小 型 微 型 计 算 机 系 统
Journal of Chinese Computer Systems
DOI:10.20009/j.cnki.21-1106/TP.2025-0083
2026 年 4 月 第 4 期
Vol.47 No.4 2026

有序距离优化与特征模糊增强的表格有序分类方法

罗正东,艾比布拉·阿塔伍拉,张国昊,韩云飞,刘金龙,王 轶,周 喜

(中国科学院 新疆理化技术研究所,乌鲁木齐 830011)
(中国科学院大学,北京 100049)
(中国科学院 新疆民族语音语言信息处理重点实验室,乌鲁木齐 830011)
E-mail:luozhengdong21@mails.ucas.edu.cn

摘 要: 表格有序分类旨在预测具有等级关系的标签,如何有效利用数据中的有序信息是该领域的关键问题.传统的表格数据方法未能充分挖掘这种信息,从而限制了表格有序分类性能.因此,该文提出了一种有序距离优化与特征模糊增强的表格有序分类方法,该方法包含两个核心模块:有序距离优化模块通过计算类间距离,并结合有序关系的非等距性和包容性来构造有序距离,利用其作为权重优化特征空间,增强特征判别性;特征模糊增强模块基于特征与标签近似映射特性,引入模糊增强机制,通过高斯噪音扰动生成新特征,从而提高模型泛化能力.该工作在 4 个有序表格数据集上进行广泛实验和分析,取得了优越性能和效果,充分验证了该方法及其有序距离优化模块和特征模糊增强模块的有效性.

关 键 词: 表格数据;有序分类;有序回归;有序距离;特征优化;特征增强
中图分类号: TP391 **文献标识码:** A **文 章 编 号:** 1000-1220(2026)04-0859-09

Ordinal Distance Optimization with Feature Fuzzy Augmentation for Tabular Ordinal Classification

LUO Zhengdong, Abibulla Atawulla, ZHANG Guohao, HAN Yunfei, LIU Jinlong, WANG Yi, ZHOU Xi
(The Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, Urumqi 830011, China)
(University of Chinese Academy of Sciences, Beijing 100049, China)
(Xinjiang Laboratory of Minority Speech and Language Information Processing, Urumqi 830011, China)

Abstract: Tabular ordinal classification aims to predict labels with rank relationships, making the effective utilization of ordinal information a crucial research challenge in this field. Traditional tabular data methods fail to fully exploit such information, thereby limiting ordinal classification performance. To address this issue, this paper proposes a tabular ordinal classification method Ordinal Distance Optimization with Feature Fuzzy Augmentation for Tabular Ordinal Classification, consisting of two key modules. The Ordinal Distance Optimization module calculates inter-class distances while considering the non-isometric and containment nature of ordinal relationships to construct an ordinal distance, which is then used as a weight to optimize the feature space and enhance feature discriminability. The Feature Fuzzy Augmentation module leverages the approximate mapping between features and labels by introducing a fuzziness augmentation mechanism, generating augmented features through Gaussian noise perturbation to improve the model's generalization ability. Extensive experiments and analyses on four ordinal tabular datasets demonstrate superior performance and effectiveness, validating the proposed method and its two key modules.

Keywords: tabular data; ordinal classification; ordinal regression; ordinal distance; feature optimization; feature augmentation

0 引 言

有序表格数据作为表格数据的重要组成部分,在诸如品质评级[1]、年龄预测[2]等实际应用场景中广泛存在.其核心特征在于类别之间具有明确的等级关系,即标签大小在一定程度上反映了类别间的有序(等级)结构.然而,传统的表格数据分类与回归方法往往未能充分挖掘和利用这种有序信息,而是将其视为普通的分类任务或回归问题,从而限制了模型在有序表格数据上的性能表现[3].因此,如何有效地分析和利用数据中类别之间的有序关系,对表格数据的有序分类研究具有巨大的潜力.

