埃3]
随机森林	估计量数量 最大深度 最大功能 最小样本叶 引导 杂质减少量最少	统一整数 [50，250] [没有 任何，2、3、4] [sqrt，sqrt，log2，无，0.1、0.2、0.3、0.4、0.5、0.6、0.7、0.8、0.9] 对数均匀分布 [1.5，50.5] [正确，错误] [0、0.01、0.02、0.05]
残差网络	层数 层大小 隐藏因素 隐藏辍学 残差丢失 学习率 权重衰减 正常化 批次大小	统一整数 [1, 8] 统一整数 [32，512] 统一整数 [1, 3] 制 服 [0，0.5] 制服 [0，0.5] 对 数均匀分布 [1埃-5，1埃-2] 对数均匀 [1埃-8，1埃-2] [batchnorm，layernorm] [16, 32]
多层感知处理器	层数 层大小 辍学 学习率 权重衰减 批次大小	统一整数 [1, 4] 统一整数 [16，1024] 制服 [0，0.5] 对 数均匀分布 [1埃-5，1埃-2] 对数均匀 [1埃-8，1埃-2] [16, 32]
岭回归	解算器 阿尔法	[svd、cholesky、lsqr、 sag] 对数均匀 [1埃-5，100]
逻辑回归	解算器 惩罚 碳	[newton-cg、lbfgs、liblinear] [无，升1，升2, elasticnet] 对数 均匀 [1埃-5，100]
联合学习的基线	源分数	制服 [0，1]

B.2. 数据集

我们对实验研究中使用的数据集提供了详细的描述。

1.动漫星球[4]该数据集包含从 Anime-Planet 网站抓取的动漫信息。任务是预测该动漫在该网站上的平均评分。

2.宝宝反斗城 (Sanjib 等人，2023）[5]从 Babies R Us 网站上抓取的婴儿用品信息。任务是预测婴儿用品的价格。

3.买买宝贝 (Sanjib 等人，2023）[6]从 Buy Buy Baby 网站抓取的婴儿用品信息。任务是预测婴儿用品的价格。

4.啤酒评分[7]该数据集包含来自 934 家不同啤酒厂的 3197 种啤酒的品尝概况和消费者评论。任务是预测不同啤酒的总体评论评分。

5.自行车德科 (Sanjib 等人，2023）[8]来自印度 bikedekho 网站的有关自行车和踏板车的信息。任务是预测自行车的价格。

6.Bikewale (Sanjib 等人，2023）[9]来自印度 bikewale 网站的有关自行车和踏板车的信息。任务是预测自行车的价格。

7.卡德科[10]该数据集包含二手车信息，以及它们在 Cardekho 网站上的标价。任务是预测价格。

8.巧克力棒评级[11]数据集包含可可批次的信息和专家评级。任务是预测评级。

9.清除语料库（Crossley 等人，2023）[12]关于小学生阅读文章摘录的一般信息。任务是预测摘录的可读性。文本特征是书名，而不是摘录。

10.咖啡评级[+=]数据集取自 coffeereview.com，其中包含各种咖啡的信息。任务是预测咖啡的评论评分。

11.公司员工[14]拥有超过1,000员工数量。任务是预测每个公司的员工数量。

12.员工薪酬及费用收入超过75000[15]年收入超过 75,000 美元的员工的薪酬和费用。任务是预测员工的薪酬。

13.员工薪资[16]有关马里兰州蒙哥马利县员工工资的信息。任务是预测员工当前的年薪范围。

14.Fifa22 球员[17]有关 Fifa22 游戏中足球运动员及其能力得分的信息。任务是预测球员的工资。

---

[4]https://www.kaggle.com/datasets/hernan4444/animeplanet-recommendation-database-2020
[5]http://pages.cs.wisc.edu/ anhai/data/784_data/bikes/csv_files/babies_r_us.csv
[6]http://pages.cs.wisc.edu/ anhai/data/784_data/bikes/csv_files/buy_buy_baby.csv
[7]https://www.kaggle.com/datasets/ruthgn/beer-profile-and- ratings-data-set
[8]http://pages.cs.wisc.edu/ anhai/data/784_data/bikes/csv_files/bikedekho.csv
[9]http://pages.cs.wisc.edu/ anhai/data/784_data/bikes/csv_files/bikewale.csv
[10]https://www.kaggle.com/datasets/sukritchatterjee/used-cars-dataset-cardekho
[11]https://www.kaggle.com/datasets/rtatman/chocolate-bar- ratings
[12]https://www.commonlit.org/blog/introducing-the-clear-corpus-an-open-dataset-to-advanceresearch-28ff8cfea84a

[+=]https://www.kaggle.com/datasets/hanifalirsyad/coffee-scrap-coffeereview
[14]https://www.kaggle.com/peopledatalabssf/free-7-million-company-dataset
[15]https://opendata.vancouver.ca/explore/dataset/employee-remuneration-and-expensesearning-over-75000/information/?disjunctive.department&disjunctive.title
[16]https://openml.org/d/42125
[17]https://www.kaggle.com/datasets/joebeachcapital/fifa-players

