现代社区成分分析：二十年后的深度表格基线

叶汉嘉 尹怀红 詹德川
南京大学人工智能学院、南京大学计算机软件新技术国家重点实验室
{yehj,yinhh,zhandc}@lamda.nju.edu.cn

抽象的

深度学习在各个领域的成功日益增长，促使人们研究其在表格数据中的应用，与传统的基于树的方法相比，深度模型已显示出令人鼓舞的结果。在本文中，我们重新审视了邻域成分分析 (NCA)，这是 2004 年推出的一种经典表格预测方法，旨在学习一种线性投影，以捕捉实例之间的语义相似性。我们发现，一些微小的修改，例如调整学习目标和集成深度学习架构，可以显著提高 NCA 的性能，使其超越大多数现代深度表格模型。此外，我们引入了一种随机邻域采样策略，可以提高我们提出的 M 的效率和预测准确性欧登NCA——在训练期间仅对邻居的子集进行采样，而在推理期间利用整个邻域。大量实验表明，我们的 M欧登NCA 在各种表格数据集的分类和回归任务中都取得了最先进的结果，其表现优于基于树的模型和其他深度表格模型，同时还减少了训练时间和模型大小。

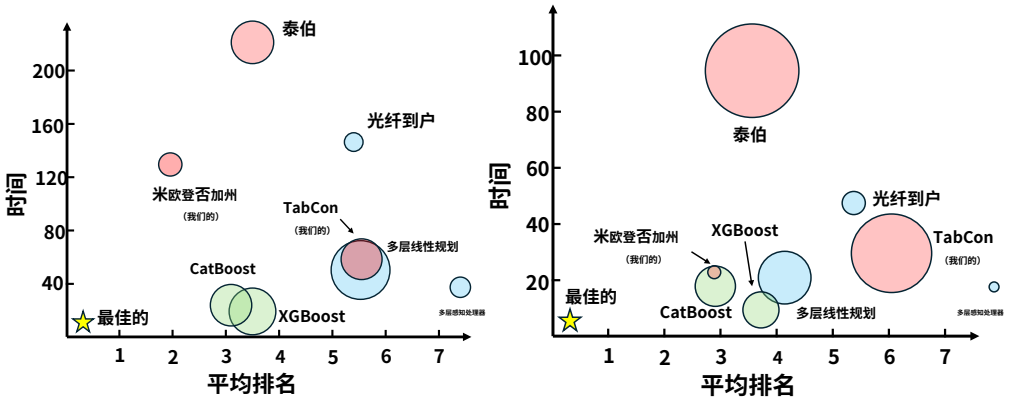
1 简介

表格数据的特点是行和列的高度结构化格式，代表单个示例和特定特征，在医疗保健 [26] 和电子商务 [39] 等领域很常见。受深度神经网络在视觉和语言等领域的成功启发 [50, 57, 17]，人们开发了许多深度模型来处理表格数据，捕捉复杂的特征交互并取得了可喜的成果 [13, 24, 42, 4, 20, 31, 10, 11, 27]。

与直接进行预测的 MLP 等参数方法相比，*非参数方法*例如最近邻利用整个训练数据集并利用实例之间的关系进行表格预测 [38]。度量学习不依赖于原始特征，而是旨在学习与嵌入空间相对应的距离度量，其中语义相似的实例彼此靠近，而不同的实例彼此远离 [62, 14, 35, 7]。这些改进的距离度量促进了非参数/基于邻域的方法的辨别过程，在处理大量类时尤其有益 [61]。类 [61]。

尽管度量学习方法与普通的最近邻方法相比表现良好，但它们在表格数据上的预测能力往往难以与基于树的非线性方法（如梯度提升决策树 (GBDT) [12, 43]）相媲美。然而，深度度量学习在图像识别 [47, 51, 52, 34]、人员重新识别 [69, 64] 和推荐系统 [28, 60] 等不同领域的成功表明，其在表格数据集方面具有尚未开发的潜力。此外，专门设计的 MLP 架构和权重正则化策略使简单的参数模型（如 MLP）在许多表格数据集上表现出色 [20, 30]，这表明类似的改进也可能使基于邻域的方法受益。

预印本。正在审核中。



（一）分类　　（b）回归

图 1：M 之间的性能-效率-大小比较欧登分类 (a) 和回归 (b) 数据集上的 NCA。代表性表格预测方法包括*古典*方法（绿色），*参数*深度方法（蓝色）和*非参数/基于邻域*根据表 1 和表 2 中所有数据集的记录，研究了 8 种深度方法（红色）。这 8 种方法的平均排名用作性能指标。我们计算了平均训练时间（以秒为单位）和模型的大小（用圆圈的半径表示，圆圈越大，模型越大）。M欧登与其他深度表格模型相比，NCA 实现了较高的训练速度，并且模型尺寸较小。TabCon 是另一个提议的基线。M欧登NCA 在两种情况下优于其他方法。

这引出了一个关键问题：“*我们能否利用现代深度学习技术改进度量学习，以在表格预测任务中取得有竞争力的表现？*”

在本文中，我们重新审视了经典的度量学习方法，即由 Goldberger 最初提出的邻域成分分析 (NCA) 等[19]。NCA 基于线性投影空间中的邻居来优化目标实例的预测准确率。通过对 NCA 进行简单修改，我们的增强方法 M欧登NCA，有效解决表格预测任务。

特别是我们的 M欧登NCA 引入了针对分类和回归任务量身定制的学习目标。我们进一步探索了一种深度神经网络架构，该架构以非线性方式将输入实例投影到潜在嵌入空间中。鉴于 NCA 的预测依赖于邻居集，这会增加训练期间的批次大小和计算负担，我们采用随机邻域采样 (SNS) 策略。SNS 在训练期间随机选择训练集的子集作为邻居，同时在测试阶段利用整个训练集。该策略不仅提高了训练效率，还提高了模型的泛化能力。我们的实验验证了 M欧登NCA 在各种数据集上的表现优于大多数现有的表格预测方法，包括基于树的方法和深度学习方法。图 1 表明 M欧登NCA 平衡了训练效率（与其他深度表格模型相比，训练时间更短）、泛化能力（平均准确率更高）和模型大小。我们论文的贡献如下：

- 我们重新审视经典的度量学习方法 NCA，并探索使用现代深度学习技术提高模型性能的方法。
- 我们提出的 M欧登NCA既能完成分类任务，也能完成回归任务，附加的采样策略提高了模型的训练效率和泛化能力。
- 跨多个数据集的大量实验表明 M欧登NCA 是表格预测任务的强大深度表格基线。

2 相关工作

使用表格数据进行学习。表格数据是各种应用中的常见格式，例如点击率预测 [45] 和时间序列预测 [1]。基于树的方法（如 XG-Boost [12]、LightGBM [32] 和 CatBoost [43]）已被证明能够有效捕获特征交互，并广泛应用于实际应用中。由于模型系列和超参数的变化

可以显著影响模型的泛化能力 [16]，一些策略侧重于自动模型选择和超参数调整，以优化不同任务的性能 [18, 25]。

深度表格数据学习。认识到深度神经网络能够从原始数据中学习特征表示并进行非线性预测的能力，最近的方法已将深度学习技术应用于表格模型 [13、24、42、9、4、30、31、11]。例如，残差网络和 Transformer 等深度架构已被用于表格预测 [20、27]。此外，还引入了数据增强策略来缓解深度模型中的过拟合 [54、6、46]。深度表格模型已在广泛的应用中展现出极具竞争力的性能，预训练的深度表格预测器已显示出推广到下游表格任务的能力 [36、27、48、41、71、70、67]。然而，研究人员已经观察到，深度模型在捕捉高阶特征交互方面仍然面临挑战，无法像基于树的模型那样有效 [23、37]。

度量学习。K 近邻 (KNN) 等方法根据整个训练集进行预测。距离度量的选择在很大程度上决定了邻居集，并影响模型的判别能力 [62, 19]。度量学习学习一种投影，使相似的实例更接近，而不同的实例更远，从而进一步改善 KNN 的分类和回归结果 [14, 61, 35, 7]。除了线性度量之外，非线性度量通常使用多种度量 [44, 65, 66]、局部组合 [59, 49, 3, 40] 或核方法 [33, 63] 来实现。度量学习最初专注于表格数据，当与图像识别 [47, 51, 52, 34]、行人重新识别 [69, 64] 和推荐系统 [28, 60] 等各个领域的深度学习技术相结合时，它已成为一种有价值的工具。最近，TabR [22] 将度量学习与 Transformer [57] 结合起来，通过基于邻域的方法增强表格预测。尽管结果令人鼓舞，但邻域选择的大量计算负担和复杂的架构限制了 TabR 的实用性。基于经典方法 NCA [19]，我们提出了一种简单的深度表格基线，可在不牺牲性能的情况下保持高效的训练速度。