在基于深度学习的数据预测领域中,高质量且具有判别性的特征表示对于提高模型性能至关重要.高质量特征表示能够有效捕捉数据的内在结构和关键信息,减少噪声和冗余,使特征空间更加紧凑和高效;而判别性特征表示则能够增强类间区分度,使不同类别在特征空间中的距离更大,从而提高模型的预测能力.特征空间优化旨在进一步提升特征表示的质量和判别性,即通过调整特征分布,使同类别特征更加紧

收稿日期:2025-02-26 收修改稿日期:2025-04-11 基金项目:新疆维吾尔自治区重点研发计划课题项目(2023B01028)资助;新疆维吾尔自治区"天山英才"项目(2022TSYCLJ0035,2023TSYCCX0046)资助;中国科学院青年创新促进会项目(2021434)资助;中国科学院"西部青年学者"项目(2022-XBQNXZ-008)资助. 作者简介:**罗正东**,男,1989 年生,博士研究生,研究方向为数据挖掘、表格有序分类;**艾比布拉·阿塔伍拉**,男,1995 年生,博士,助理研究员,研究方向为自然语言处理、数据分析;**张国昊**,男,1997 年生,博士研究生,研究方向为联邦学习、数据分析;**韩云飞**,男,1990 年生,博士,副研究员,CCF 会员,研究方向为数据处理与分析;**刘金龙**,男,1983 年生,硕士,副研究员,研究方向为大数据分析、软件测试;**王 轶**,男,1986 年生,博士,研究员,CCF 高级会员,研究方向为区块链、大数据治理;**周 喜**(通信作者),男,1978 年生,博士,研究员,研究方向为自然语言处理、大数据分析.

凑,而使不同类别特征在特征空间中更加可分,从而增强模型的判别能力. 常用的特征优化策略包括 Triplet loss[4]、L-softmax[5]、Am-softmax[6]、Center loss[7] 和隐含狄利克雷分布 LDA[8] 等. 这些方法在优化特征分布、提升类别判别性方面取得了显著成效,但主要针对传统分类任务,未能充分考虑类别之间的有序关系. 近年来有序分类方法[9]通过将标签作为类间特征权重,以优化特征空间的判别性分布,并在某些任务中取得了良好效果. 但这些方法主要应用于图像、文本等领域,对有序表格数据尚缺乏针对性的优化研究. 本文专门针对表格数据中的有序分类任务展开研究,发现该类数据的标签不仅用于划分类别,同时也隐含等级关系. 然而,这种标签本质上是一种离散且粗糙的有序关系,并不能准确反映类别间的真实等级距离,导致基于标签的特征优化策略难以充分挖掘类间的有序特性. 为此,本文通过计算类间距离,更精细地表示类间有序关系距离,将其作为类间特征优化的权重,可进一步优化特征空间的结构,使得特征分布更加符合任务需求,示意图见图 1. 图 1 左侧示例了 3 个类别,其中两类之间的距离为单位 1. 图 1 中间展示了这 3 个类别之间的不等距关系,同时计算得出的有序权重也各不相同. 最终,将图 1 左侧的类间距离与图 1 中间对应的有序权重相乘,得到图 1 右侧调整后的类间距离变化. 经过该优化过程,特征分布会更加合理,最终可提升预测模型的性能,为有序表格数据的分析与预测提供更有力的支持.

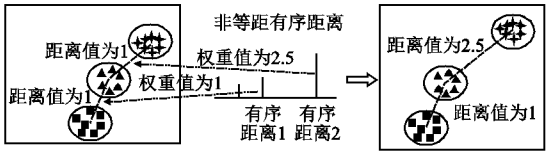

图 1 有序距离权重优化特征空间

Fig. 1 Ordinal distance weight optimization feature space

尽管特征空间优化能够有效提升特征表示的质量和判别性,但由于有序分类任务的特性,特征 x 与有序标签 y 之间的关系并非严格的数学函数映射,而是一种近似映射. 具体而言,y 只能取离散的整数值,即使 x 在一定范围发生微小变化,其对应的 y 可能仍然保持不变. 基于这一特性,阈值法[10,11]成为一种常见有序分类方法,主要包括累积链接模型[12]、支撑向量机[13]、判别式学习[14]等,这些方法通过将特征映射到一维连续变量空间,并根据预测值所在分段空间确定所属类别. 本质上,这是一种区间映射过程,即通过 x 预测目标 $\hat{y}$(整数或浮点值),然后根据 $\hat{y}$ 所在区间及阈值决定最终的离散有序类别 y. 换一种角度思考:既然阈值法是关注标签 y 的区间变化,那么是否能通过特征 x 的区间变化体现近似映射? 基于特征 x 与标签 y 的近似映射特性,当 x 在一个区间微小变化时,仍然会映射到相同 y,这可视为一种模糊映射过程. 与阈值法不同,该过程是关注特征 x 在区间内的微小变化而阈值法关注标签 y 的区间划分. 根据这一观察,本文提出一种新颖的特征模糊增强策略,即通过对特征 x 添加符合其分布的高斯噪音,实现特征 x 的微小扰动,生成新的增强特征 x',使其仍映射到相同的有序标签 y. 从一定程度上,新特征 x'可视为对原特征 x 的扩充增强,示意图见图 2. 通过对原特征和增强特征的联合学习,可以进一步丰富特征的多样性,捕捉更多潜在的模式,从而提高模型的泛化能力. 通过这种方式,模型不仅能够更好地适应训练数据中的多种变化,还能在面对未知数据时展现出更强的鲁棒性和准确性. 这种特征增强的策略,有助于减少过拟合现象,提升模型在不同场景下的预测性能,最终带来更优的整体表现.