15.Filmtv 电影[18]从意大利电影评论网站 Filmtv Movies 抓取的电影和评分信息。任务是预测公众对电影的投票。

16.期刊评分JCR科学期刊及其来自期刊引文报告的描述特征。任务是预测期刊的影响因子。

17.期刊评分 SJR来自 Scimago 期刊排名的科学期刊及其描述性特征。任务是预测期刊的 H 指数。

18.日本动漫[19]日本动漫列表及其相关信息。任务是预测动漫的得分。

19.韩剧[20]来自 mydramalist 网站的韩剧列表及其基本信息。任务是预测韩剧的评分。

20.米其林[21]米其林餐厅指南精选的餐厅列表及其他详细信息。任务是预测餐厅的奖项。

21.ML/DS 薪资[22]机器学习和数据科学行业从业人员的工资和基本信息。任务是预测从业人员的工资。

22.电影收入[23]2017 年 7 月或之前上映的电影的元数据。任务是预测票房收入范围。

23.博物馆[24]关于美国博物馆的一般信息。任务是预测博物馆的收入。

24.戏剧家[二十五]从 mydramalist 网站抓取的亚洲电视剧的一般信息。任务是预测亚洲电视剧的收视率。

25.NBA选秀[二十六]有关 1989 年至 2021 年所有 NBA 选秀球员的信息。任务是预测球员的“替换价值”。

26.处方药[二十七]数据包含在加利福尼亚州引入市场的新处方药，其批发采购成本 (WAC) 超过了 Medicare Part D。任务是预测引入时的 WAC。

27.拉面评分[二十八]该数据集包含来自多个国家的各种拉面的评分和特征。任务是预测拉面的评分范围。

28.罗杰・艾伯特[二十九]该数据集包含著名评论家罗杰特・艾伯特的电影评分。任务是预测评分范围。

29.烂番茄（Sanjib 等人，2023）[三十]包含可在烂番茄电影评级网站上找到的电影信息。任务是预测电影的评分值。

30.Spotify[31]有关 Spotify 曲目的一般信息以及一些相关音频特征。任务是预测专辑的受欢迎程度。

[18] https://www.kaggle.com/datasets/stefanoleone992/filmtv-movies-dataset/data
[19] https://www.kaggle.com/datasets/dbdmobile/myanimelist-dataset
[20] https://www.kaggle.com/datasets/noorrizki/top-korean-drama-list-1500
[21] https://www.kaggle.com/datasets/ngshiheng/michelin-guide-restaurants-2021
[22] https://ai-jobs.net/salaries/download/salaries.csv
[23] https://www.kaggle.com/rounakbanik/the-movies-dataset
[24] https://www.kaggle.com/datasets/markusschmitz/museums
[二十五] https://www.kaggle.com/datasets/rajchinagundi/mydramalist-complete-dataset
[二十六] https://www.kaggle.com/datasets/mattop/nba-draft-basketball-player-data-19892021
[二十七] https://data.ca.gov/uk/dataset/prescription-drugs-introduced-to-market
[二十八] https://www.kaggle.com/datasets/ankanhore545/top-ramen- ratings-2022
[二十九] https://github.com/gabrielcs/movie- ratings-prediction
[三十] http://pages.cs.wisc.edu/ anhai/data/784_data/movies1/csv_files/rotten_tomatoes.csv
[31] https://www.kaggle.com/datasets/maharshipandya/-spotify-tracks-dataset

31.美国事故[三十二]2016 年至 2020 年美国城市事故信息。根据该数据集，我们将执行两项任务：(1) 美国城市的事故数量范围 (2) 报告事故的严重程度。

