注意注意可解释的表格学习

Sercan O¨.Arık，TomasPfister
谷歌云人工智能
加利福尼亚州森尼维尔
soarik@google.com, tpfister@google.com

摘要

我们提出了一种新的高性能和可解释的规范深度表格数据学习架构，TabNet。TabNet使用顺序注意从每个决策步骤中选择哪些特征，使可解释性和更有效的学习，因为学习能力用于最显著的特征。我们证明了TabNet在广泛的非性能饱和的表格数据集上优于其他变体，并产生了可解释的特征属性和对其全局行为的洞察力。最后，我们演示了对表格数据的自监督学习，当未标记数据丰富时，显著提高了性能。

介绍

深度神经网络(DNNs)在图像方面取得了显著的成功(He等人。，文本(Lai等。和音频(Amodei等人。2015).对于这些，有效地将原始数据编码为有意义的表示的规范架构，推动了快速进程。一种在规范体系结构中尚未取得如此成功的数据类型是表格数据。

尽管它是现实世界人工智能中最常见的数据类型（因为它由任何分类和数字特征组成），(Chui等人。2018年)，对表格数据的深度学习仍未得到充分探索，集成决策树(DTs)的变体仍然主导着大多数应用(Kaggle2019a)。为什么首先，因为基于dt的方法有一定的好处：(i)它们对于具有近似超平面边界的决策流形具有代表性的效率，这在表格数据中很常见；(ii)它们的基本形式是高度可解释的(e。g.通过跟踪决策节点)，有流行的事后解释方法的集成形式，例如。(伦德berg，Erion和Lee2018)——这在许多现实应用中是一个重要的问题；(iii)它们的训练速度很快。其次，因为以前提出的DNN架构并不太适合表格数据：例如。堆叠卷积层或多层感知器(MLPs)被极大地过度参数化——缺乏适当的归纳偏差往往导致它们无法找到表格决策流形的最优解决方案(古德费罗、Bengio和考维尔2016；沙维特和西格尔2018；Xu等人。2019).

为什么深度学习值得探索表格数据？一个明显的动机是预期的性能提高-

特别是对于大型数据集。2017).此外，与树学习不同的是，dnn能够实现表格数据进行梯度分散的端到端学习，这有很多的好处：(i)有效地编码多种数据类型，如图像和表格数据；(ii)减轻特征工程的需要，这是目前基于树的表格数据学习方法的一个关键方面；(iii)从流数据中学习，也许最重要的是(iv)端到端模型允许表示学习，从而使许多有价值的应用场景成为可能，包括数据高效的领域适应（古德费罗、本吉奥和考维尔2016年）、生成建模（雷德福、梅茨和钦塔拉2015年）和半监督学习(Dai等人。2017).

我们为表格数据提出了一种新的规范DNN架构，TabNet。主要贡献总结如下：

1. *TabNet输入没有任何预处理的原始表格数据，并使用基于梯度下降的优化进行训练，使其能够灵活地集成到端到端学习中。*
2. TabNet使用顺序注意从每个决策步骤中选择哪些特征，使可解释性和更好的学习，因为学习能力用于最显著的特征(见图。1).这个特征选择是基于实例的，e。g.每个输入可能不同，不同于其他实例级的特征选择方法，如(Chen等。*2019年)，TabNet采用了单一的深度学习架构来进行特征选择和推理。*
3. 以上的设计选择导致了两个有价值的特性：(i) *在不同领域的分类和回归问题上，TabNet优于或与其他表格学习模型；(ii)TabNet支持两种可解释性：局部可解释性可视化特征的重要性和它们的组合，以及全局可解释性量化每个特征对训练模型的贡献。*
4. 最后，对于表格数据，我们首次通过使用无监督的预训练来预测掩蔽特征，显示出了显著的性能改善(见图。2).

相关工作

特征选择：特征选择主要是指根据特征的预测有用性，明智地选择特征子集。常用的技术，如为-

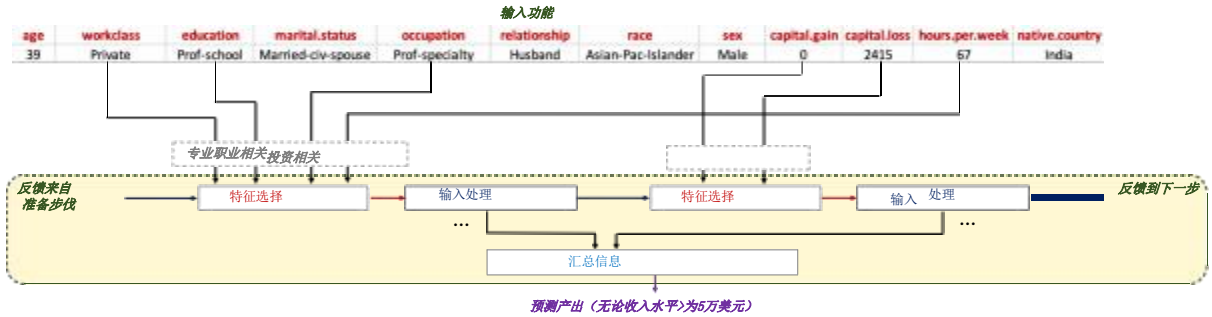

图1：TabNet的稀疏特征选择为成人人口普查收入预测的样本(Dua和Graff2017)。稀疏特征选择使可解释性和更好的学习，因为容量被用于最显著的特征。TabNet使用多个决策块，它们专注于处理输入特征的一个子集以进行推理。两个决策块分别处理与专业职业和投资相关的特征，以预测收入水平。