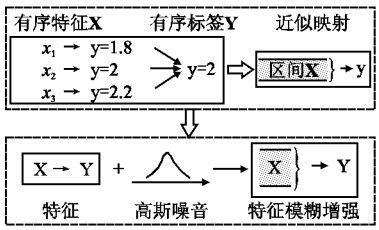

图 2 特征与标签间的近似映射关系及特征模糊增强

Fig. 2 Approximate mapping relationship between features and labels and feature fuzzy augmentation

总而言之,为进一步挖掘有序表格数据中的类间有序(等级)关系,提升特征表示的质量并增强模型的泛化能力,本文提出了一种有序距离优化与特征模糊增强的表格有序分类方法(Ordinal Distance Optimization with Feature Fuzzy Augmentation For Tabular Ordinal Classification,简称 ODFA). 该方法主要包含有序距离优化和特征模糊增强两个核心模块:有序距离优化模块通过计算类间距离进一步精确刻画类别之间的有序关系,并将其作为类间特征优化的权重,引导特征空间分布更符合任务需求,从而提升模型的判别能力. 特征模糊增强模块结合有序分类任务中特征与标签之间的近似映射特性,通过向特征添加噪声进行微小扰动,生成新的增强特征,使模型学习更加鲁棒、更加多样的特征,从而提高泛化性能. 最终,本文方法有效提升了表格有序分类的预测性能. 本文主要贡献点如下:

1)本文提出了一种有序距离优化策略,通过计算更精细的类间有序距离并将其作为权重来优化特征空间,使特征分布更加符合有序分类任务的需求,从而提升模型的判别能力;

2)本文引入了一种特征模糊增强机制,通过对特征施加微小扰动生成增强特征,增加特征表示的多样性和模型泛化能力,提高模型的稳健性;

3)本文在 4 个有序表格数据集上开展广泛实验和分析:在所有数据集上取得最优准确率 ACC 性能,在 3 个数据集上取得最优均方根误差 RMSE 性能. 进一步的分析实验表明,本文方法在提升预测性能同时具备良好的泛化能力.

1 相关工作

1.1 表格数据分类方法

表格数据广泛应用电力信息管理[15]、漏洞检测[16]和可视化[17]等领域,现有的表格数据分类方法大致可分为传统机器学习方法和深度学习方法两大类.

在表格数据领域,基于传统机器学习的方法在许多任务中取得了显著成效,尤其是在高维稀疏数据处理和特征选择方面展现了强大的能力. 其中,树模型因其卓越的特征处理能力和高效的性能表现而备受青睐. 例如,XGBoost[18] 通过梯

度提升框架和正则化技术,能够有效应对高维稀疏数据,并在各类竞赛和实际应用中表现出色;LightGBM[19]则通过基于直方图的决策树算法和叶子优先的分裂策略,进一步提升了训练效率和模型性能;CatBoost[20]则专注于处理分类型特征,通过有序目标编码和对称树结构,显著提升了模型在表格数据中的表现. 多层感知机(Multilayer Perceptron,MLP)[21]作为一种经典的神经网络模型,能够通过多层非线性变换捕捉数据中的复杂关系,尽管其对超参数较为敏感,但在某些表格数据任务中仍具有竞争力. 逻辑回归(Logistic Regression,LR)[22]作为一种简单且有效的线性模型,因其可解释性强和计算效率高,常被用于二分类和多分类任务中,尤其是在特征维度较低的场景中表现优异.