3 初步

在本节中，我们首先介绍表格数据的任务学习。然后，我们简要概述表格数据的度量学习方法，特别是 NCA 方法。

3.1 使用表格数据进行学习

带标签的表格数据集格式为 $否$ 示例（表格中的行）和 d 特征/属性（表格中的列）。实例 $十_{我}$ 其描绘的是 d 特征值。特征有两种：数值（连续）特征和分类（离散）特征。给定 $十_{我,}$ 作为

$杰$ 实例的第个特征 $十_{我}$，我们使用 $十_{数量我,} \in$ R和 $十_{猫\quad 我,}$ 表示数值（*例如.*，高度人）和分类（*例如.*，一个人的性别）特征值。
分类特征通常以独热方式转换，*IE.*, $十_{猫}$ $_{我,} \in \{0，1\}^{钾_{杰}}$ 和指数值为 1 表示 $钾_{杰}$ 分类值。我们假设实例 $十_{我} \in$ R^{d}wlog 稍后会探讨其他编码策略。每个实例都与一个标签相关联 $是_{我}$， 在哪里 $是_{我} \in [碳] = \{1，...，C\}$ 在多类别分类任务中 $是_{我} \in$ R在回归任务中。

给定一个表格数据集 $德 = \{（十_{我}, y_{我}）\}^{否}_{我=1}$，我们的目标是学习一个模型 f 在 $\prod$ $德$ 地图 $十_{我}$ 到它的标签 $是_{我}$. 我们衡量 f 联合可能性 $德$，*IE.*,最大限度 $_{f}$ $_{(十_{我},y_{我}) \in D}$ 错($是_{我}$/f（$十_{我}$))。
目标可以以预测标签的负对数似然的形式重新表述：

$$分钟_{f} \sum_{(十_{我},y_{我}) \in D} - 对数 \Pr(是_{我}/f（十_{我})) = \sum_{(十_{我},y_{我}) \in D} \ell（y，\hat{y}_{我} = f（十_{我}））。\tag{1}$$

预测标签之间的差异 $\hat{y}_{我}$ 和真正的标签 $是_{我}$ 可以通过损失来衡量 ℓ（·，·),*例如.*，交叉熵。我们期望学习到的模型 f 能够将其能力扩展到从与 $德$。f 可以用经典方法（例如 SVM、MLP 或基于树的方法）来实现。

K 最近邻 (KNN) 是分类和回归领域最具代表性的非参数表格模型之一，它根据最近邻居的标签进行预测 [8, 38]。换句话说，该模型 f 通过以下方式进行预测 f（$十_{我}$，$德$） 给出一个例子 $十_{我}$，KNN计算 $十_{我}$ 以及其他情况 $德$. 假设 $钾$ 最近的邻居是 $否$（$十_{我}$；$德$) =

{（十1, y1），...，（十钾, y钾）}.然后，标签是我的十我根据邻居集中的标签进行预测否（十我；德）。 用于分类任务ŷ我是标签的多数投票否（十我；德）而 是回归任务中这些标签的平均值。

3.2 表格数据的度量学习

距离分布（十我，十杰）在 KNN 中确定最近邻居集否（十我；德），这是其关键因素之一。不使用欧几里得距离，而是使用一对之间的马哈拉诺比斯距离。（十我，十杰）公制米是：

$$分布_{米}（十_{我}，十_{杰}）=\sqrt{（十_{我}-十_{杰}）_{T}米（十_{我}-十_{杰}）}。 \quad (2)$$

矩阵米被认为是半正定的。分布2 米变成欧几里得距离（表示经过分布（十我，十杰））当我们设置度量米作为单位矩阵。米可以分解为预测米=二T和大号∈Rd×d这样马哈拉诺比斯距离就变成了投影空间中的欧几里得距离大号. 大多数距离度量学习方法学习度量米或者 大号分为两个阶段。首先基于标签构建以对或三元组形式出现的辅助信息，指导实例之间的相似性和比较关系 [62, 14, 61, 7]。然后，在学习到的度量的帮助下，KNN 做出更准确的最终预测。

邻域成分分析 (NCA)。NCA 是一种单阶段方法，它直接将投影大号实例的后验概率十被归类为是[19]：

$$镨(是\hat{}=是|x，D,大号)=\sum_{十_{我}\in D\wedge 是_{我}=是}\frac{经验-分布_{2}（大号_{T}x，大号_{T}十_{我}）}{\sum_{十_{杰}\in D，十_{杰}\neq 十_{我}}经验-分布_{2}（大号_{T}x，大号_{T}十_{杰}）}。 \quad (3)$$

因此，一个实例的后验十属于阶级是取决于其相似性（通过投影空间中的负平方欧几里得距离来衡量大号）与其邻居的类是在德。公式 3 中的预测不是平等地考虑邻域中的所有实例，而是模仿 KNN 的软版本，其中训练集中的所有实例都经过加权（较近的邻居具有更大的权重）以进行最终决策。在小样本学习等领域也探索了类似的策略 [58]。

4 米欧登全国咖啡协会

在本节中，我们描述了改进版 NCA 的学习目标、架构和训练策略。IE., 米欧登NCA（缩写为M-NCA）。

学习目标。我们使用非线性变换φ，映射输入十我进入具有维度的潜在空间d′. 假设标签是我在回归任务中是连续值，在分类任务中是独热形式。然后我们将公式 3 修改为

$$ŷ_{我}=\sum_{十_{我}\in D}\frac{指数（-分布（\phi（十），\phi（十_{我}））}{\sum_{十_{杰}\in D，十_{杰}\neq 十_{我}}指数（-ds（\phi它十），\phi（十_{杰}））}是。 \quad (4)$$

公式 4 预测相似的实例（基于它们在嵌入空间中的距离，由φ）将有接近的输出。在分类场景中，公式 4 推广了公式 3，并使用每个样本中邻居的加权平均值来预测目标实例的标签碳课程。ŷ我是一个概率向量，可以用作{镨(ŷ=是|x，D, φ）}是∈[碳]在这种情况下。在回归场景中，一个实例的预测是其邻域标量标签的加权和。通过将公式 3 与公式 1 相结合，我们设置损失ℓ在公式 1 中，分类问题中为负对数似然损失，回归问题中为均方误差。通过最小化公式 1，嵌入φ拉取训练集中的同类实例德关闭并推开其他实例。在测试阶段，整个训练集中测试实例的邻居德用于做出预测。

架构。我们设置映射φ默认为线性层，IE.,φ（十我) =线性(十我），包含一个线性投影和偏差。我们进一步提高了十我具有多个非线性层。

具体来说，我们将单层非线性映射定义为多个运算符的序列[20]，即一维批量归一化[29]、线性层、ReLU激活、dropout[53]，

和另一个线性层。换句话说，输入 $十_{我}$将被改造为

$$克（十_{我}）=线性（Dropout（（ReLU（线性（BatchNorm（十_{我}））））））。 \quad (5)$$

除了原始的线性层之外，还应用了零层或多层块来实现最终的映射。最终ϕ应用额外的批量归一化来校准嵌入。经验结果表明，在学习潜在嵌入空间方面，批量归一化优于其他归一化策略，例如层归一化 [5]。

随机邻域抽样。随机梯度下降通常用于优化深度神经网络——对一批实例进行采样，并计算批量损失以进行反向传播。然而，在训练过程中需要计算成对距离，其中批次中的实例与整个训练集之间的距离*德*被认为估计损失函数公式 4，引入了较高的计算负担。

加快M的训练速度欧登NCA，我们提出了一种随机邻域抽样（SNS）策略，其中子集*迪*训练集*德*被随机抽取作为候选实例，并计算距离。换句话说，*迪*替代品*德*在公式 4 中，只有

标签*迪*可以用来预测给定实例的标签。该模型在*全部的*训练集*德*在推理阶段。

我们通过实证观察发现，SNS 不仅提高了 M欧登NCA 因为用于反向传播的示例较少，但 SNS 也提高了学习度量的泛化能力。改进主要来自于映射ϕ在训练过程中，通过更困难的预测任务来学习，因此ϕ适应测试场景中嘈杂且不稳定的邻域。实验中研究了采样率和其他采样策略的影响。