32.美国总统（Cvetkov-Iliev 等人，2023) 2020 年美国总统大选的投票统计数据以及美国各县的信息。任务是预测美国各县的投票数字范围。

33.二手车 24[33]有关二手车的信息。任务是预测价格。

34.意大利二手奔驰汽车[三十四]数据集包含意大利二手车销售信息。任务是预测价格。

35.UsedCars.com包含二手车 usedcars.com 信息的数据集。任务是预测价格。

36.巴基斯坦二手车[三十五]数据集包含巴基斯坦二手车销售信息。任务是预测价格。

37.沙特阿拉伯二手车[三十六]数据集包含来自 Syarah 网站的有关在沙特阿拉伯销售的二手车的信息。任务是预测二手车的价格。

38.电子游戏销售[三十七]该数据集包含销量超过 100,000 份的视频游戏列表（从 vgchartz.com 抓取）。任务是预测视频游戏的全球销量。

39.威士忌酒[三十八]whiskyanalysis.com 提供威士忌的基本信息和品尝信息。任务是预测威士忌的元评论范围。

40.维基利克[三十九]Wikiliq 网站上可以找到的有关酒精的信息。我们进行了两项任务，以预测 (1) 啤酒和 (2) 烈酒的价格。

41.波兰[40]有关波兰市场上的葡萄酒的信息。任务是预测价格。

42.葡萄酒网[41]从 wine.com 网站抓取的葡萄酒信息。我们进行了两项任务，预测 (1) 葡萄酒评级和 (2) 葡萄酒价格。

43.葡萄酒爱好者[四十二]来自 winemag.com 的有关葡萄酒和品酒师的信息。我们进行了两项任务，预测 (1) 葡萄酒评级和 (2) 葡萄酒价格。

44.葡萄酒[43]从 Vivino 网站上抓取的葡萄酒瓶信息。我们进行了两项任务，预测 (1) 葡萄酒评级和 (2) 葡萄酒价格。

45.喊叫[四十四]Yelp Open 学术研究数据集。我们从原始数据集中提取了有关餐厅的信息。任务是预测餐厅的星级范围。https://www.yelp.com/dataset

46.佐马托[四十五]在 zomato 网站上找到的餐厅信息和评论。任务是预测每家餐厅的评分范围。

[三十二] https://smoosavi.org/datasets/us_accidents
[33] https://www.kaggle.com/datasets/avikasliwal/used-cars-price-prediction
[三十四] https://www.kaggle.com/datasets/bogdansorin/second-hand-mercedes-benz-registered-2000-2023-ita
[三十五] https://www.kaggle.com/datasets/mustafaimam/used-car-prices-in-pakistan-2021
[三十六] https://www.kaggle.com/datasets/turkibintalib/saudi-arabia-used-cars-dataset
[三十七] https://www.kaggle.com/datasets/gregorut/videogamesales
[三十八] https://whiskyanalysis.com/index.php/database/
[三十九] https://www.kaggle.com/datasets/limtis/wikiliq-dataset
[40] https://www.kaggle.com/datasets/skamlo/wine-price-on-polish-market
[41] https://www.kaggle.com/datasets/manyregression/updated-wine-enthusiast-review
[四十二] https://www.kaggle.com/datasets/manyregression/updated-wine-enthusiast-review
[43] https://www.kaggle.com/datasets/joshuakalobbowles/vivino-wine-data
[四十四] https://www.yelp.com/dataset
[四十五] https://www.kaggle.com/datasets/anas123siddiqui/zomato-database?select=restaurant.csv

B.3. 来自多表实验的数据集

对于多表实验，我们从上面的列表中提取与相同主题相关的表组：

葡萄酒价格：Wina Poland、WineEnthusiasts、WineVivino、Wine.com

葡萄酒评分：WineEnthusiasts、WineVivino、Wine.com

啤酒：啤酒评级，Wikiliq-啤酒

二手车：二手车 24、意大利奔驰二手车、UsedCars.com、巴基斯坦二手车、沙特阿拉伯二手车

影片：Filmtv 电影，烂番茄

戏剧：韩剧，Mydramalist

动漫：动漫星球，日本动漫

婴儿用品：买买宝贝，宝宝反斗城

自行车销售：Bikedekho、Bikewale

员工薪酬：公司员工，员工薪酬和费用收入超过 75000，ML/DS 薪资

餐厅评分：Zomato、米其林、Yelp

期刊分数：期刊评分 JCR、期刊评分 SJR

C. 扩展结果

C.1. CARTE 与 42 个基线在单表学习上的性能比较

作为结果的扩展第4.2节，图 9显示了 CARTE 与 42 个基线的总体比较，其中还考虑了 Scikit-Learn 的 HistGradientBoosting（血红蛋白（HGB）、分类变量的目标编码（米奇・巴雷卡，2001）（TarEnc)、利用 Fasttext 语言模型的外部信息（金融时报）和 intfloat/e5-small-v2 (王等人，2022）（法学硕士 (LLM)和'装袋'策略。我们看到 CARTE 在回归和分类任务的所有基线中都取得了显著的领先。此外，有趣的是，bagging 策略对神经网络模型有积极影响，而对线性或集成基线（基于树的模型或 TabPFN）的影响有限。这可能暗示，使用不同的训练/验证分割进行 bagging 是其他深度学习架构的重要设置，尤其是对于有限的训练大小。

C.2. TabLLM 数据集的详细结果

表 5展示了数据集规格以及 CARTE 与基线之间的性能比较的详细结果Hegselmann 等人（2023）。数据集通常包含高比例的数值特征（四个数据集）或低基数的分类列（八个数据集）。在这种情况下，TabPFN 往往优于其他方法。然而，对于数据集“bank”，CARTE 优于其他基线。具体来说，该数据集与 51 个包含数值和分类特征的数据集一致，并且后者的基数相对较高。从某种意义上说，CARTE 位于 TabPFN（适合数值特征）和 TabLLM（将信息表示为标记）的中间，具有一个注意力架构，旨在处理数值和字符串。