随着深度学习技术的发展,其在表格数据领域的应用逐渐增多,并成为表格数据预测方法的重要分支. 基于决策树的方法[23]通过结合深度学习与树模型的优势,能够更好地捕捉表格数据中的非线性关系和高阶交互特征. 基于 Attention 机制的方法[24,25]通过动态分配特征权重,能够更有效地捕捉关键特征及其相互关系,从而提升模型的预测性能. 迁移学习在表格数据深度学习中的应用[26]则通过将预训练模型的知识迁移到目标任务中,显著减少了对大规模标注数据的依赖,尤其在数据稀缺的场景中表现出色. 基于检索机制的表格深度学习[1]则通过检索相似样本,增强了模型对复杂表格数据的理解和泛化能力.

尽管这些方法为表格数据的分类任务提供了多样化的解决方案,但在处理有序分类问题时,仍然面临如何有效利用类别间有序关系的挑战,导致性能受限. 因此,针对有序表格数据的特殊需求,如何充分挖掘和利用类间有序关系,优化特征表示和模型结构,成为提升表格数据有序分类性能的重要课题.

1.2　有序分类方法

有序分类(亦称有序回归)是机器学习中的一个重要研究领域,旨在通过模型预测新样本的有序标签. 作为介于分类与回归之间的特殊任务,有序分类既不同于传统分类任务,也区别于传统回归任务. 与传统分类任务相比,有序分类不仅需要区分不同类别,还需充分考虑类别间的有序关系;而与回归任务不同,有序分类的标签仅表示有序等级,无法精确量化类别间的距离[27]. 因此,有序分类在许多实际应用中具有独特的挑战和研究价值.

目前,针对有序分类的研究方法主要可归纳为 3 类:1)朴素法[28],这类方法通过简化假设将有序分类问题转化为传统的分类或回归问题进行处理,虽然实现简单,但往往忽略类别间的有序信息,导致模型性能受限;2)有序二元分解法[3,29],该方法将有序目标变量(标签)分解为多个二元分类任务,通过逐级预测来捕捉类别间的有序关系;3)阈值法[10],这类方法将有序目标变量映射到一维连续空间,并通过预测值所在的区间来确定最终类别,从而更好的利用类别间的有序信息. 尽管现有方法在一定程度上解决了有序关系的挖掘与利用问题,但如何更精细地刻画类间的有序关系并优化特征空间,依然是表格有序分类中的一大挑战.

最近,专门针对表格有序分类任务的 TabCGOK 方法[30]通过检索与样本特征相似的多个特征,并结合 Attention 机制将类间有序距离与检索到的相似特征融合为一个补偿特征,最终通过将补偿特征与原始样本特征融合来提升模型性能. 与 TabCGOK 方法不同,本文提出了一种新的思路,即通过利用有序距离优化特征空间分布,从而提升模型性能. 这种特征空间优化策略进一步丰富了表格有序分类领域的研究,为提高模型的预测能力提供了新路径.

2　有序距离优化与特征模糊增强模型(ODFA)

为进一步充分利用类间有序关系以及特征与有序标签之间的近似映射特性,本文提出一种有序距离优化与特征模糊增强的模型 ODFA. 如图 3 所示,ODFA 模型框架包括两个主