图2：自我监督的表格学习。现实世界的表格数据集有相互依赖的特征列，例如，教育水平可以从职业中猜到性别，或者可以从关系中猜到性别。无监督表示学习的蒙面自监督学习的结果在一个改进的编码器模型的监督学习任务。

病房选择和套索正则化(Guyon和Elisseeff2003)基于整个训练数据的属性特征重要性，并被称为全局方法。*实例级特征选择是指为每个输入单独选择特征，在(Chen等人中研究*。使用一个解释模型，最大化所选特征和响应变量之间的互信息，并在(Yoon，Jordon和vanderSchaar2019)中通过使用演员-评论家框架模拟优化选择的基线。与这些不同的是，TabNet在端到端学习中采用具有可控稀疏性的软特征选择——单一模型联合执行特征选择和输出映射，从而以紧凑的表示具有优越的性能。

基于树的学习：DTs是常用的表格数据学习。它们的突出优势是有效地选择具有最多统计信息收益的全球特征（格拉布谢夫斯基和扬科夫斯基，2005年）。为了提高标准dt的性能，一种常见的方法是集成以减少方差。在集成方法中，随机森林(Ho1998)使用随机数据的随机子集选择的特征来生长许多树。XGBoost(Chen和Guestrin2016)和LightGBM(Ke等。是在最近的数据科学竞赛中占据主导地位的两种近期的集成DT方法。我们对不同数据集的实验结果表明，在深度学习提高表示能力的同时，在保留其特征选择特性的同时，该模型可以优于基于树的模型。将dnn集成到DTs中：用DNN构建块表示DTs，如(Humbird、彼得森和麦克拉伦2018)，在表示中产生冗余和低效的学习。软（神经）DTs(Wang，Aggarwal和Liu2017；康齐德等。使用可微的决策函数，而不是不可微的轴对齐的分割。然而，失去自动的特征选择往往会降低性能。在(杨、莫里洛和Hoseredales2018)中，提出了一个软箱函数来模拟dnn中的DTs，通过不有效地枚举所有可能的决策模拟DTs。(Ke等。2019年)提出了一种通过明确利用表达性特征组合的DNN架构，然而，学习是基于从梯度增强的DT中转移知识。端野

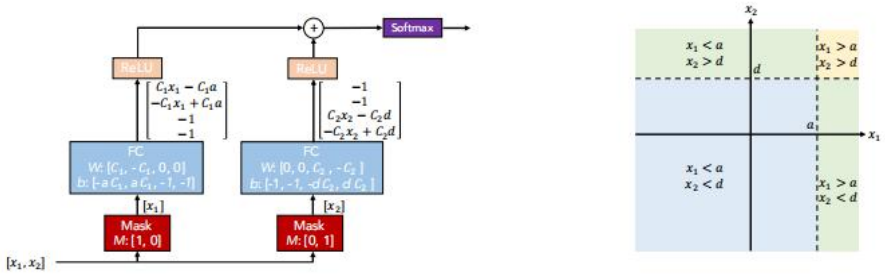

图3：使用传统的DNN块（左）和相应的决策流形（右）进行的dt类分类的说明。通过对输入端使用乘法稀疏掩模来选择相关特征。对所选的特征进行线性变换，在添加偏差（表示边界）后，ReLU通过归零区域来进行区域选择。多个区域的聚合是基于加法的。作为C_1和C_2越大，决策边界就越清晰。

以及其他人2018)提出了一种DNN架构，通过从原始块自适应增长，同时表示学习到边缘、路由函数和叶节点。

TabNet的不同之处在于，它通过顺序注意嵌入了具有可控稀疏性的软特征选择。

自监督学习：无监督表示学习改进了监督学习，特别是在小数据情况下(Raina等。2007).最近的文本工作(Devlin等人。和图像(Trinh，Luong和Le2019)数据显示了显著进展——受无监督学习的明智选择目标（蒙面输入预测）和基于注意力的深度学习的驱动。

表格学习的TabNet

DTs可以成功地从真实世界的表格数据集中学习。通过一个特定的设计，传统的DNN构建块可以用于实现类似于dt的输出流形，例如。见图。3).在这种设计中，个体特征选择是获得超平面形式的决策边界的关键，可以推广到特征的线性组合，其中系数决定了每个特征的比例。TabNet是基于这种功能的，它优于dt，同时通过精心的设计：(i)使用从数据中学习的稀疏实例特征选择；(ii)构建一个连续的多步骤体系结构，每一步都对基于所选特征的部分决策有贡献；(iii)通过对所选特征的非线性处理提高学习能力；(iv)通过更高的维度和更多的步骤模拟集成。

图4显示了用于编码表格数据的TabNet架构。我们使用原始的数值特征，并考虑了具有可训练嵌入的分类特征的映射。我们不考虑任何全局特征归一化，而只是应用批处理归一化(BN)。我们通过了相同的二维特征f2必$^{bd\times}$到每个决策步骤，其中B是批处理大小。TabNet的编码是基于使用N的顺序多步处理步骤决策步骤。我th步骤从（i-1）中输入处理后的信息th步骤，决定使用哪些特性，并将处理过的特性表示聚合到整个决策中。自上而下的顺序形式的注意的想法的灵感来自于其在处理视觉和文本数据(Hudson和曼宁2018)和强化学习-(Mott等人。而在高维输入中搜索相关信息的一个小子集。