备注1：与其他深度表格方法的关系。最近的深度表格模型 TabR [22] 也利用了最近邻，并以邻域为基础的方式进行预测。然而，我们的 M欧登NCA 和 TabR。首先，我们的方法比 TabR 简洁得多，后者采用深度神经网络和类似 Transformer 的架构来搜索整个训练集。我们的 M欧登NCA 验证了采用公式 4 中的预测策略，在大多数情况下只有线性层才能取得有竞争力的结果。其次，M欧登NCA 比 TabR 更高效，需要的训练时间和内存更少。这种效率源于其更简单的架构和随机采样策略。此外，M 不是像 TabR 那样使用实例特定的预测，而是欧登NCA 学习嵌入空间并以更有效的方式进行预测。图 1 展示了 M 的优势欧登与 TabR 相比，NCA 在性能、效率和模型大小方面均更胜一筹。

备注2：学习深度表格转换的另一种方法。我们的 M欧登NCA 利用 NCA 实现公式 4 中的目标函数，其中学习到的表格嵌入的质量ϕ（$十$）直接影响分类/回归性能$十$. 如 3.2 节所述，另一种常见的学习策略ϕ涉及两个阶段的过程：首先，ϕ优化匹配映射实例之间的比较关系，例如三元组，然后基于此应用简单的分类器，如 KNN 和 LRϕ. 由于小批量$乙$在训练过程中，每次采样 100 个样本，我们研究了监督对比损失 [51, 34]，其中监督是在小批量内生成的。目标是学习映射ϕ是：

$$\ell（\phi\,|乙）=-\sum_{（十_{我},y_{我}）\in B}\sum_{十_{杰}\in 磷（十_{我}）}日志\frac{指数（\phi（十）_{我}\,\phi（十_{杰}）\,）/\tau}{\sum_{十_{钾}\in Q,\,十_{钾}\neq十_{我}}指数（\phi（十_{我}）\cdot\phi（十_{钾}）/\tau）}，\quad (6)$$

$$磷（十_{我}）=\{十_{杰}|\forall（十_{杰},y_{杰}）\in B\wedge是_{杰}=是_{我}\}。 \quad (7)$$

$磷$（$十_{我}$）是小批量中与 具有相同标签的实例集$十_{我}$。τ是用于校准损失函数的温度参数。通过最小化公式 6，ϕ还可以将同类实例拉到一起，将不同类的实例分开。通过学习到的ϕ，我们应用 KNN 进行分类/回归推理。我们在回归场景中离散化标签值，并将此基线方法表示为 Tabular Contrastive (TabCon)。

5 实验

我们验证了 M 的性能欧登标准表格分类和回归任务中的 NCA（M-NCA）。然后，我们分析 M欧登NCA 与消融研究。

表 1：所有方法在 12 个分类数据集上的平均准确率（越高越好）。每个数据集上的最佳结果以粗体显示，第二好的结果以下划线显示。我们还报告了所有数据集中不同方法的平均排名（越低越好）。

↑M-NCA	TabCon	XGBoost	CatBoost	MLP	FT-T	TabR	KNN	MLP公共关系	肺动脉高压	LMNN	NCA
广告	87.29	86.97	87.21	87.48	85.80	85.89	86.86	84.61	86.97	85.53	83.21	83.87
CU	79.99	77.72	81.73	81.03	75.54	80.65	79.61	70.08	80.75	75.89	72.12	73.78
发光	96.12	87.44	92.06	91.69	85.49	87.50	96.46	84.27	86.93	85.10	83.56	85.15
安永	99.41	82.88	71.02	71.88	61.29	71.49	98.20	59.46	74.27	60.13	54.98	66.50
他	39.84	38.32	37.86	38.26	38.34	38.49	40.80	32.91	38.95	37.64	29.36	31.30
你好	73.09	72.36	72.73	72.66	72.29	72.92	72.85	66.78	72.64	72.48	61.44	66.94
知识产权	87.96	87.65	88.64	87.74	84.86	84.95	87.23	83.72	87.84	84.83	83.81	84.10
JA	74.08	72.17	72.55	72.39	71.97	72.38	73.44	65.67	72.10	71.52	59.12	64.64
马萨诸塞州	87.90	87.22	87.67	87.87	87.08	87.73	87.91	84.65	87.17	86.94	83.68	84.73
加时赛	82.37	78.94	82.45	82.21	81.47	80.82	82.38	78.15	81.24	81.20	77.37	78.85
肺动脉高压	89.29	87.88	87.66	87.97	86.88	87.86	89.10	86.86	87.13	85.54	85.75	87.05
威斯康星州	74.74	73.72	74.39	74.59	72.50	72.64	74.22	77.10	72.50	72.16	73.58	72.60
秩	2.000	5.917	3.833	3.750	8.000	5.417	3.083	10.000	5.667	9.250	11.333	9.750

5.1 设置

数据集。我们验证我们的 M欧登对从 OpenML [56] 和 Kaggle 收集的 12 个分类数据集和 10 个回归数据集进行 NCA。附录中的表 5 列出了数据集的详细统计数据。12 个分类数据集中有 4 个包含两个以上的类别。

评估。我们主要遵循 [20, 22] 中的设置来评估所有方法。我们将每个数据集随机分成三个分区，分别以 64%/16%/20% 的比例进行训练/验证/测试。对于每个数据集，我们使用 15 个随机种子训练给定模型，并报告测试集上的平均性能。我们使用准确率进行分类（越高越好），使用均方根误差 (RMSE) 进行回归（越低越好）。我们还遵循 [16, 37] 并报告所有方法和数据集的平均性能排名（越低越好）。

比较方法。我们比较 M欧登NCA 有三种不同的方法。首先，我们比较经典的参数表格预测方法，包括基于树的方法 XGBoost[12] 和 CatBoost [43]。然后，我们比较参数深度表格模型。我们遵循 [20] 中的架构来实现 MLP，MLP公共关系[22] 是 MLP 的一个变体，它利用了数值特征的分段线性编码 [21]。FT-Transformer (FT-T) [20] 为每个属性学习 token，然后使用 transformer 进行预测。PTaRL 是一种基于空间校准的原型表格方法 [68]。对于基于邻域的表格方法，我们考虑 *钾*NN、度量学习方法（如 LMNN [61] 和 NCA [19]）、TabR [22]，以及我们的表格对比基线 TabCon。[1]

实施细节。我们按照 [20] 对所有数据集进行预处理。对于所有深度方法，我们将批处理大小设置为 1024。所有方法的超参数均基于训练和验证集通过 Optuna [2] 搜索，这些训练和验证集经过了 [20, 22] 超过 100 次试验。在最后 15 个种子期间，表现最佳的超参数是固定的。由于 SNS 的采样率有效地提高了性能并降低了训练开销，我们将其视为超参数并在 [0.05, 0.6] 范围内搜索。像 LMNN 这样的度量学习方法是基于度量学习库 [15] 实现的。我们对分类特征使用独热编码，对数值特征使用 [22] 中提出的 PLR（精简版）嵌入（PLR 嵌入的简化版本（周期性嵌入、线性层和 ReLU 的组合）[21]）。为了公平比较，在 MLP 中也应用了相同的编码策略公共关系和 TabR。

5.2 主要结果

M 之间的比较结果欧登NCA 和其他用于分类和回归任务的方法分别列于表 1 和表 2，其中 M欧登 NCA 在两种情况下都取得了最佳平均排名。

[1]我们的 M 代码欧登NCA 网址：https://github.com/qile2000/LAMDA-TALENT。

表 2：10 个回归数据集上所有方法的平均 RMSE（越低越好）。每个数据集上的最佳结果以粗体显示，第二好的结果用下划线表示。数据集名称旁边的科学符号表示结果的尺度，*例如*.,×10 意味着所有结果的最终值应该是10.我们还报告所有数据集的平均排名（越低越好）。

↓	M-NCA	TabCon	XGBoost	CatBoost	MLP	FT-T	TabR	KNN	MLP公共关系	磷灰石
人工智能×10-3	1532年	1522年	1527年	. 1465	.1558	.1565	.1546	.2431	1522年	1563年
双×102	.7057	.7352	.7180	.7264	.7773.	7466.	6657	1.087	.7155	.7193
加州	.4212	.4581	.4328	.4360	.5074	.4701。	4157	.5843	.4690	.5086
厘米×103	.4618	.5011	.4809	.4929	.5131.	5187.	5091	.8754	.4655	.5211
CP×10	.2381	.2473	.2404	.2365	.2490。	2310。	2337	.8825	.2468	.2487
福×10	.7272	.7436	.7384	.7393	.7900	.7885	.7398	.8140	.7428	.7896
何×105	.3041	.3156	.3026	.3051	.3161.	3074.	3133	.3663	.3093	.3149
侯×105	.4291	.4963	.4797	.4546	.5221	.4791。	4149	.6889	.4921	.5118
洛杉矶×103	.4135	.4529	.4448	.4506	.4789	.4609	.4606	.4968	.4753	.4777
米娅×105	.8284	.9322	.8740	.8034	.8973.	8717.	8585	1.508	.8562	.9302
秩	2.200	6.100	3.500	3.100	8.300	5.900	3.600	10.000	4.500	7.800