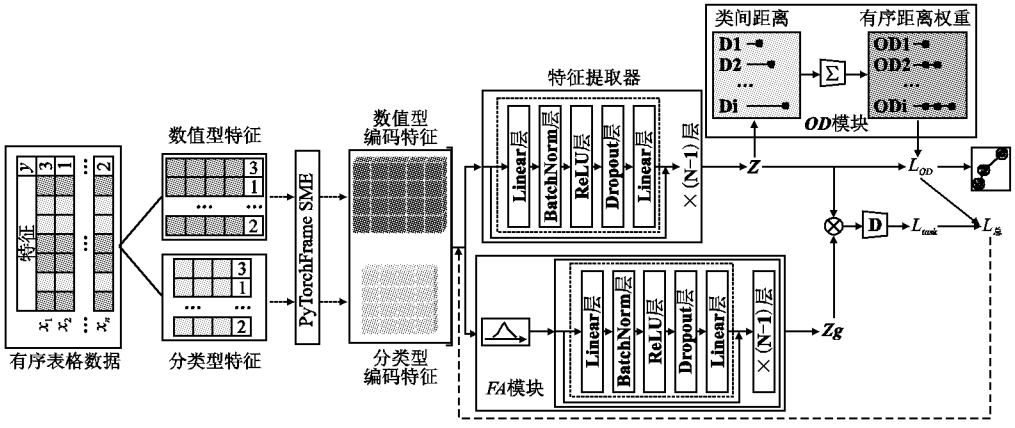

图 3　本文方法框架图

Fig. 3　Framework of our method

要模块:有序距离优化模块(OD)和特征模糊增强模块(FA). 在 ODFA 模型中,有序距离优化模块 OD 通过计算类间距离,并将其作为权重来放大类间的距离,从而优化特征空间分布;特征模糊增强模块 FA 则通过在特征上添加高斯噪音,生成新的增强特征,丰富特征的多样性,从而有效提升模型的鲁棒性和泛化能力. 图中 PyTorchFrame 表示 PyTorch-Frame[31]框架的数据处理过程中的语义预处理(Semantic type)、张量化(Materialization)、编码(Encoding)这 3 个步骤的简称. Σ 表示累积求和操作,D 表示解码器(Decoder). 特征提取器可以是各种骨干网络结构,具体选择需根据任务进行

调整. 在推理阶段,考虑到有序分类既具有分类任务的特点,又与回归任务有所关联,故采用(ACC)和均方根误差(RMSE)作为评估标准来综合评估模型的性能. 本文方法的各模块详细介绍如下文.

2.1 问题定义

表格有序分类旨在根据给定的表格数据输入 x 预测有序标签 y. 在有序表格数据集 $D=\{(X,Y)=(x_i,y_i),i=1,2,\cdots,N\}$ 中,各类标签具有有序等级关系,即标签集 $Y=\{C_1,C_2,\cdots,C_N\}$ 满足 $C_1\prec C_2\prec,\cdots,\prec C_N$,其中符号"$\prec$"表示有序关系.

2.2 有序距离优化模块 OD

本文将 PyTorchFrame 框架里数据处理过程中的语义预处理(Semantic type)、张量化(Materialization)、编码(Encoding)这 3 个步骤统称为 SME,表格数据 X 经过 SME 后得到张量形式 X_T,然后 X_T 经过特征提取器 F 后,得到特征 Z:

$$X_T=\mathrm{SME}(X) \tag{1}$$

$$Z=\mathrm{F}(X_T) \tag{2}$$

计算类间距离:

$$D_i=Z_{ci}-Z_{c1} \tag{3}$$

其中,Z_{ci}表示第 i 类的中心;N 表示类别总数;D_i 表示第 i 类中心到第一类中心的距离;$i=1,\cdots,N$.

由于有序关系具备有序性、包容性和非等距性,可知标签差值相等的各类间距离并不等,且等级较高的类包含等级较低的类的相关知识. 因此,采用特征中心类间距作为有序距离,通过对前项低等级类距离的累加来体现有序关系的包容性[30]. 类间有序距离的计算为:

$$OD_i=y_1+\sum_{i=1}^{N}D_i \tag{4}$$

其中,OD_i 表示第 i 类的有序距离;y_1 表示第一类标签值,引入 y_1 的目的是防止第一类距离为 0,避免对后续的计算产生干扰,并且引入 y_1 可增大权重,使类间扩大更明显.

利用有序距离作为类间权重,拉开类间特征距离:

$$\omega_{ij}=\|OD_i-OD_j\|_2 \tag{5}$$

$$L_{OD}=-\frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{i\neq j}\omega_{ij}\|Z_{ci}-Z_{cj}\|_2 \tag{6}$$

其中 ω_{ij}表示类别 i 和 j 之间的有序权重.

总之,针对该模块的设计目标,本文旨在通过有序距离作为权重来增大类间特征距离. 首先,计算各类特征中心之间的距离,接着结合有序关系的有序性、包容性和非等距性,通过累加前项各类之间的距离来构造有序距离. 最终,使用该有序距离作为类间特征优化的权重,调整特征空间的结构,使得不同类别之间的特征距离更具判别性,从而提升模型的分类能力.

2.3 特征模糊增强模块 FA

本文方法的另一个创新模块是特征模糊增强 FA. 正如引言中所述,本文分析特征 x 与 y 之间的近似映射关系,根据特征 x 略有变化仍会映射到相同 y 这一模糊映射过程,巧妙设计增加与原特征分布相同的高斯噪音实现特征在微小区间变化. 计算公式为:

$$X_g=X_T+\lambda\cdot\eta,\eta\sim N(\mu_{X_T},\Sigma_{X_T}) \tag{7}$$

其中,η 是从与特征 X_T 分布相同的高斯噪声;λ 是噪声的缩放因子,用于控制噪声的强度. 本文中 $\lambda=0.01$.