C.3. CARTE 组件的消融研究

为了研究 CARTE 各个组成部分的作用，我们进行了额外的实验，排除或改变了相关组成部分。图 10显示了在排除或切换组件的情况下，训练规模从 32 增加到 1,024 的学习曲线。对于使用 Minhash 的图形构建，我们使用 Skrub 的 Minhash 编码器更改了特征初始化步骤（斯克鲁布，2024)，它们通过将 MinHash 方法应用于字符串的 n-gram 分解来编码字符串分类特征。图中显示，每个方法对于获得 CARTE 的性能都至关重要。在

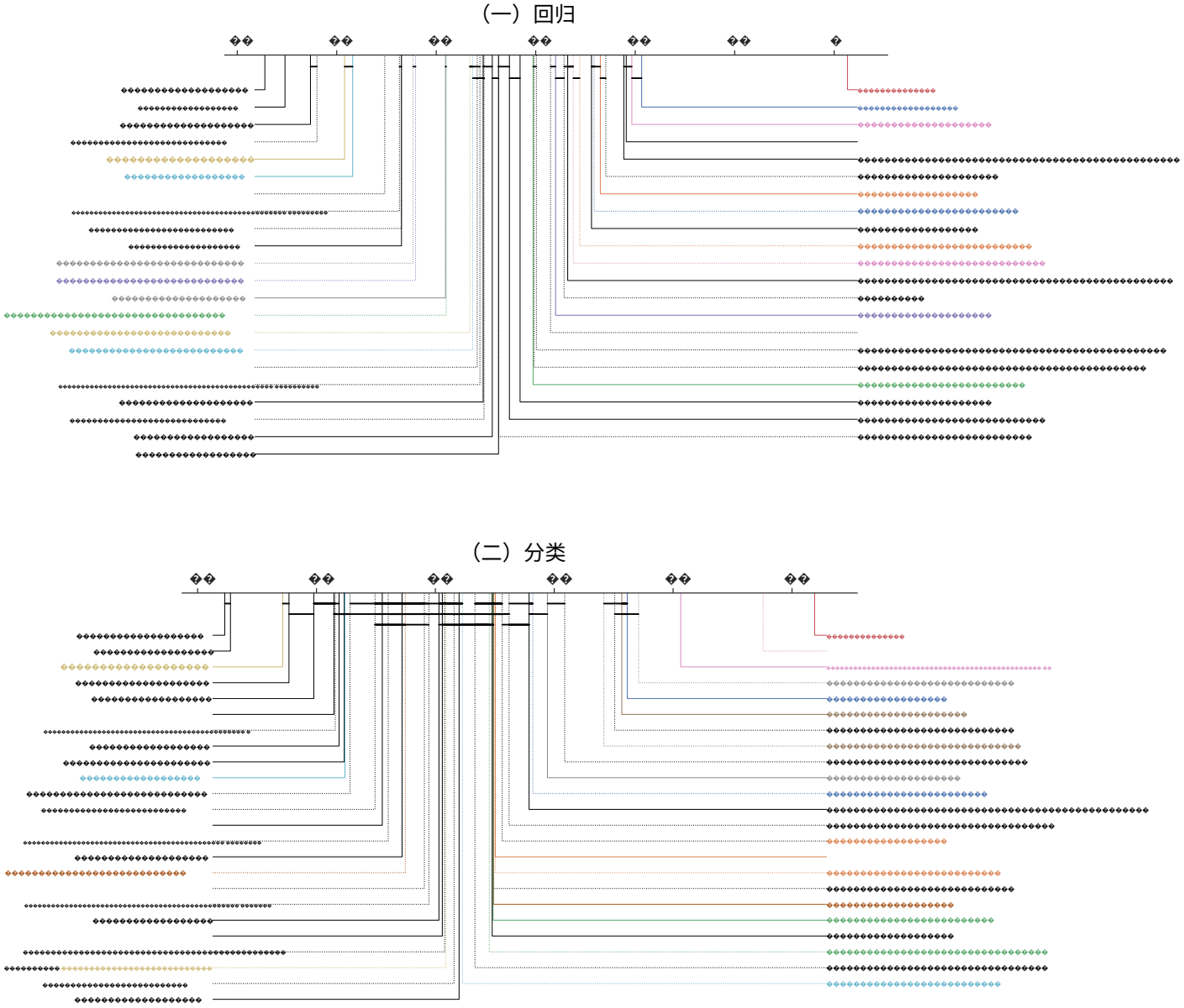

图 9.在单个表格上对 CARTE 与 42 条基线进行比较。CARTE 的关键差异图和（a）回归（b）分类任务的 42 个基线。除了图 4（使用相同的颜色），我们包括了 HistGradientBoosting、目标编码和来自语言模型的外部信息等附加基线。该图显示，CARTE 在回归和分类任务的所有基线中都取得了显著的领先。此外，与线性或集成基线相比，bagging 策略为神经网络带来了更大的好处，这表明它对其他深度学习架构来说是一种重要的设置。