特征选择：我们使用了一个可学习的掩模M[i]2必$^{bd\times}$用于显著特征的软选择。通过对最显著特征的稀疏选择，决策步骤的学习能力不会浪费在无关的特征上，从而使模型的参数效率更高。掩蔽是乘法的。我们使用一个细心的变压器(见图。4)使用前一步处理后的特征获得掩模，a[i-1]：M[i]=稀疏矩阵(P[i-1]·h_i(a[i — 1])).稀疏归一化(马tins和Astudillo2016)通过将欧几里得投影映射到概率单纯形上来鼓励稀疏性，它在性能上更优越，并与稀疏特征选择的目标相一致

能力注意对$\sum_{j=1}^{D}$M[i]b，j= 1.h_i是一个可训练的函数，如图所示。4.使用FC层，然后是BN层。P[i]是之前的比例项，表示一个特定的特征以前被使用了多少：P[i]=$\prod_{j=1}^{u_i}$(V-M[j])，其中V是一个松弛参数——当V=1时，一个特性被强制要求只在一个决策步骤中使用，并且随着V的增加，在多个决策步骤中使用一个特性提供了更多的灵活性。P[0]被初始化为所有的[0]，$1^{bd\times}$，没有任何先前的蒙面功能。如果一些特征未使用（如在自监督学习中），相应的P[0]条目为0，以帮助模型的学习。为了进一步控制所选特征的稀疏性，我们提出了熵形式的稀疏性正则化(爷爷和Bengio2004)，L稀疏=

对 对 对$\sum_{i=1}^{N_{steps}}\sum_{b=1}^{B}\sum_{j=1}^{D}\frac{-M_{b,j}[i]\log(M_{b,j}[i]+\epsilon)}{N_{steps}\cdot B}$，其中e是a对于数值稳定性的数目很小。我们将稀疏性正则化添加到总体损失中，其系数为λ稀疏.稀疏性为大多数特征是冗余的数据集提供了一个有利的归纳偏差。

特征处理：我们使用一个特征转换器来处理过滤后的特征(见图。4)，然后拆分决策步骤的输出和后续步骤的信息，[d[i]，a[i]]=f_i(M·f)，其中有2个必$^{十亿\times}$d和[i]2必$^{十亿\times}$a.为了实现高容量的参数高效和鲁棒学习，特征转换器应该包含在所有决策步骤中共享的层（因为在不同的决策步骤中输入相同的特征），以及与决策步骤相关的层。图4显示了实现

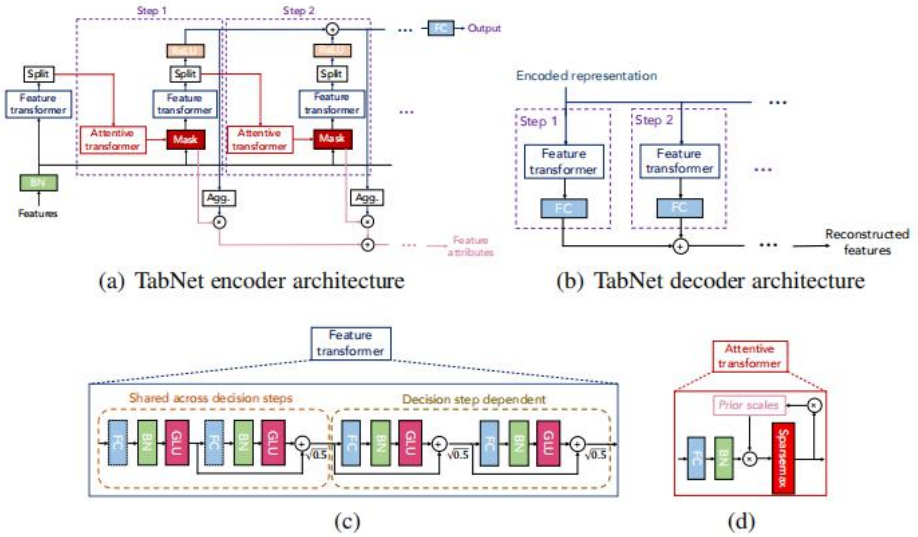

图4：(a)TabNet编码器，由特征变压器、注意变压器和特征掩蔽组成。分割块将处理过的表示划分给后续步骤的注意变压器以及整体输出。对于每一步，特征选择掩码提供了关于模型功能的可解释信息，并且可以将掩码聚合以获得全局特征重要属性。(b)TabNet解码器，由每一步的特征变压器块组成。(c)一个特征变压器块的例子-4层网络显示，其中2是在所有的决策步骤中共享的，2是依赖于决策步骤的。每一层都由一个全连接(FC)层、BN和GLU非线性组成。(d)一个注意变压器块的例子——单层映射使用先前的尺度信息进行调制，该信息聚合了在当前决策步骤之前每个特征被使用的数量。利用电火花线(马丁斯和Astudillo2016)对系数进行了归一化，从而得到了显著特征的稀疏选择。