然后,使用与 OD 模块的特征提取器结构相同,但参数不同的 F′提取特征:

$$Z_g=\mathrm{F}'(X_g) \tag{8}$$

融合增强特征 Z'与原特征 Z,公式为:

$$Z_{fused}=\max(Z,Z_g) \tag{9}$$

其中,max(·,·)表示对两个特征张量进行点对点的最大值选择,这个操作的具体意义是,对于每个位置上的元素,保留多个张量中的较大值,从而捕捉到特征张量中更显著的信息,同时抑制噪声引起的冗余影响.

此处,特征模糊增强模块 FA 的损失函数 L_{task}为:

$$\mathrm{pred}=\mathrm{Decoder}(Z_{fused}) \tag{10}$$

$$L_{task}=\mathrm{lossFunction}(pred,Y) \tag{11}$$

在本文中,对分类准确率 ACC 评估时,则 lossFunction 为交叉熵损失函数;对回归均方根误差 RMSE 评估时,则 lossFunction 为均方误差损失函数. Y 为真实标签;Decoder 为 Linear(ReLU(LayerNorm(·)))结构.

总之,本模块通过利用特征与标签之间的近似映射关系,设计了特征模糊增强机制. 具体而言,模型通过向原始特征添加与其分布相同的高斯噪音,生成微小扰动后的增强特征. 然后将其与原特征进行融合,从而丰富特征表示的多样性. 这一过程不仅能够有效提升特征的鲁棒性,避免对噪声的过度敏感,还能增强模型的泛化能力,使其在面对未见过的数据时表现更加稳健和可靠.

2.4 目标优化

本文所设计的模型旨在同时实现分类和回归任务的准确预测,并通过有序距离优化特征空间,以增加类间特征的区分度. 因此,模型的总体优化目标 $L_{总}$ 可定义为:

$$L_{总}=\alpha\cdot L_{task}+(1-\alpha)\cdot L_{\mathrm{OD}} \tag{12}$$

其中,L_{task}是任务损失函数(若评估 ACC 则为交叉熵损失,若评估 RMSE 则为均方误差损失),用于确保模型在分类或回归任务中的预测准确性;L_{OD}是有序距离优化特征空间的损失函数,用于优化类间特征的分布,提升类间判别能力;α 是超参数,用于平衡任务损失和有序距离优化损失的权重. 在训练过程中,通过调整超参数 α,模型可以在保证任务准确率和均方误差的同时进一步优化特征空间结构,从而使得类别间的区分度更加明显,提高模型在有序分类任务中的性能表现.

3 实验与分析

3.1 实验数据集

为验证所提方法的效果,本文选用公开有序表格数据集 Wine_Quality[1]、Abalone[2]、Cmc[3] 和 Eucalyptus[4] 作为基准数据

[1] https://openml.org/d/287

[2] https://archive.ics.uci.edu/dataset/1/abalone

[3] https://www.openml.org/d/23

[4] https://www.openml.org/d/188

集. 按照 PyTorchFrame[31] 数据集划分方法,本文将各数据集

表 1　有序表格数据简要统计

Table 1　Summary statistics of ordinal tabular datasets

数据集	样本数	特征维度	类别数	描述
Wine_Quality	6,497	11	7	红酒品质预测
Abalone	4,177	7	8	动物年龄预测
Cmc	1,473	9	3	避孕效果预测
Eucalyptus	736	19	5	效能等级预测

划分为训练集、验证集和测试集,划分比例为 7∶1∶2. 各数据集的简介见表 1 所示.

3.2　实验设置

本文实验环境为 Python 3.9、PyTorch 2.3,硬件为 NVIDIA A100 GPU. 为消除因随机性带来的实验偏差,采用 15 轮随机种子进行实验,seed = range(15). 噪声因子参数 λ = 0.01,批次大小 Batch size = 256,学习率 lr = 0.001,最大训练轮次 epochs = 500,早停机制为 50 次. 由于有序分类任务具有分类与回归之间的中间性质,因此在实验中同时使用准确率(Accuracy,ACC)和均方根误差(Root Mean Square Error,RMSE)作为评估指标. 具体实验中,公式中为 Revisiting_ResNet[30] 的特征提取模块,作为本文方法中的骨干网络模块.