表 5.TabLLM 数据集的详细结果。TabLLM 中提出的 CARTE 与三个基线的数据集规范和性能比较（Hegselmann 等人，2023）。数据集包含大量数值特征或基数较低的分类列。从比较结果可以看出，TabPFN 表现最佳，其次是 CARTE、XGBoost 和 TabLLM。

TabLLM 数据集规范

数据集	分数 数字列	平均的 基数
银行	0.40	38.44
血	1.00	0.00
卡尔霍斯	1.00	0.00
车	0.00	3.50
信用	0.33	3.79
糖尿病	1.00	0.00
心	0.45	2.67
收入	0.33	12.38
丛林	1.00	0.00

性能比较

数据集	方法	拍摄张数				
		三十二	64	128	256	512
银行	菜單	0.81±0.03	0.83±0.03	0.87±0.04	0.89±0.03	0.90±0.01
	XGBoost	0.76±0.03	0.83±0.02	0.85±0.03	0.88±0.01	0.90±0.01
	表PFN	0.76±0.03	0.82±0.03	0.86±0.02	0.89±0.00	0.90±0.00
	法学硕士	0.64±0.06	0.69±0.03	0.82±0.05	0.87±0.01	0.88±0.01
血	菜單	0.68±0.01	0.68±0.01	0.72±0.02	0.72±0.0	0.71±0.01
	XGBoost	0.67±0.06	0.68±0.05	0.71±0.06	0.70±0.07	0.67±0.06
	表PFN	0.70±0.04	0.73±0.04	0.75±0.04	0.76±0.04	0.76±0.03
	法学硕士	0.68±0.04	0.68±0.04	0.68±0.06	0.70±0.08	0.68±0.04
卡尔霍斯	菜單	0.79±0.02	0.83±0.03	0.85±0.04	0.87±0.05	0.89±0.05
	XGBoost	0.79±0.04	0.82±0.04	0.87±0.01	0.90±0.01	0.92±0.01
	表PFN	0.85±0.03	0.89±0.01	0.91±0.01	0.92±0.00	0.93±0.00
	法学硕士	0.77±0.08	0.77±0.04	0.81±0.02	0.83±0.01	0.86±0.02
车	菜單	0.87±0.06	0.94±0.07	0.98±0.03	0.99±0.03	1.00±0.02
	XGBoost	0.82±0.03	0.91±0.02	0.95±0.01	0.98±0.01	0.99±0.01
	表PFN	0.92±0.02	0.97±0.00	0.99±0.01	1.00±0.00	1.00±0.00
	法学硕士	0.91±0.02	0.96±0.02	0.98±0.01	0.99±0.00	1.00±0.00
信用	菜單	0.67±0.03	0.68±0.01	0.70±0.02	0.75±0.01	0.77±0.02
	XGBoost	0.66±0.03	0.67±0.06	0.68±0.02	0.73±0.02	0.75±0.03
	表PFN	0.69±0.07	0.70±0.07	0.72±0.06	0.75±0.04	0.75±0.02
	法学硕士	0.72±0.06	0.70±0.07	0.71±0.07	0.72±0.03	0.72±0.02
糖尿病	菜單	0.76±0.06	0.79±0.02	0.81±0.01	0.82±0.01	0.81±0.00
	XGBoost	0.69±0.08	0.73±0.05	0.78±0.05	0.80±0.03	0.80±0.01
	表PFN	0.77±0.03	0.82±0.03	0.83±0.03	0.83±0.03	0.81±0.02
	法学硕士	0.68±0.04	0.73±0.03	0.79±0.04	0.78±0.02	0.78±0.04
心	菜單	0.90±0.02	0.91±0.02	0.92±0.02	0.93±0.01	0.93±0.01
	XGBoost	0.88±0.04	0.91±0.01	0.91±0.01	0.90±0.01	0.92±0.01
	表PFN	0.91±0.02	0.92±0.02	0.92±0.02	0.92±0.01	0.92±0.02
	法学硕士	0.87±0.06	0.91±0.01	0.90±0.01	0.92±0.01	0.92±0.01
收入	菜單	0.84±0.09	0.84±0.02	0.85±0.03	0.87±0.01	0.88±0.01
	XGBoost	0.79±0.03	0.82±0.02	0.84±0.01	0.87±0.01	0.88±0.00
	表PFN	0.80±0.04	0.82±0.04	0.84±0.01	0.86±0.01	0.87±0.01
	法学硕士	0.84±0.01	0.84±0.02	0.86±0.01	0.87±0.00	0.89±0.01
丛林	菜單	0.71±0.03	0.80±0.02	0.81±0.02	0.86±0.02	0.90±0.02
	XGBoost	0.78±0.03	0.81±0.02	0.84±0.02	0.87±0.01	0.91±0.01
	表PFN	0.78±0.02	0.81±0.01	0.84±0.01	0.88±0.01	0.91±0.00
	法学硕士	0.71±0.02	0.78±0.02	0.81±0.02	0.84±0.01	0.89±0.01