表 3：不同架构方法的比较结果。NCA 不能应用于回归任务。

	泰伯	NCA	线性层范数残差	M-NCA	
加州 ↓	.4157	-.4212	.4210	.4237	.4212
广告 ↑	86.86	83.87	87.18	87.20	87.25	87.29
米娅×105 ↓	.8585	-.8325	.9070	.8418	.8284
肺动脉高压 ↑	89.10	87.05	89.09	88.70	89.78	89.29

对于表 1 中的分类结果，M欧登NCA 在大多数情况下比基于树的方法（如 XGBoost）取得更好的效果，这表明利用深度神经网络 M欧登NCA 具有更强的非线性预测能力。可以清楚地观察到，虽然 LMNN 和 NCA 等经典度量学习方法在某些情况下可以改进 KNN，但它们很难达到与基于树和深度的方法相媲美的性能。与 FT-T 和 MLP 等深度表格方法相比公共关系, 米欧登NCA 仍然保持其优势。最相关的比较方法是 TabR，它提取训练集中所有邻居的特征，并使用 Transformer 的变体进行预测。我们的 M欧登NCA 优于 TabR，架构更简单，训练时间更短。结果表明 M欧登NCA 是一种有效的表格深度基线。我们的 TabCon 基线也证明了其学习映射的有效性，在所有情况下都优于经典度量学习方法。TabCon 与 FT-T 竞争，甚至在某些情况下超越 XGBoost，例如“HE”和“PH”。M 的优势欧登TabCon 上的 NCA 可能是由于 M 中学习到的映射和预测性能之间的直接联系所致欧登NCA. 表 2 中的回归案例也存在类似现象。欧登NCA 是唯一一个在回归数据集上始终优于基于树的方法的深度模型。经典的度量学习方法不能直接应用于回归。我们的 M欧登NCA 显示出优于两种竞争方法 MLP公共关系和 TabR。图 1 显示了代表性分类和回归方法在所有数据集上的性能、训练时间和模型大小的比较。虽然某些模型（例如 TabR）获得了强大的性能，但它们需要更长的训练时间。我们的 M欧登NCA在各种模型评价标准之间取得了良好的平衡。

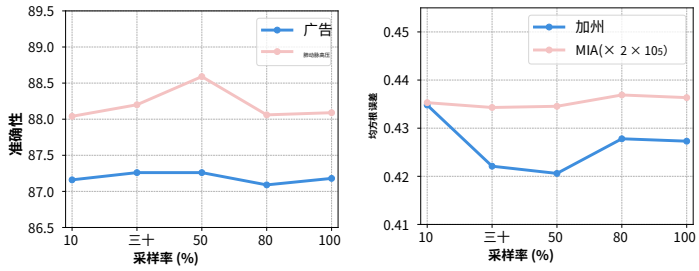
5.3 消融研究

我们分析了所提出的 M 的不同属性欧登NCA 基于三个数据集，其中 CA/MIA 用于回归，AD/PH 用于分类。

线性与深度架构。我们首先研究架构的设计ϕ在 M欧登NCA，我们添加一个或多个块层*克*（·）基于线性投影。层数

表 4：M 的结果欧登具有不同距离函数的 NCA 变体，用于在学习到的嵌入空间中获得邻里关系。

	欧几里得 欧几里得2	ℓ1-范数余弦内积		
加州 ↓	. 4212 . 4271	.4184	.4264	.4528
广告 ↑	87.29 87.04	87.25	87.18	87.07
米娅 ×105 ↓	.8284 .8637	.8624	.8627	.8497
肺动脉高压 ↑	89.29 89.38	89.08	89.10	88.23



（一）分类　　（二）回归

图2：随机邻域采样策略中{10%、30%、50%、80%、100%}不同采样率时性能的变化。

设置为超参数，可以根据验证集确定。我们考虑三种选择。首先，我们设置φ作为线性投影，其中投影空间的维数是超参数。我们还用层归一化代替了块中的批量归一化。最后，我们装备φ使用从块的输入到其输出的另一个残差链接。

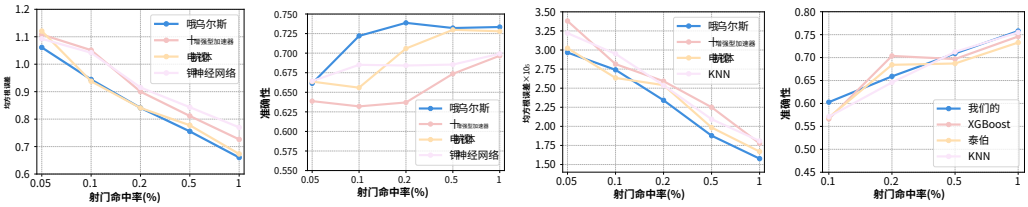
结果列于表 3 中，我们还包含了 TabR/NCA 的结果以供参考。尽管 NCA 使用的是线性投影，但我们发现 M 的线性版本欧登NCA（第四列）在两个分类数据集上取得了更好的表现。例如，在 AD 上，M 的线性版本欧登NCA 的表现比 NCA 高出约 4%。比较结果表明 M 的目标函数和训练策略欧登NCA 是必不可少的。与最后一列 M 相比欧登NCA 利用深度神经网络，我们发现在 MIA 等数据集上使用非线性深度架构是必要的。线性版本可以实现有竞争力甚至更好的结果，因为投影维度是在较小的超参数空间中搜索的，并且对于 M欧登NCA 根据验证集从零层到更多层搜索附加层的数量。

第五列的变体在块中使用了另一种归一化策略，即层归一化，而不是在 M 中使用批量归一化欧登NCA。我们根据经验发现，批量归一化的平均表现更好，尤其是在分类任务中。在比较最后两列中的残差和 MLP 块时，我们发现直接使用多个非线性层表现更好一些，因此我们选择公式 5 中的 MLP 实现作为 M欧登NCA。

距离函数的影响。目标实例的预测标签 $十_{我}$由学习到的嵌入空间中邻居的标签决定，该空间由φ距离函数是确定嵌入空间中实例之间成对关系的关键，影响 w

√公式 4 中的八。在 M欧登NCA，我们选择欧几里得距离分布欧盟委员会（φ（十我），φ（十杰)) = （φ（十我）－φ（十杰））ㄒ（φ（十我）－φ（十杰)) = 〃φ（十我）－φ（十杰）〃2我们还利用其他距离功能，*例如.*，平方欧几里德距离分布2　　欧盟委员会（φ（十我），φ（十杰)),这ℓ1-范数距离分布（φ（十我），φ（十杰)) = 〃φ（十我）－φ（十杰）〃1，（负）余弦相似度分布（φ（十我），φ（十杰)) = －（我十杰）／（〃十我〃2〃十杰〃2）以及（负）内积分布（φ（十我），φ（十杰)) = -十ㄒ　我十杰。这表 4 列出了使用不同距离函数的结果，其中欧氏距离平均取得了良好的结果。尽管ℓ1-norm 也表现良好，但它会比欧几里得距离引入更大的计算成本。

采样率的影响。由于在学习到的嵌入空间中计算距离的计算成本巨大，我们的 M欧登NCA 采用随机邻域采样 (SNS) 策略，其中每个小批量仅随机采样一部分训练集数据。



（一）CA↓ （坏的↑） （三）MIA↓ （四）PH↑

图 4：在训练集示例的不同百分比中，少量训练示例的性能变化。由于 PH 数据集只有一个样本，即 0.05% 的命中百分比，因此我们仅显示命中百分比从 0.1% 到 1% 的模型性能。

测试阶段使用整个训练集。由于使用了采样数据的子集，训练时间成本大大降低。我们在图 2 中改变了比例并评估了模型的相应测试性能。从图 2 中我们观察到，30%-50% 的训练集比例对 M 更有帮助欧登NCA 比使用完整训练集更有效。SNS 不仅提高了训练效率，还提高了模型的泛化能力。图表表明，采样训练数据的比例很重要，这是一个针对 M 进行调整的超参数欧登NCA。

采样策略的影响。如前所述，SNS 在计算公式 4 的损失时会为每个小批量随机抽取一个训练数据子集。我们还研究了在采样过程中纳入更丰富的信息是否可以进一步提高模型的分类/回归能力，*例如*.，实例的标签。