作为两个共享层和两个决策步长依赖层的连接。每一层FC层之后都是BN和门控线性单元(GLU)的非线性度(Dauphin等。2016年)，最终采用标准化的标准化残余连接。归一化$\sqrt{0}.5$通过确保整个网络的方差没有显著变化，有助于稳定学习(Gehring等人。2017).为了更快的训练，我们使用了大批量的BN。因此，除了应用于输入特性的一个，我们使用鬼BN(Hoffer，Hubara，和Soudry2017)形式，使用虚拟批大小B_V和动量m_B.对于输入特征，我们观察到低方差平均的好处，因此避免了幽灵BN。最后，受到决策树类聚合的启发，如图所示。3、我们构建了整体的决策嵌入如$d_{出局}=\sum_{i=1}^{N_{steps}}\text{ReLU}(d[i])$.我们应用一个线性映射$W_{final}$来获取输出映射。[1]

可解释性：TabNet的特征选择掩模可以阐明在每一步中所选择的特征。如果$M_{b,j}[i]$=为0，然后是j^{th}b的特征th样本对决定没有贡献。如果f_i是一个线性函数，系数M吗$_{b,j}[i]$将对应于f的特征重要性$_{b,j}$.虽然每个决策步骤都采用非线性处理，但它们的输出稍后会以线性的方式组合。我们的目标是除了在对每个步骤的分析之外，还要量化聚合特征的重要性。在不同的步骤中组合这些口罩需要一个系数来权衡决策中每一个步骤的相对重要性。我们只是提出$\eta_b[i]=\sum_{c=1}^{N_d}\text{ReLU}(d_{b,c}$表示在i处的总决策贡献th对b的决策步骤th样品直观地说，如果$d_{b,c}[i]<0$，然后所有特性th决策步骤应该有0对整体决策的贡献。随着其值的增加，它在整体线性组合中发挥着更高的作用。在每个决策步骤中缩放决策掩码$\eta_b[i]$，我们提出了聚合特征重要性掩模，$M_{agg-b,j}=\sum_{i=1}^{N_{steps}}\eta_b[i]M_{b,j}[i]\Big/\sum_{j=1}^{D}\sum_{i=1}^{N_{steps}}\eta_b[i]M_{b,j}[i].$[2]

表格自监督学习：我们提出了一个解码器从标签中重建表格特征的架构-Net编码表示。解码器由特征转换器组成，在每个决策步骤中都有FC层。对输出结果进行求和，得到重构的特征。我们提出了预测其他特征列的任务。考虑一个二进制掩码$b\in\{0,1\}^{B\times D}$.TabNet编码器输入(1-S)，解码器输出重构的特征S。$\hat{f}$我们在编码器中初始化P[0]=(1-S)，使模型只强调已知特征，解码器最后的FC层与S相乘，输出未知特征。我们考虑了自监督阶段的重建损失：

$$\sum_{b=1}^{B}\sum_{j=1}^{D}\left|(\hat{f}_{b,j}-f_{b,j})\cdot\text{所罗门群岛}_{b,j}\right|^2\Big/\sum_{b=1}^{B}(f_{b,j}-1/B\sum_{b=1}^{B}f_{b,j})^2.$$ 规范化

[1]对于离散输出，我们在训练期间另外使用softmax(在推理期间使用argmax)。

[2]规范化用于确保工$M_{agg-b,j}=1.$

表1：平均值和std。对6个合成数据集的接收工作特征曲线(AUC)下的测试面积。2018年)，用于TabNetvs。其他基于特征选择的DNN模型：无sel。：使用所有功能而不需要任何功能选择，
全局：仅使用全局显著特征，树集合(Geurts，Ernst，和Wehenkel2006)，套索正则化模型，L2X(Chen等。2018年)和INVASE(Yoon，Jordon和范德Schaar2019年)。粗体数字表示每个数据集的最佳值。

模型	*测试AUC*					
	赛恩1	Syn2	赛恩3	赛恩4	赛恩5	赛恩6
没有选择	.578±.004	.789±.003	.854±.004	.558±.021	.662±.013	.692±.015
树	.574±.101	.872±.003	.899±.001	.684±.017	.741±.004	.771±.031
套索正则化的	.498±.006	.555±.061	.886±.003	.512±.031	.691±.024	.727±.025
L2X	.498±.005	.823±.029	.862±.009	.678±.024	.709±.008	.827±.017
内陷	**.690±.006**	.877±.003	**.902±.003**	**.787±.004**	.784±.005	.877±.003
全球	.686±.005	.873±.003	.900±.003	.774±.006	.784±.005	.858±.004
表网	.682±.005	**.892±.004**	.897±.003	.776±.017	**.789±.009**	**.878±.004**

与总体标准差的地面真实值是有益的，因为特征可能有不同的范围。我们样本$S_{b,j}$独立于一个参数为p的伯努利分布s，每次迭代。

实验

我们研究TabNet的范围广泛的问题，其中包含回归或分类任务，特别是与已发布的基准测试。对于所有的数据集，分类输入被映射到一个具有可学习嵌入的单维可训练标量[3]和数字列的输入没有经过和预处理。[4]我们使用标准分类(softmax交叉熵)和回归（均方误差）损失函数，我们训练直到收敛。TabNet模型的超参数在一个验证集上进行了优化，并在附录中列出。TabNet的性能对大多数超参数都不是很敏感。在附录中，我们还介绍了各种设计和关键超参数选择指南的消融研究。对于我们引用的所有实验，我们使用与原始工作相同的训练、验证和测试数据。采用Adam优化算法(KingmaandBa2014)和Glorot均匀初始化方法对所有模型进行训练。[5]