特别地，有趣的是，在排除边缘信息和注意层的情况下，观察到显著的下降。由于这两者对于利用给定表中的上下文都是必不可少的，因此这意味着捕获上下文对于实现强大的预测性能至关重要。此外，CARTE 和 Minhash 之间的性能差距证实，字符串级模型（也捕获语义相似性）很重要，而语言模型的使用对于有效利用外部信息至关重要，尤其是在表中给出的信息有限的情况下。

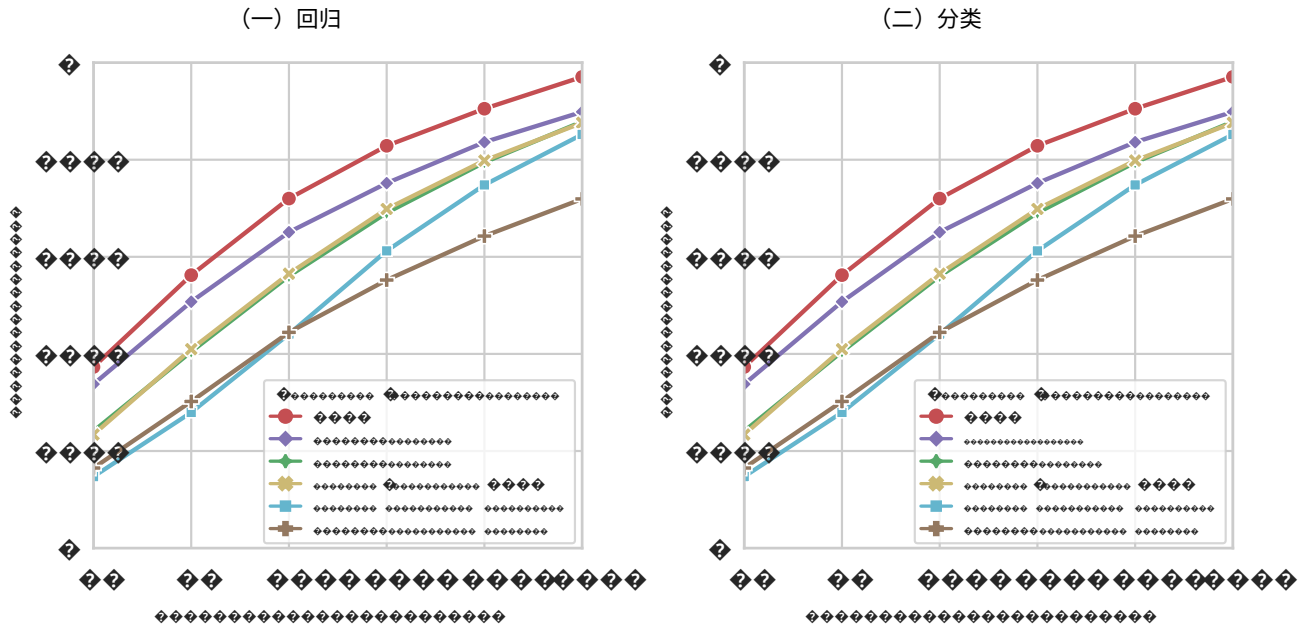

图 10.对 CARTE 的各个组件进行消融。在排除或切换组件的各种情况下，训练规模从 32 到 1,024 的学习曲线。每个对于获得 CARTE 的性能都至关重要。特别是，排除边缘信息和注意层后，性能会显著下降，而边缘信息和注意层对于利用给定表中的上下文至关重要。此外，使用 Minhash（没有任何语义内容的字符串级表示）的结果塞尔达和瓦罗夸克斯，2022)表明语言模型对于有效利用外部信息至关重要。

C.4. 过度平滑的影响

图 11表明从图神经网络 (GNN) 的较深层提取的表示对于下游任务的预测用处较小四十六。我们将其解释为过度平滑的影响，这是 GNN 中一个众所周知的问题（陈等人，2020 年；Rusch 等人，2023）。