除了之前使用的完全随机采样策略外，我们还考虑了另外两种采样策略。第一种是按类随机采样，这意味着给定一个比例，我们从训练集中的每个类中采样并将它们组合在一起。此策略利用训练标签信息，并保留采样子集中将存在的所有类的实例。此外，我们还考虑基于实例之间成对距离的采样策略。由于实例的邻居在公式 4 中可能贡献更多（具有更大的权重），因此给定一个小批量，我们首先使用嵌入函数计算批量中实例与所有训练集之间的欧几里得距离 ϕ在当前时期。然后我们根据成对距离值的倒数对训练集进行采样。具体来说，给定一个实例 $十_{我}$，我们提供特定实例的邻域候选，并且 $十_{杰}$在训练集中的样本是基于概率的 ~1/（分鐘（ϕ（$十_{我}$），ϕ（$十_{杰}$））$_{\tau}$。 τ是用于校准分布的非负超参数。距离计算需要模型的前向传递ϕ由于基于距离的采样策略是针对所有训练实例进行采样，而基于实例的邻域使得损失与训练数据的范围很广有关，因此基于距离的采样策略训练速度慢，计算负担大。

图 3 列出了不同采样策略在两个分类数据集上的比较结果。我们发现基于标签的采样策略无法提供进一步的改进。虽然基于距离的策略在某些情况下有所帮助，但改进是有限的。从性能和效率的综合考虑，我们选择在 M 中使用 vanilla 随机采样欧登NCA。

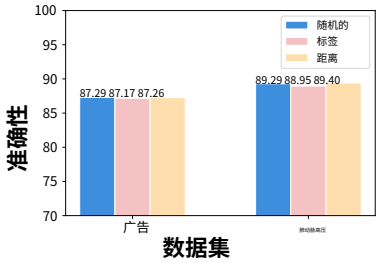


图3：不同采样策略的影响，即随机策略、基于标签的策略和基于距离的策略。

利用少量数据进行学习。我们研究了训练样本数量有限时性能的变化。结果如图 4 所示。给定一个数据集，我们通过随机抽取训练数据子集来合成小样本学习场景，其中无线电变化来自{0。05%，0。1%，0。2%，0。5%，1%}由于训练集规模较小，我们学习所有方法

使用默认超参数。结果重复 10 次，并报告其平均值。基于图 4 所示的少样本性能，我们发现我们的 M欧登与其他表格预测方法相比，NCA 具有竞争力。

6 结论

利用学习到的嵌入空间中的关系进行预测是机器学习中经典的度量学习概念。虽然传统的度量学习方法在表格数据集上的表现通常不如基于树的方法，但我们的研究重新审视并增强了最具代表性的方法之一 NCA。改进的 M欧登NCA 已成为深度表格预测任务的强大基线，在分类和回归能力方面通常优于其他方法，同时还减少了模型大小和训练时间。我们相信我们的研究将鼓励进一步研究表格数据领域的经典方法。通过对其核心概念进行微小但有效的修改，这些方法可以焕发活力，提供重大改进和新颖的解决方案。

参考

[1] Nesreen K Ahmed、Amir F Atiya、Neamat El Gayar 和 Hisham El-Shishiny。时间序列预测机器学习模型的实证比较。*计量经济学评论*《癌症》杂志, 29 (5-6): 594–621, 2010.

[2] 秋叶拓哉、佐野正太郎、柳濑俊彦、太田健、小山正德。 Optuna：下一代超参数优化框架。在*知识发现*，2019年。

[3] Joseph St. Amand 和 Jun Huan。稀疏组合局部度量学习。在*知识发现*，2017年。

[4] Sercan Ö. Arik 和 Tomas Pfister。Tabnet：专注的可解释表格学习。在*美国航空学会联合会*，2021年。

[5] Lei Jimmy Ba、Jamie Ryan Kiros 和 Geoffrey E. Hinton。层归一化。*钴*，abs/1607.06450，2016年。

[6] Dara Bahri、Heinrich Jiang、Yi Tay 和 Donald Metzler。Scarf：使用随机特征损坏的自监督对比学习。*国际肾病研究联合会*，2022年。

[7] Aurélien Bellet、Amaury Habrard 和 Marc Sebban。*度量学习*. Morgan & Claypool 出版社，2015 年。

[8] 克里斯托弗・毕晓普。*模式识别和机器学习*.施普林格出版社，2006 年。

[9] Vadim Borisov、Tobias Leemann、Kathrin Seßler、Johannes Haug、Martin Pawelczyk 和 Gjergji Kasneci。深度神经网络和表格数据：一项调查。*IEEE 神经网络和学习系统学报*，abs/2110.01889:1–21，2022年。

[10] Chun-Hao Chang、Rich Caruana 和 Anna Goldenberg。NODE-GAM：用于可解释深度学习的神经广义加性模型。*国际肾病研究联合会*，2022年。

[11] Jintai Chen、Kuanlun Liao、Yao Wan、Danny Z. Chen 和 Jian Wu。Danets：用于表格数据分类和回归的深度抽象网络。*美国航空学会联合会*，2022年。

[12] Tianqi Chen 和 Carlos Guestrin。Xgboost：可扩展的树提升系统。*知识发现*，2016 年。

[13] Heng-Tze Cheng、Levent Koc、Jeremiah Harmsen、Tal Shaked、Tushar Chandra、Hrishi Aradhye、Glen Anderson、Greg Corrado、Wei Chai、Mustafa Ispir、Rohan Anil、Zakaria Haque、Lichan Hong、Vihan Jain、Xiaobing Liu，和赫马尔・沙阿。推荐系统的广泛和深度学习。在*远程遥控*，2016 年。

[14] Jason V. Davis、Brian Kulis、Prateek Jain、Suvrit Sra 和 Inderjit S. Dhillon。信息理论度量学习。*国际激光医学联合会*，2007 年。

[15] William de Vazelhes、CJ Carey、Yuan Tang、Nathalie Vauquier 和 Aurélien Bellet。metriclearn：Python 中的度量学习算法。*机器学习研究杂志*, 21(138): 1–6, 2020。

[16]曼努埃尔·费尔南德斯·德尔加多、伊娃·塞纳达斯、塞南·巴罗和迪纳尼·戈麦斯·阿莫林。我们是否需要数百个分类器来解决现实世界的分类问题？*机器学习研究杂志*《细胞与分子生物学杂志》, 15(1): 3133–3181, 2014.

[17] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。BERT：用于语言理解的深度双向转换器的预训练。*抗凝血酶原*，2019年。

[18] Matthias Feurer、Aaron Klein、Katharina Eggensperger、Jost Tobias Springenberg、Manuel Blum 和 Frank Hutter。高效而强大的自动化机器学习。*神经信息处理系统*，2015 年。

[19] Jacob Goldberger、Sam T. Roweis、Geoffrey E. Hinton 和 Ruslan Salakhutdinov。邻里成分分析。*神经信息处理系统*，2004年。

[20] Yury Gorishniy、Ivan Rubachev、Valentin Khrulkov 和 Artem Babenko。重新审视表格数据的深度学习模型。*神经成像与成像系统*，2021年。

[21] Yury Gorishniy、Ivan Rubachev 和 Artem Babenko。关于表格深度学习中数值特征的嵌入。*神经成像与成像系统*，2022年。

[22]尤里·戈里什尼、伊万·鲁巴切夫、尼古拉·卡尔塔舍夫、丹尼尔·什伦斯基、阿基姆·科捷尔尼科夫和阿特姆·巴本科。 Tabr：表格深度学习将于 2023 年与最近邻居相遇。*国际肾病研究联合会*，2024年。

[23] Léo Grinsztajn、Edouard Oyallon 和 Gaël Varoquaux。为什么基于树的模型在典型的表格数据上仍然优于深度学习？*神经成像与成像系统*，2022年。

[24] 郭惠峰，唐瑞明，叶云明，李振国，何秀强。 Deepfm：用于 CTR 预测的基于分解机的神经网络。在*国际计算机辅助教学*，2017年。

[25] Isabelle Guyon、Lisheng Sun-Hosoya、Marc Boullé、Hugo Jair Escalante、Sergio Escalera、Zhengying Liu、Damir Jajetic、Bisakha Ray、Mehreen Saeed、Michèle Sebag 等。 automl挑战系列分析。*自动机器学习*, 177: 177–219, 2019年。

[26] Md. Rafiul Hassan、Sadiq Al-Insaif、Muhammad Imtiaz Hossain 和 Joarder Kamruzzaman。一种用于预测 IVF 治疗后妊娠结果的机器学习方法。 *神经计算与应用*《临床神经病学杂志》2019年第32期, 32(7): 2283–2297, 2020年。

[27] Noah Hollmann、Samuel Müller、Katharina Eggensperger 和 Frank Hutter。Tabpfn：一款可在一秒钟内解决小型表格分类问题的转换器。*国际肾病研究联合会*，2023年。