基于实例的功能选择

显著特征的选择对于高性能至关重要，特别是对于小数据集。我们考虑了6个表格数据集。(包括10k个训练样本)。数据集的构造方式只有特征的子集决定输出。对于Syn1Syn3，所有实例的显著特征都是相同的(e。g.，Syn2的输出取决于特性X_3-X_6)，而全局特征选择，如已知的显著特征，将给予高性能。对于Syn4-Syn6，显著的特征是依赖于实例(例如，对于Syn4，输出依赖于任何一个X_1-X_2或X_3-X_6这取决于X的值11)，使得全局特征选择不优。表1显示，TabNet的性能优于其他系统(树集成(Geurts、恩斯特和温克尔2006)、套索正则化、L2X(Chen等。)，并与INVASE(Yoon，Jordon和vanderSchaar2019)持平。对于Syn1-Syn3，TabNet的性能接近于全局特性选择——它可以找出哪些特性是全局重要的。对于Syn4-Syn6，消除了实例级的冗余特征，TabNet改进了全局特征选择。其他方法均采用具有43k个参数的预测模型，参数总数为101k的INVASE，由于另外两个模型在演员-批评者的框架。TabNet是一个单一的架构，Syn1-Syn3的大小为26k，Syn4-Syn6的大小为31k。紧凑的表示是TabNet有价值的属性之一。

在真实世界的数据集上的性能

表2：森林覆盖类型数据集的性能。

模型	*测试精度（%）*
XGBoost	89.34
LightGBM	89.28
猫的提升	85.14
AutoML表	94.95
表网	**96.99**

森林覆盖类型(Dua和Graff2017)：任务是从地图变量中对森林覆盖类型进行分类。表2显示，TabNet优于基于集成树的方法，这些方法可以实现稳定的性能(Mitchell等。2018).我们还考虑了AutoML表(AutoML2019)，一个基于模型集成的自动搜索框架，包括DNN，梯度增强DT，AdaNet(Cortes等人。和集成(AutoML2019)与非常彻底的超参数搜索。没有细粒度超参数搜索的单一表网的性能优于它。

扑克牌(Dua和Graff2017)：任务是从扑克牌的等级属性进行分类。输入-输出关系是确定性的和

[3]在某些情况下，高维嵌入可能会略微提高性能，但对单个维度的解释可能会变得具有挑战性。

[4]特别设计的功能工程，例如。变量高偏态分布的对数变换，可能会进一步改善结果，但我们将它排除在本文的范围之外。

[5]我们将发布一个开源的实现。

表3：扑克手感应数据集的性能。

模型	*测试精度（%）*
dt	50.0
mlp	50.0
深神经DT	65.1
XGBoost	71.1
LightGBM	70.0
猫的提升	66.6
表网	**99.2**
基于规则的	100.0

手工制作的规则可以获得100%的准确性。然而，传统的dnn、DTs，甚至它们的深度神经DTs的混合变体（杨、莫里洛和医院达莱斯2018年）严重遭受数据不平衡的影响，无法学习所需的排序和排序操作（杨、莫里洛和医院达莱斯2018年）。调优的XGBoost、CatBoost和LightGBM显示出了非常轻微的改进。TabNet的性能优于其他方法，因为它可以通过其深度进行高度非线性的处理，而不会由于实例级的特征选择而进行过拟合。

表4：在Sarcos数据集上的性能。三个TabNetmod考虑了不同尺寸的els。

模型	*测试MSE*	*型号大小*
随机森林	2.39	16.7K
随机DT	2.11	28K
mlp	2.13	0.14M
自适应神经树	1.23	0.60M
渐变增强树	1.44	0.99M
TabNet-S	**1.25**	**6.3K**
TabNet-M	**0.28**	**0.59M**
TabNet-L	**0.14**	**1.75M**

Sarcos(维贾亚库马尔和Schaal2000)：这项任务是对一个拟人化的机器人手臂的逆动力学进行回归。(Tanno等人。2018年)表明，一个非常小的模型可以获得良好的性能。在非常小的模型规模范围下，TabNet的性能与Tanno等。）的参数增加了100倍。当模型大小不受约束时，TabNet几乎达到了一个数量级的较低的测试MSE。

表5：在希格斯玻色子数据集上的性能。两个TabNet模型用-S和-M表示。

模型	*测试acc。(%)*	*型号大小*
稀疏进化MLP	**78.47**	**81K**
梯度增强树-S	74.22	0.12M
梯度增强树-M	75.97	0.69M
mlp	78.44	2.04M
梯度增强树-L	76.98	6.96M
TabNet-S	78.25	81K
TabNet-M	**78.84**	**0.66M**

希格斯玻色子(Dua和Graff2017)：这项任务是区分希格斯玻色子过程和。背景由于其更大的尺寸（1050万个实例），dnn即使在非常大的集成中也优于DT变体。TabNet在更紧凑的表示方式上优于mlp。我们还比较了最先进的进化稀疏化算法(Mocanu等人。将非结构化稀疏性集成到训练中。由于其紧凑的表示，TabNet在相同数量的参数下产生的性能几乎与稀疏进化训练相似。与稀疏进化训练不同，TabNet的稀疏性是结构化的——它不会降低操作强度(Wen等。并能有效利用现代多核处理器。