四十六在这里，它们在 scikit-learn 的 HistGradientBoosting 预测器中使用。

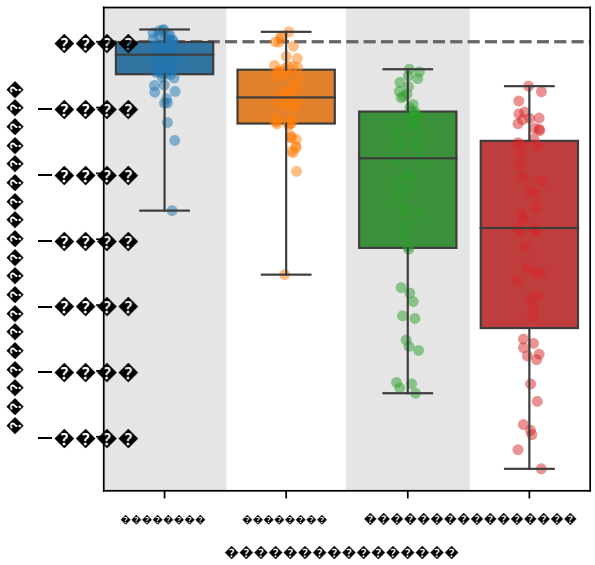

图 11.过度平滑的效果：比较从 2 个 GNN 中提取的表示的预测 nd，4日，8日和 12日层，从第一层开始构建。我们发现，在 GNN 中走得越深，表示对于下游任务的用处就越小。我们将其解释为过度平滑的效果。

C.5. 实体匹配实验的细节

实验采用与单表实验相同的实验设置进行。表6给出实体匹配实验中使用的每个数据集背后的具体结果图 5. 具体数据集为

- CE = 公司员工：32% 的公司与 YAGO 匹配
- MV = 电影收入：84% 的电影与 YAGO 匹配
- US-Acc = 美国事故：67% 的城市与 YAGO 匹配
- US-Elec = 美国选举：98% 的县与 YAGO 匹配

C.6. 跨多表联合学习的模式匹配结果

图 12直接比较了 275 种不同情况下有无模式匹配的 CARTE 的性能（源数据数量从一个源到五个源）。每个点代表给定大小的训练数据对具有不同训练/测试分割的数据集的平均分数的比较。如果点位于对角线下方，则表示 x 轴的性能更高。图中显示点沿对角线对齐，表示有无模式匹配的 CARTE 之间的性能相似（也带有对平均值差异的双侧 t 检验的值为 0.728）。结果表明，在跨多个表传输时，CARTE 不需要模式匹配。

表 6.实体匹配实验详细结果：每个数据集上的单独分数。缩写如下：O-Original 条目、M-Matched 条目、R-Reduced 数据集和 F-Full 数据集。

	CatBoost-MR	CatBoost-MF	CatBoost-OR	CatBoost-OF
CE-32	0.673±0.036	0.672±0.063	0.683±0.052	0.668±0.062
CE-64	0.707±0.021	0.72±0.013	0.702±0.025	0.718±0.017
CE-128	0.734±0.007	0.739±0.024	0.731±0.011	0.745±0.01
CE-256	0.739±0.008	0.747±0.014	0.74±0.009	0.749±0.008
CE-512	0.744±0.005	0.758±0.005	0.744±0.005	0.758±0.004
CE-1024	0.752±0.006	0.763±0.004	0.752±0.006	0.764±0.003
MV-32	0.4±0.058	0.398±0.049	0.394±0.058	0.403±0.042
MV-64	0.436±0.043	0.449±0.045	0.426±0.059	0.453±0.035
MV-128	0.484±0.027	0.492±0.028	0.482±0.018	0.495±0.019
MV-256	0.511±0.011	0.515±0.017	0.51±0.017	0.523±0.012
MV-512	0.545±0.007	0.55±0.007	0.545±0.009	0.552±0.008
MV-1024	0.559±0.007	0.574±0.007	0.563±0.005	0.573±0.005
美国-Acc-32	- 0.016±0.063-0.018±0.064-0.023±0.076-0.02±0.058 0.007±			
美国-Acc-64	0.055-0.01±0.069 0.028±0.036-0.023±0.087			
美国-Acc-128	0.082±0.026	0.055±0.029	0.084±0.028	0.057±0.026
美国-Acc-256	0.129±0.018	0.089±0.031	0.129±0.015	0.089±0.025
美国-Acc-512	0.163±0.02	0.12±0.022	0.163±0.021	0.121±0.02
美国-Acc-1024	0.214±0.007	0.157±0.01	0.217±0.005	0.155±0.009
US-Elec-32	0.31±0.133	0.318±0.142	0.34±0.118	0.285±0.161
美国电气-64	0.433±0.062	0.441±0.068	0.449±0.038	0.445±0.056
US-Elec-128	0.512±0.019	0.505±0.02	0.511±0.02	0.51±0.02
US-Elec-256	0.547±0.009	0.543±0.011	0.546±0.009	0.544±0.011
US-Elec-512	0.572±0.007	0.571±0.01	0.57±0.009	0.571±0.008
US-Elec-1024	0.586±0.004	0.586±0.006	0.586±0.005	0.587±0.005