[28] Cheng-Kang Hsieh、Longqi Yang、Yin Cui、Tsung-Yi Lin、Serge J. Belongie 和 Deborah Estrin。协作度量学习。*万维网*，2017年。

[29] Sergey Ioffe 和 Christian Szegedy。批量标准化：通过减少内部协变量偏移来加速深度网络训练。*国际激光医学联合会*，2015 年。

[30] Arlind Kadra、Marius Lindauer、Frank Hutter 和 Josif Grabocka。经过良好调整的简单网络在表格数据集上表现出色。在*神经成像与成像系统*，第 23928–23941 页，2021 年。

[31] Liran Katzir、Gal Elidan 和 Ran El-Yaniv。Net-dnf：表格数据的有效深度建模。*国际肾病研究联合会*，2021年。

[32] 柯国林，孟奇，托马斯·芬利，王泰丰，陈伟，马卫东，叶其伟，刘铁岩。 Lightgbm：一种高效的梯度提升决策树。在*神经信息处理系统*，2017年。

[33] Dor Kedem、Stephen Tyree、Kilian Q. Weinberger、Fei Sha 和 Gert RG Lanckriet。非线性度量学习。在*神经信息处理系统*，2012 年。

[34] Prannay Khosla、Piotr Teterwak、Chen Wang、Aaron Sarna、Yonglong Tian、Phillip Isola、Aaron Maschinot、Ce Liu 和 Dilip Krishnan。监督对比学习。在*神经成像与成像系统*，2020年。

[35] Brian Kulis.度量学习：一项调查。*机器学习的基础和趋势*,5(4), 2013。

[36] Roman Levin、Valeriia Cherepanova、Avi Schwarzschild、Arpit Bansal、C. Bayan Bruss、Tom Goldstein、Andrew Gordon Wilson 和 Micah Goldblum。使用深度表格模型进行迁移学习。*国际肾病研究联合会*，2023年。

[37] Duncan C. McElfresh、Sujay Khandagale、Jonathan Valverde、Vishak Prasad C.、Ganesh Ramakrishnan、Micah Goldblum 和 Colin White。神经网络在表格数据上何时优于提升树？*神经成像与成像系统*，2023年。

[38] Mehryar Mohri、Afshin Rostamizadeh 和 Ameet Talwalkar。*机器学习基础*麻省理工学院出版社，2012年。

[39] Lennart J Nederstigt、Steven S Aanen、Damir Vandic 和 Flavius Frasincar。Floppies：一种用于从电子商务商店的表格数据中大规模填充产品信息的本体框架。*决策支持系统*, 59: 296–311, 2014。

[40] Yung-Kyun Noh、Byoung-Tak Zhang 和 Daniel D. Lee。用于最近邻分类的生成局部度量学习。*IEEE 模式分析与机器智能学报*, 40(1): 106–118, 2018。

[41] Soma Onishi、Kenta Oono 和 Kohei Hayashi。Tabret：针对看不见的列进行基于 Transformer 的表格模型预训练。*钻*，绝对/2303.15747，2023年。

[42] Sergei Popov、Stanislav Morozov 和 Artem Babenko。用于表格数据深度学习的神经无意识决策集成。*国际肾病研究联合会*，2020年。

[43]Liudmila Ostroumova Prokhorenkova、Gleb Gusev、Aleksandr Vorobev、Anna Veronika Dorogush 和 Andrey Gulin。 Catboost：具有分类特征的无偏增强。在*神经成像与成像系统*，2018年。

[44] Qi Qian, Rong Jin, Shenghuo Zhu, 和 Yuanqing Lin. 通过多阶段度量学习实现细粒度视觉分类。*计算机视觉与图像处理*，2015 年。

[45] Matthew Richardson、Ewa Dominowska 和 Robert Ragno。预测点击次数：估算新广告的点击率。*万维网*，2007 年。

[46] Ivan Rubachev、Artem Alekberov、Yury Gorishniy 和 Artem Babenko。重新审视表格深度学习的预训练目标。*钻*，abs/2207.03208，2022 年。

[47] Florian Schroff、Dmitry Kalenichenko 和 James Philbin。Facenet：用于人脸识别和聚类的统一嵌入。*计算机视觉与图像处理*，2015 年。

[48] Junhong Shen、Liam Li、Lucio M Dery、Corey Staten、Mikhail Khodak、Graham Neubig 和 Ameet Talwalkar。跨模态微调：先对齐，再细化。在*国际激光医学联合会*，2023年。

[49]袁石、Aurélien Bellet 和 Fei Sha。稀疏组合度量学习。在*美国航空学会联合会*，2014 年。

[50] Karen Simonyan 和 Andrew Zisserman。用于大规模图像识别的超深卷积网络。*国际肾病研究联合会*，2015 年。

[51] Kihyuk Sohn。使用多类 n 对损失目标改进深度度量学习。在*神经信息处理系统*，2016年。

[52] Hyun Oh Song、Yu Xiang、Stefanie Jegelka 和 Silvio Savarese。通过提升结构化特征嵌入进行深度度量学习。在*计算机视觉与图像处理*，2016 年。

[53] Nitish Srivastava、Geoffrey E. Hinton、Alex Krizhevsky、Ilya Sutskever 和 Ruslan Salakhutdinov。Dropout：一种防止神经网络过拟合的简单方法。*机器学习研究杂志*《自然》杂志, 15(1): 1929–1958, 2014.

[54] Talip Ucar、Ehsan Hajiramezanali 和 Lindsay Edwards。Subtab：用于自监督表示学习的表格数据子集特征。在*神经成像与成像系统*，第 18853–18865 页，2021 年。

[55] 劳伦斯・范德马滕和杰弗里・辛顿。使用 t-sne 可视化数据。*机器学习研究杂志*,9(11), 2008。

[56] Joaquin Vanschoren、Jan N Van Rijn、Bernd Bischl 和 Luis Torgo。 Openml：机器学习中的网络科学。*ACM SIGKDD 探索简讯*, 15(2):49–60, 2014。

[57] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。你只需要关注。*神经信息处理系统*，2017年。

[58] Oriol Vinyals、Charles Blundell、Tim Lillicrap、Koray Kavukcuoglu 和 Daan Wierstra。一次性学习的匹配网络。*神经信息处理系统*，2016 年。

[59] Jun Wang、Alexandros Kalousis 和 Adam Woznica。用于最近邻分类的参数局部度量学习。在*神经信息处理系统*，2012 年。

[60] 魏天军, 马江宏, Tommy WS Chow. 协作残差度量学习. 在*信号*，2023年。

[61] Kilian Q. Weinberger 和 Lawrence K. Saul. 大边距最近邻分类的距离度量学习。*机器学习研究杂志*, 10: 207–244, 2009。

[62] Eric P. Xing、Andrew Y. Ng、Michael I. Jordan 和 Stuart Russell。距离度量学习及其在带边信息聚类中的应用。*神经信息处理系统*，2002 年。

[63]Zhixiang Eddie Xu、Kilian Q. Weinberger 和 Olivier Chapelle，核机的距离度量学习。*钆*，绝对/1208.3422，2012年。

[64] 杨迅, 王猛, 陶大成. 利用特权信息的度量学习进行行人重新识别。*IEEE 图像处理学报*《细胞与分子生物学杂志》, 27(2): 791–805, 2018。

[65] 叶汉佳、詹德川、司学敏、蒋远和周志华。是什么让物体变得相似：一种统一的多度量学习方法。*神经信息处理系统*，2016 年。

[66] Han-Jia Ye, De-Chuan Zhan, Nan Li, and Yuan Jiang. 学习多个局部指标：全局考虑有所帮助。*IEEE 模式分析与机器智能学报*《自然》杂志, 42(7): 1698–1712, 2020年。

[67] 叶汉佳，周其乐，詹德川，通过元表示对异构表格数据进行无需训练的泛化。*钆*，绝对/2311.00055，2023年。

[68]叶航亭，范伟，宋晓庄，郑顺，赵何，丹丹郭，常毅。 Ptarl：通过空间校准进行基于原型的表格表示学习。在*国际肾病研究联合会*，2024年。

[69] 董毅，雷震，廖胜才，李斯坦。用于人员重新识别的深度度量学习。在*国际医学医学学会联合会*，2014 年。

[70] 周其乐、叶汉佳、王乐业和詹德川，《解锁表格数据深度模型中标记的可转移性》。*钆*，绝对/2310.15149，2023年。

[71] Bingzhao Zhu、Xingjian Shi、Nick Erickson、Mu Li、George Karypis 和 Mahsa Shoaran。Xtab：表格转换器的交叉表预训练。在*国际激光医学联合会*，2023年。