表6：罗斯曼商店销售数据集的性能。

模型	*测试MSE*
mlp	512.62
XGBoost	490.83
LightGBM	504.76
猫的提升	489.75
表网	**485.12**

罗斯曼商店销售(Kaggle2019b)：该任务是从静态和时变的特性来预测商店销售。我们观察到TabNet优于常用的方法。时间特征(例如。天)获得了很高的重要性，并且在销售动态不同的假期等情况下，可以观察到实例级特性选择的好处。

可解释性

合成数据集：图。5显示了表1中合成数据集的聚合特征重要性掩码。[6]Syn2上的输出只依赖于X_3-X_6我们观察到，无关特征的聚合掩模几乎都为零，而TabNet只关注相关的特征。对于Syn4，输出依赖于任何一个X_1-X_2或X_3-X_6这取决于X的值11.TabNet可以产生精确的实例级特征选择——它分配了一个掩码来聚焦于指示器X11，并将几乎所有零的权重分配给不相关的特征（两个特征组以外的特征）。

真实世界的数据集：我们首先考虑一个简单的任务
蘑菇可食性预测(Dua和Graff2017)。TabNet在该数据集上达到了100%的测试精度。众所周知(Dua和Graff2017)，“气味”是最具区别性的特征——只有“气味”，一个模型可以获得>98.5%的测试精度(Dua和Graff2017)。因此，我们期望它具有很高的特性重要性。TabNet为其分配的重要性分数为43%，而其他方法如石灰（里贝罗、辛格和格斯特林2016）、综合梯度（孙达拉扬、塔利和严2017）和深度提升（施里库玛、格林赛德和昆达杰2017）分配的重要性分数低于30%(易卜拉欣等人。2019).接下来，我们考虑成人人口普查的收入。表网

[6]为了更好地说明这里，模型用10M样本而不是10K进行训练，因为我们获得了更清晰的选择掩模。

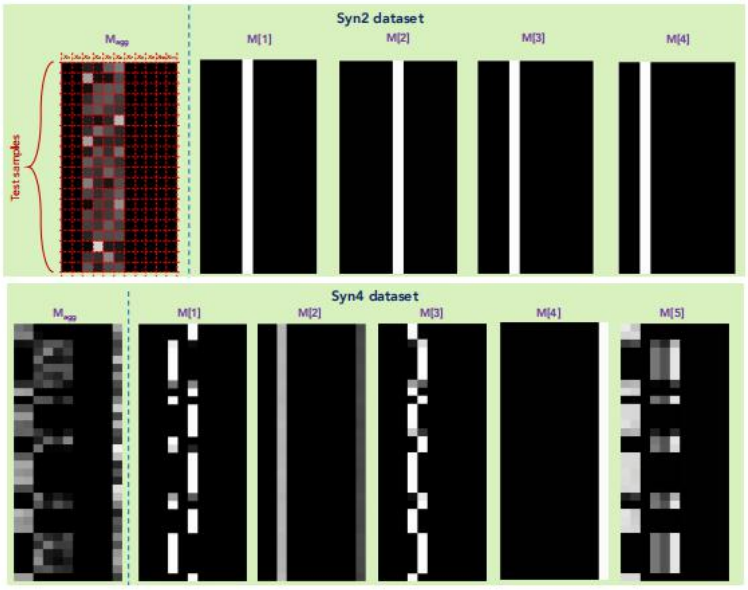

图5：特征重要性掩码M[i](表示在i处的特征选择th和聚合特征重要性掩码M_{agg}在Syn2和Syn4上显示了全局实例级的特征选择(Chen等人。2018).颜色越亮，表示值越高。E.g.对于Syn2，只有X_3-X_6使用。

.

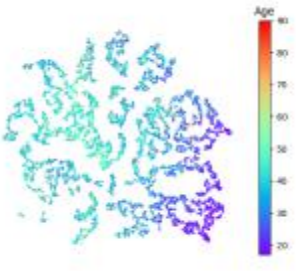

图6：成人决策流形的T-SNE的前两个维度和顶部特征“年龄”的影响。

产量特征重要性排名与众所周知的(伦德伯格，Erion和Lee2018；Nbviewer2019)（见附录）相同的问题，图。6显示了不同年龄组之间的明显分离，如“年龄”是TabNet最重要的特征。

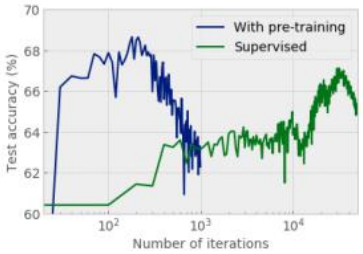

图7：在带有10k个样本的希格斯数据集上的训练曲线。

自我监督学习

表7：平均值和std。使用Tabnet-M模型的希格斯粒子的精度（超过15次运行），改变训练数据集的大小以进行监督微调。

培训数据集大小	*测试精度（%）*	
	监督	*有预培训*
1k	57.±47 1.78	**61.37 0.88±**
10k	66.66 0.88±	**68.06 0.39±**
100k	72.±92 0.21	**73.±19 0.15**