	卡特-MR	CARTE-MF	卡特-奥	点菜	肯-R	肯-F
CE-32	0.699±0.023	0.69±0.034	0.693±0.023	0.692±0.029	0.518±0.11	0.459±0.119
CE-64	0.729±0.019	0.744±0.025	0.733±0.01	0.747±0.022	0.612±0.077	0.434±0.284
CE-128	0.755±0.007	0.763±0.012	0.755±0.01	0.763±0.014	0.708±0.019	0.409±0.305
CE-256	0.762±0.009	0.776±0.01	0.763±0.008	0.781±0.007	0.738±0.02	0.368±0.812
CE-512	0.773±0.011	0.785±0.007	0.778±0.012	0.789±0.008	0.757±0.006	0.267±0.494
CE-1024	0.783±0.008	0.793±0.006	0.787±0.01	0.798±0.006	0.772±0.008	0.486±0.301
MV-32	0.3±0.057	0.313±0.083	0.329±0.066	0.322±0.095	0.301±0.057	0.318±0.033
MV-64	0.452±0.044	0.471±0.027	0.458±0.025	0.461±0.038	0.42±0.035	0.369±0.055
MV-128	0.521±0.022	0.523±0.022	0.519±0.023	0.515±0.018	0.493±0.024	0.384±0.08
MV-256	0.556±0.022	0.562±0.02	0.554±0.021	0.555±0.014	0.543±0.014	0.464±0.089
MV-512	0.594±0.013	0.597±0.013	0.595±0.014	0.595±0.011	0.589±0.012	0.517±0.04
MV-1024	0.62±0.008	0.622±0.008	0.62±0.008	0.618±0.009	0.616±0.007	0.544±0.032
美国-Acc-32	0.061±0.054	0.053±0.055	0.054±0.067	0.048±0.045	0.062±0.094	0.047±0.045
美国-Acc-64	0.112±0.057	0.114±0.046	0.122±0.056	0.105±0.051	0.146±0.038	0.051±0.09
美国-Acc-128	0.155±0.06	0.136±0.053	0.16±0.058	0.14±0.051	0.175±0.025	0.117±0.026
美国-Acc-256	0.225±0.024	0.197±0.023	0.232±0.02	0.2±0.024	0.225±0.029	0.152±0.014
美国-Acc-512	0.278±0.008	0.237±0.01	0.275±0.015	0.235±0.014	0.27±0.012	0.173±0.029
美国-Acc-1024	0.303±0.01	0.263±0.008	0.304±0.008	0.265±0.012	0.298±0.004	0.205±0.006
US-Elec-32	0.387±0.082	0.387±0.083	0.393±0.08	0.393±0.082	0.149±0.193	0.209±0.159
美国电气-64	0.467±0.031	0.467±0.032	0.465±0.021	0.465±0.022	0.432±0.089	0.454±0.073
US-Elec-128	0.52±0.021	0.52±0.021	0.52±0.022	0.52±0.022	0.564±0.038	0.571±0.035
US-Elec-256	0.552±0.011	0.553±0.011	0.546±0.013	0.546±0.013	0.625±0.019	0.628±0.011
US-Elec-512	0.586±0.008	0.586±0.008	0.58±0.007	0.581±0.007	0.667±0.006	0.664±0.01
US-Elec-1024	0.615±0.007	0.615±0.007	0.612±0.007	0.613±0.007	0.7±0.009	0.697±0.005

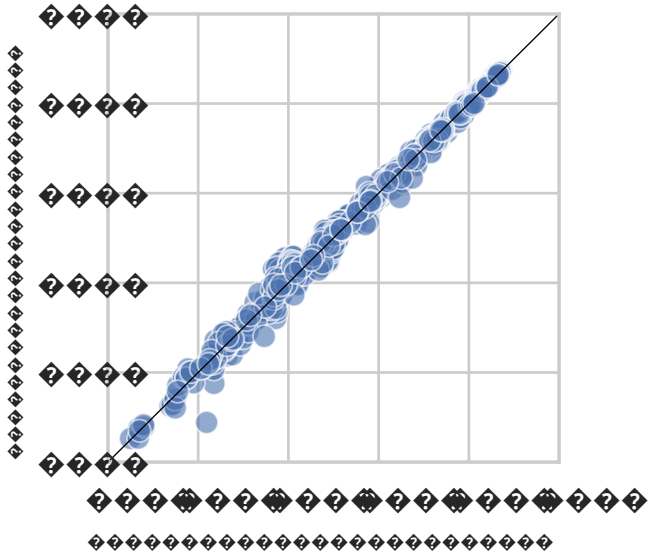

图 12.有无模式匹配的 CARTE 性能比较：图中直接比较了 275 种不同情况下有无模式匹配的 CARTE 的性能。对角线下方的点表示该方法在 x 轴上的性能更好。我们看到点沿着对角线对齐，表明 CARTE 在联合学习上的两种方法的性能相似（对值为 0.728）。结果支持CARTE不存在模式匹配。