附录由五部分组成：

- 附录A：数据集和实施细节。
- 附录B：额外的实验结果。

附录 A 数据集

在本节中，我们介绍训练之前对数据集应用的预处理步骤，以及我们使用的数据集的详细描述。

A.1 数据预处理

我们遵循 [20] 并对所有方法的数据集进行预处理。例如，我们使用标准化（包括均值减法和缩放）来规范化每个数值数据集。我们还对所有分类特征应用独热编码。

A.2 数据集信息

数据集来自不同的领域，包括 OpenML [56] 和 Kaggle。这些数据集通常用于表格预测领域 [23, 20]。描述如表 5 所示。为了减少随机性，我们使用了样本量大于 1000 的数据集。

表 5：所有数据集的描述。有 8 个二分类数据集、4 个多分类数据集和 10 个回归数据集。

	缩写	任务类型	数据大小	数量	类别	类别数量
成人	广告	分类	48842	6	8	2
副翼	人工智能	回归	13750	40	0	-
加州住房	加州	回归	20640	8	0	-
CPMP-2015	厘米	回归	2108	22	2	-
CPU 实际值	CP	回归	8192	21	0	-
电	发光	分类	45312	7	1	2
眼部运动	安永	分类	10936	24	3	3
送餐时间	福	回归	45593	6	2	-
海伦娜	他	分类	65196	二十七	0	100
希格斯·斯莫尔	你好	分类	98050	二十八	0	2
房屋	侯	回归	20640	8	0	-
房屋_16H	何	回归	22784	16	0	-
詹尼斯	JA	分类	83733	54	0	4
共享单车需求挑战	双	回归	10886	3	6	-
KDDCup09_追加销售	CU	分类	5128	三十四	15	2
kdd_ipums_la_97-小	知识产权	分类	5188	20	0	2
笔记本电脑价格数据集	洛杉矶	回归	4441	9	0	-
魔幻望远镜	马萨诸塞州	分类	19020	9	0	2
迈阿密住房2016	米娅	回归	13932	16	0	-
奥托集团产品	加时赛	分类	61878	93	0	9
音素	肺动脉高压	分类	5404	5	0	2
葡萄酒	威斯康星州	分类	2554	4	0	2

我们随机抽取 20% 的实例来构建测试集。其余 80% 的实例用于拆分。在训练集中，我们随机抽取 20% 的实例作为验证集。验证集用于调整超参数和执行早期停止。我们选择了在调整超参数后在验证集上表现最佳的模型进行评估。

A.3 硬件

大部分实验都是在 Tesla V100 GPU 上进行的，包括与时间和内存开销计算相关的实验。

表 6：MLP 的结果和我们的 M欧登NCA 对数值属性进行有或没有 PLR 编码。

	无 PLR 的 MLP	带 PLR 的 MLP	无 PLR 的 M-NCA	带 PLR 的 M-NCA
加州 ↓	.5074	.4690	.4266	.4212
广告 ↑	85.80	86.97	86.77	87.29
米娅 ×105 ↓	.8973	.8562	.8759	.8284
肺动脉高压 ↑	86.88	87.13	88.41	89.29

表 7：采用不同预测方式的 TabCon 变体的结果。TabCon 的默认选择是基于学习到的嵌入的 KNN。

	加州 ↓	广告 ↑	米娅 ×105 ↓	肺动脉高压 ↑
TabCon	.4581	86.97	.9322	87.88
TabCon（左）	.4724	87.09	.9017	86.69

A.4 TabCon 的实现

TabCon 采用监督对比损失进行两阶段训练。在训练阶段，映射ϕ，定义与 M 相同欧登为了简单起见，NCA 被训练来将原始特征投影到潜在空间中，如公式 7 所示。在计算监督对比损失之前，每个映射ϕ（$\cdot$）归一化为单位长度。在验证阶段，一个简单的预测器（*例如*.、KNN 和 LR）根据在训练阶段学习到的映射进行进一步训练。根据简单预测器的性能，在验证阶段应用早期停止。我们使用带有欧几里得距离的 KNN 作为默认预测器。

由于 Supervised Contrastive Loss 最初是为分类任务提出的，因此在计算损失之前，我们会对回归任务的标签进行离散化。具体来说，我们会根据样本的数值将其分成不同的 bin，并使用分位数作为 bin 的边界，以保证每个 bin 包含相似数量的样本。bin 的数量设置为超参数。

附录 B 附加实验

B.1 额外的消融研究

PLR 嵌入的影响。M 的结果欧登表 1 和表 2 中的 NCA 利用了数值特征的 PLR 嵌入 [21]（其中所有分类特征都以独热形式处理）。我们比较了 M欧登表 6 中 NCA 以及有或无 PLR 的 MLP。

结果表明，我们的 M欧登NCA 也可以利用 PLR 的优势实现进一步的改进。然而，与 MLP 不同的是，M欧登没有 PLR 的 NCA 仍然具有竞争力。例如，在回归数据集 CA 中，M欧登不带 PLR 的 NCA 表现优于 MLP 和 MLP公共关系. 结果验证了 M 的能力欧登NCA来自于其目标、架构和训练策略，而主要不是来自于PLR编码策略。

TabCon 的其他实现。如 A.4 小节所述，基于学习到的嵌入构建的预测器有不同的选择ϕ（$\cdot$）使用监督对比损失。我们在这里考虑两种变体：第一种是 KNN，第二种是用于分类的逻辑回归和用于回归的线性回归。我们分别将这两种预测方法表示为 TabCon 和 TabCon (LR)。我们在表 7 中比较了这两种变体的结果。结果表明，虽然不同的方法在数据集上的表现各不相同，但它们的整体性能是相当的。鉴于 KNN 的训练开销较低，我们将其用作默认配置。

可视化结果。为了更好地分析 M 的属性欧登NCA，我们将学习到的嵌入可视化ϕ（$\cdot$）的 M欧登NCA、TabCon 和 TabR 使用 TSNE [55]。如图 5 所示，所有深度表格方法都将嵌入空间转换为比原始特征更有助于分类或回归。TabCon 学习到的嵌入空间将同一类的样本聚类在一起，将不同类的样本分开，通常聚类

同一类实例归为一个簇。然而，它仍然难以区分一些难以区分的样本。TabR 和 M欧登另一方面，NCA 将同一类的样本划分为多个簇，确保相似的样本彼此更接近。此策略与 KNN 的预测机制一致，其中通过将具有相似邻居的实例聚类在一起而不是聚类到单个簇中来实现良好的性能。M 学习到的嵌入空间欧登NCA 比 TabR 学习到的更具判别性。主要原因是 TabR 利用类似 Transformer 的架构在进行预测之前修改每个实例的嵌入，这使得学习到的嵌入空间与 M 相比判别性较低欧登NCA。

B.2 运行时和内存使用情况估计

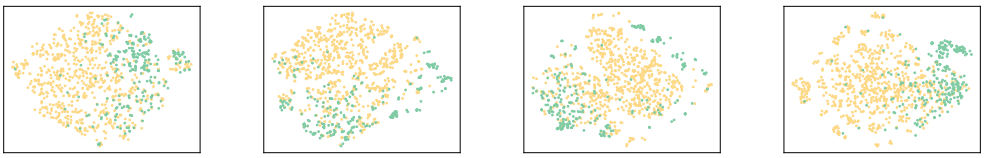
我们在图 1 中对运行时间和内存使用情况进行了比较。以下是我们进行估算的步骤。首先，我们对验证集上的所有模型进行了 100 次迭代调整，保存了迄今为止找到的最佳参数。接下来，我们使用调整后的参数对模型进行了 15 次迭代，并保存了验证集上的最佳检查点。使用调整后的模型在训练和验证阶段运行一个种子所需的平均时间来估计模型的运行时间。使用保存的检查点的大小来估计模型的大小。我们在表 8 中展示了基准数据集中运行时间和内存使用量估计的平均结果。

表 8：基准数据集上不同调整模型的训练时间和内存使用量估计。平均排名表示基于性能指标（回归的 RMSE 和分类的准确度）的这些模型的平均性能排名。

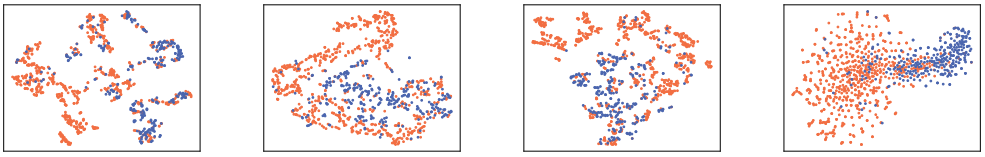
模型	M-NCA	TabCon	MLP	MLP公共关系	FT-T	TabR	XGBoost	CatBoost
训练时间（秒）	87.5	53.16	30.36	38.94	111.91	173.34	4.53	20.48
大小 (MB)	3.36	8.54	2.92	9.53	3.19	9.13	7.44	6.92
平均排名	2.41	5.80	7.63	4.83	5.32	3.37	3.62	3.02