表7显示，无监督预训练显著提高了有监督分类任务的性能，特别是在未标记数据集比标记数据集大得多的情况下。如图所示。7在无监督预训练下，模型的收敛速度要快得多。非常快的收敛可以用于持续学习和领域适应。

结论

我们提出了一种新的用于表格学习的深度学习架构。TabNet使用顺序注意机制来选择语义上有意义的特征子集，在每个决策步骤中进行处理。实例式的特征选择能够高效学习，因为模型容量被充分用于最显著的特征，并通过选择掩模的可视化产生更多可解释的决策。我们证明了TabNet在来自不同领域的表格数据集上优于以前的工作。最后，我们展示了无监督的预训练对快速适应和提高性能的显著好处。

致谢

与尹金成的讨论，长T。感谢乐、宋贤、凯鹏华盈、张思昭、库兹涅佐夫、陈星、高杉和摩尔。

伦理影响

表格数据是真实世界人工智能中最常见的数据类型(Chui等。2018).表格数据学习问题出现在医疗保健、能源、金融、零售、零售、制造、物理科学等关键人工智能应用中。TabNet是一种新型的深度神经网络架构，可以提高表格数据学习的性能，同时也为其推理提供了可解释的见解。为了强调其广泛的适用性，在本文中，我们展示了TabNet在环境科学、物理、零售、机器人和公共部门等广泛应用领域的强大结果。

除了强大的性能之外，TabNet还提供了关于其本地和全局推理的可解释的见解。在一些主要的表格数据应用程序中，由于监管原因或非技术用户的期望，透明度至关重要。例如，医生应该知道为什么人工智能模型会建议某种特定的治疗方法，或者信贷员应该知道为什么人工智能模型会将客户标记为高违约风险。事实上，由于这个原因，深度学习无法深入到这种透明敏感的表格学习应用程序中。我们相信TabNet将是这个方向上的一个重要贡献（尽管它并不构成完整的解决方案）。TabNet的特征归因掩码可以阐明模型对每个实例分别用于推理的特征，并有助于为决策者建立对模型的信任，也可以为监管当局提供指导。这些见解也可以被数据科学家用来通过特征工程来提高模型的性能。

最后但并非最不重要的是，我们首次展示了对表格数据的自监督学习的潜力。自我监督学习是近年来最活跃的人工智能研究领域之一，但几乎所有的文献都集中于文本或图像数据。我们证明了真实世界的表格数据集也有一个基于TabNet的自监督学习框架可以有效利用的结构化信息。我们在无监督的预培训中显示了显著的性能改善，我们希望这一方向能够提高人工智能对人工标签非常昂贵的应用程序的渗透（例如，医疗保健、金融或零售，因为人类标签人员应该拥有特定领域的专业知识）。

参考文献

Amodei, D.; Anubhai；E；情况下， C.; 卡斯珀，J.；等人。2015.深度语音2：英语和普通话中的端到端语音识别。*arXiv:1512.02595*。

AutoML。2019.AutoML表-谷歌云。紫外线https://cloud.谷歌。com/automl-tables/.

陈J.；宋，L.；温赖特，M。J.；乔丹，M。I.2018.学习解释：一个关于模型解释的信息理论的视角。*arXiv:1802.07814*。

陈，t.；和格斯特林，C。2016.XGBoost：一个可伸缩的树形提升系统。*在KDD。*

崔先生；Manyika，J.；Miremadi，M.；亨克，n.；钟，R。以及其他人2018.来自人工智能前沿的笔记。*麦肯锡全球机构*

Cortes，c.；贡扎尔沃，X.；库兹涅佐夫，诉；Mohri，M.；和杨，S。2016.人工神经网络的自适应结构学习。*arXiv:1607.01097*。

戴z；杨，z；杨F.；科恩，W。W.；和萨拉克胡蒂诺夫。2017.好的半监督学习，需要一个坏的学习。*arxiv:1705.09783*。

多芬，Y。N.；风扇， A.; 奥利，M.；和格兰奇尔，D.2016.使用门控卷积网络进行语言建模。*arXiv:1612.08083*。

德夫林，J。；张，m.；李，K.；和图塔诺瓦，K。2018.对语言理解的深度双向转换器的预培训。*arXiv:1810.04805*。

Dua, D.; 和格拉夫， C.2017.UCI机器学习存储库。URL http://archive.ics.uci.edu/ml.

Gehring，J。；奥利，M。；灰色， D.; Yarats, D.; 和多芬，Y。N.2017.卷积序列到序列的学习。*arXiv:1705.03122*。

Geurts；恩斯特；和温克尔，l。2006.非常随机的树。*机器学习63（1）：3-42。*国际标准期刊编号1573-0565.

好朋友，我；Bengio，Y.；和考维尔。2016.*深度学习*麻省理工学院出版社。

格拉布谢夫斯基；和扬考斯基，N。2005.具有决策树准则的特征选择。*在他的。*

Grandvalet, Y.；和孟加拉，Y。2004.基于熵最小化的半监督学习。*在NIPS。*

盖，我；和艾丽舍夫。2003.介绍变量和特征选择。JMLR 3: 1157 – 1182.

他K。张x；任年代；和孙，J。2015.图像识别中的深度残差学习。*arXiv:1512.03385*。

赫斯特内斯，J.；Narang，美国；阿达拉尼，N.；Diamos，G。F.；六月，H.；Kianinejad，H.；帕特里，M。M.A.；杨，Y.；和周，Y。2017.根据经验证明，深度学习尺度是可预测的。*arXiv:1712.00409*。

Ho, T.K.1998.构造决策森林的随机子空间方法。PAMI 20(8): 832 – 844.

你好；胡巴拉，我；和Soudry，D。2017.训练时间更长，泛化效果更好：缩小神经网络大批量训练中的泛化差距。*arXiv:1705.08741*。

哈德逊，D。A.; 和曼宁， C.D.2018.用于机器推理系统的组合注意网络。*arXiv:1803.03067*.

Humbird, K.D.; 彼得森，J。L.；麦克拉伦，R。G.2018.使用决策树的深度神经网络初始化。*IEEE跨神经网络和学习系统。*

易卜拉欣M.；路易Modarres，C；和佩斯利，J。W.2019.神经网络的全球解释：绘制预测的景观。*arxiv:1902.02384 .*

卡格尔。2019a.卡格尔的历史数据科学趋势。
https://www.kaggle.com/shivamb/data-science-trendson-kaggle.访问时间：2019年4月20日。

卡格尔。2019b.罗斯曼商店销售。
https://www.kaggle.com/c/罗斯曼-商店销售。访问：20191110。

柯，G.；孟，Q.；芬利，T.；王，T.；陈，W.；etal。2017.LightGBM：一种高效的梯度提升决策树。*在NIPS*。

柯G.；张J.；徐z；边J.；和刘。.-2019.TabNN：一种针对表格数据的通用神经网络解决方案。URL https://openreview.net/forum?id=r1eJssCqY7.

金玛，D。P.；和Ba，J。2014.一种随机优化的方法。*在ICLR*。

康齐德，p；FiterauM.；犯罪；和Bul，S。R.2015.深层神经决策森林。*在ICCV*。

赖s；徐l.；刘K.；和赵，J。2015.用于文本分类的递归卷积神经网络。*在AAAI*。

伦德伯格，S。M.；ErionG。G.；和李，S。2018.树集合的一致的个性化特征属性。*arXiv:1802.03888 .*

马丁，A。F.T.；和阿斯图希略，R。F.2016.从Softmax到稀疏模型：注意和多标签分类的稀疏模型。*arXiv:1602.02068 .*

米切尔，R.；Adinets，一个；饶，T.；和弗兰克，E。2018.可扩展的GPU加速学习。*arXiv:1806.11248 .*

莫卡， D.；Mocanu，E。；石头，P.；阮；P.；Gibescu，M.；和Liotta， A.2018.受网络科学的启发，具有自适应稀疏连接性的人工神经网络的可扩展训练。*自然通信9*。

丛林 A.；佐兰公司， D.；克扎诺夫斯基；维斯特拉，D.；和德， D.J.2019.S3TA：一个软的、空间的、顺序的、自顶向下的注意模型。URL https://openreview.net/forum?id=B1gJOoRcYQ.

nbveer。2019.Nbveer上的笔记本。紫外线 https://nbviewer.朱皮特。org/github/dipanjanS/data_科学_为了_all/blob/master/tds_模型_解释_xai/Human-interpretableMachineLearning-DS.ipynb#.

雷德福德；梅茨，l.；和钦塔拉，S。2015.使用深度卷积生成对抗网络的无监督表示学习。*arXiv:1511.06434 .*

Raina。战役 A.；李，H.；包装器， B.；和Ng，A.Y.2007.自学学习：从未标记的数据中转移学习。*在ICML*。

里贝罗；辛格；和Guestrin，C。2016.我为什么要相信你？：解释任何分类器的预测。*在KDD*。

沙维特，我；和西格尔，E。2018.正则化学习网络：对表格数据集的深度学习。

尖叫， A.；绿边，P.；还有昆达杰， A.2017.通过传播激活差异来学习重要的特征。*arXiv:1704.02685 .*

孙达拉拉詹先生；他说；和燕，Q。2017.深度网络的公理化归因。*arXiv:1703.01365 .*

Tanno，R.；阿鲁库马兰，k.；亚历山大，D。C.；Criminisi，A.；和诺， A.V.2018.自适应神经树。*arXiv:1807.06699 .*

Trinh。H.；Luong，m.；和勒，Q。V.2019.自拍：进行图像嵌入的自我监督预训练。*arXiv:1906.02940 .*

维jayakumarS；Schaal，S。2000.局部加权投影回归：一种高维空间增量实时学习的O(n)算法。*在ICML*。

王，s.；Aggarwal，C.；和刘，H。2017.利用随机森林来激发神经网络并改进它。在*sdm*。

温W.；吴C.；王Y.；陈Y.和李，H。2016.学习深度神经网络中的结构化稀疏性。*arXiv:1608.03665 .*

徐，L。；斯库拉里杜，M.；和韦拉马切内尼，K。2019.使用条件GAN建模表格数据。*arXiv:1907.00503 .*

YangY.；莫里洛，我。G.；和酒店，T。M.2018.深度神经决策树。*arXiv:1806.06988 .*

Yoon，J.；乔丹，J.；和范德沙尔，M。2019.INVASE：使用神经网络进行实例级变量选择。*在ICLR*。