B.3 带标准差的完整结果

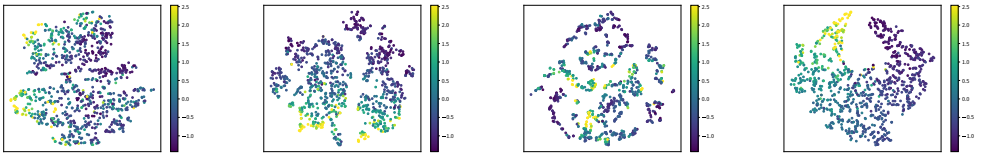
在本节中，我们展示了表 5 中所述数据集的整体结果及其标准差。分类任务的结果如表 9 所示。回归任务的结果如表 10 所示。



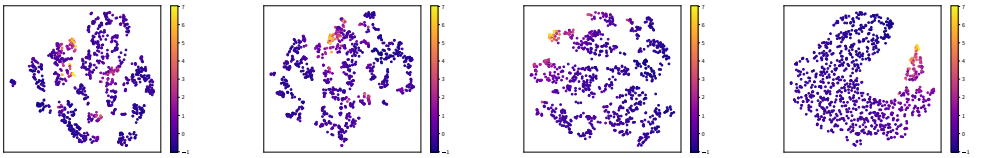
（一）广告 ↑原始特征　（坏的 ↑泰伯　（三）广告 ↑米欧登全国咖啡协会　（三）广告 ↑TabCon



（一） PH ↑原始特征　（二） PH ↑泰伯　（三）公共卫生 ↑米欧登全国咖啡协会　（三）公共卫生 ↑TabCon



（一）CA ↓原始特征　（二）CA ↓泰伯　（三）CA ↓米欧登全国咖啡协会　（三）CA ↓TabCon



（一）MIA ↓原始特征　（二）MIA ↓泰伯　（三）MIA ↓米欧登全国咖啡协会　（三）MIA ↓TabCon

图 5：不同方法的嵌入空间的可视化。

表 9：15 个种子中 12 个分类数据集的测试准确率和标准差。最佳结果以粗体突出显示。

↑	M-NCA	TabCon	XGBoost	CatBoost	MLP	光纤到户泰伯		KNN	多层感知器公共关系	肺动脉高压	LMNN	NCA
公元 87.29 年	86.97	87.21		87.48	85.80	85.89	86.86	84.61	86.97	85.53	83.21	83.87
	±0.07	±0.11	±0.05	±0.06	±0.14	±0.16	±0.10	±0.00	±0.12	±0.21	±0.00	±0.00
CU	79.99	77.72	81.73	81.03	75.54	80.65	79.61	70.08	80.75	75.89	72.12	73.78
	±0.55	±1.85	±0.17	±0.38	±0.57	±0.61	±1.18	±0.00	±0.67	±0.62	±0.00	±0.00
发光	96.12	87.44	92.06	91.69	85.49	87.50	96.46	84.27	86.93	85.10	83.56	85.15
	±0.11	±0.23	±0.21	±0.18	±0.28	±0.25	±0.12	±0.00	±0.61	±0.39	±0.00	±0.00
安永	99.41	82.88	71.02	71.88	61.29	71.49	98.20	59.46	74.27	60.13	54.98	66.50
	±0.33	±9.49	±0.42	±0.34	±0.98	±0.85	±0.54	±0.00	±2.28	±1.20	±0.00	±0.00
何氏	39.84	38.32	37.86	38.26	38.34	38.49	40.80	32.91	38.95	37.64	29.36	31.30
	±0.24	±0.23	±0.09	±0.11	±0.25	±0.22	±0.18	±0.00	±0.12	±0.27	±0.00	±0.00
你好	73.09	72.36	72.73	72.66	72.29	72.92	72.85	66.78		72.64	72.48	61.44	66.94
	±0.13	±0.17	±0.08	±0.11	±0.12	±0.17	±0.13	±0.00	±0.21	±0.14	±0.00	±0.00
知识产权	88.96	87.65	88.64	87.74	84.86	84.95	87.23	83.72	87.84	84.83	83.81	84.10
	±0.28	±0.56	±0.20	±0.08	±0.42	±0.34	±0.76	±0.00	±0.33	±0.28	±0.00	±0.00
JA	74.08	72.17	72.55	72.39	71.97	72.38	73.44	65.67		72.10	71.52	59.12	64.64
	±0.14	±0.37	±0.08	±0.09	±0.17	±0.16	±0.22	±0.00	±0.19	±0.27	±0.00	±0.00
马萨诸塞州	87.90	87.22	87.67	87.87	87.08	87.73	87.91	84.65		87.17	86.94	83.68	84.73
	±0.24	±0.22	±0.22	±0.17	±0.32	±0.28	±0.22	±0.00	±0.22	±0.17	±0.00	±0.00
加时赛	82.37	78.94	82.45	82.21	81.47	80.82	82.38	78.15	81.24	81.20	77.37	78.85
	±0.14	±0.49	±0.12	±0.12	±0.15	±0.28	±0.14	±0.00	±0.14	±0.21	±0.00	±0.00
肺动脉高压	89.29	87.88	87.66	87.97	86.88	87.86	89.10	86.86	87.13	85.54	85.75	87.05
	±0.63	±0.45	±0.53	±0.49	±0.72	±0.40	±0.49	±0.00	±0.49	±0.63	±0.00	±0.00
威斯康星州	74.74	73.72	74.39	74.59	72.50	72.64	74.22	77.10		72.50	72.16	73.58	72.60
	±1.55	±1.55	±0.85	±0.91	±0.50	±0.80	±1.29	±0.00	±1.35	±0.82	±0.00	±0.00
排名	2.000	5.917	3.833	3.750	8.000	5.417	3.083	10.000	5.667	9.250	11.333	9.750

表 10：15 个种子中 10 个回归数据集的测试 RMSE 和标准差。最佳结果以粗体突出显示。

↓	M-NCA	TabCon	XGBoost	CatBoost	多层感知处理器	光纤到户泰伯		KNN	多层感知处理器公共关系	
人工智能 ×10⁻³	.1532	.1522	1527年	.1465	.1558	.1565	.1546	.2431	1522年	1563年
	±.001	±.002	±.001	±0.00	±.002	±.003	±.002	±0.00	±.001	±.002
双 ×10²	.7057	.7352	.7180	.7264	.7773	.7466	.6657	1.087	.7155	.7193
	±.007	±.011	±.006	±.007	±.007	±.011	±.006	±0.00	±.003	±.005
加州	.4212	.4581	.4328	.4360	.5074	.4701	.4157	.5843	.4690	.5086
	±.005	±.007	±.002	±.005	±.005	±.004	±.005	±0.00	±.005	±.010
厘米 ×10³	.4618	.5011	.4809	.4929	.5131	.5187	.5091	.8754	.4655	.5211
	±.010	±.012	±.004	±.013	±.006	±.015	±.023	±0.00	±.020	±.004
CP ×10	.2381	.2473	.2404	.2365	.2490	.2310	.2337	.8825	.2468	.2487
	±.003	±.003	±.007	±.004	±.004	±.003	±.003	±0.00	±.006	±.003
福 ×10	.7272	.7436	.7384	.7393	.7900	.7885	.7398	.8140	.7428	.7896
	±.001	±.001	±0.00	±.001	±.001	±.001	±.013	±0.00	±.001	±.001
何 ×10⁵	.3041	.3156	.3026	.3051	.3161	.3074	.3133	.3663	.3093	.3149
	±.004	±.003	±.001	±.002	±.004	±.005	±.006	±0.00	±.003	±.003
侯 ×10⁵	.4291	.4963	.4797	.4546	.5221	.4791	.4149	.6889	.4921	.5118
	±.002	±.020	±.008	±.002	±.009	±.005	±.005	±0.00	±.004	±.004
洛杉矶 ×10³	.4135	.4529	.4448	.4506	.4789	.4609	.4606	.4968	.4753	.4777
	±.014	±.007	±.003	±.003	±.005	±.005	±.007	±0.00	±.013	±.005
米娅 ×10⁵	.8284	.9322	.8740	.8034	.8973	.8717	.8585	1.508	.8562	.9302
	±.017	±.014	±.016	±.008	±.011	±.023	±.012	±0.00	±.012	±.017
秩	2.200	6.100	3.500	3.100	8.300	5.900	3.600	10.000	4.500	7.800