3.3　模型学习过程

为了更清晰地展示本文方法的训练过程,本节通过伪代码描述了模型的关键学习过程,以便能够直观地理解两个关键模块的操作原理及其作用.

模型训练主要过程:

算法 1. 有序距离优化与特征模糊增强模型优化

输入:模型 $M \leftarrow (F, F', \text{Decoder}; \theta)$;数据样本 D　/* F 和 F' 均为 *Revisiting_ResNet* 结构,θ 为模型参数 */

超参:噪声缩放因子 λ;epoch 次数 T;学习率 lr

输出:更新参数 θ 后的模型 $\acute{M}$

for $t \leftarrow 1$ *to* T **do**
  for (X, Y) in *Batch*(D) **do**　/* 选择 *Batch* 数据 */
    $X_T = SME(X)$　/* 编码特征 */
    $Z \leftarrow F(X_T)$;　/* 模型提取特征 */
    $D_i \leftarrow Z_{ci} - Z_{c1}$　/* 第 i 类中心到第 1 类中心的距离 */
    $OD_i \leftarrow y_1 + \sum_{i=1}^{N} D_i$　/* 计算第 i 类有序距离 */
    $\omega_{ij} \leftarrow \| OD_i - OD_j \|_2$　/* 类别距离作为类间权重 */
    /* 计算有序距离优化损失 */
    $L_{OD}(\theta) = -\frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{i \neq j} \omega_{ij} \| Z_{ci} - Z_{cj} \|_2$
    $X \leftarrow X + \text{random\_like}(X) * \lambda$　/* 添加噪声 */
    $Z_g \leftarrow F'(X_g)$; $Z_{fused} = max(Z, Z_g)$　/* 特征增强 */
    $p \leftarrow \text{Decoder}(Z_{fused})$
    $L_{FA}(\theta) = \text{Cross} - \text{Entropy}(p, Y)$　/* 计算分类损失 */
    $L_{总}(\theta) \leftarrow (1-\alpha) \cdot L_{OD}(\theta) + \alpha \cdot L_{FA}(\theta)$
    $\acute{\theta} \leftarrow \theta - lr \cdot \nabla L_{总}(\theta)$　/* 模型参数更新 */
  end
end

return $\acute{M} = (SME, F, F', \text{Decoder}; \acute{\theta})$

3.4　实验结果与分析

本文方法与 3 类主流基线方法进行了比较:1)基于树模型的表格学习方法 XGBoost[18]、CatBoost[19]、LightGBM[20];2)基于检索机制的表格深度学习方法 SAINT[32]、TabR[1]、TabCGOK[30];3)PyTorchFrame[31] 框架中的表格深度学习方法 TabTransformer[25]、TabNet[24]、FTtransformer[33]、ExcelFormer[34]、Trompt[35]、MambaTab[36]、Revisiting_ResNet[33]. 准确率(ACC)和均方根误差(RMSE)的实验结果分别见表 2 和表 3,表中下划线表示在 PyTorchFrame 框架中性能最优值,黑色加粗性能表示在 3 类方法中最优值. 此外,通过计算平均差异效果,进一步清晰展示了本文方法与各基线方法的整体性能差异,定义为:平均差异 =(本文方法性能-基线方法性能)/数据集个数.

表 2　有序表数据集上准确率 ACC(%)性能比较(↑值越大性能越好)

Table 2　Performance comparison of accuracy ACC (%) on ordinal tabular dataset (↑ larger value better performance)

基线框架	基线方法	Wine_Quality	Abalone	Cmc	Eucalyptus	平均差异
树模型	XGBoost	65.52	32.94	56.76	66.22	15.61
	CatBoost	66.08	32.06	56.53	72.97	14.06
	LightGBM	64.74	33.33	56.08	68.47	15.31
检索机制	SAINT	63.44	34.13	56.31	72.07	14.48
	TabR	66.23	33.49	56.31	72.97	13.72
	TabCGOK	66.91	34.37	57.21	73.42	12.99
PyTorchFrame	TabTransformer	59.64	34.42	56.50	61.00	18.08
	TabNet	81.92	55.84	58.35	64.52	5.81
	FTtransformer	75.30	38.11	50.50	45.90	18.51
	ExcelFormer	56.20	35.35	53.38	57.06	20.47
	Trompt	63.49	39.77	59.32	70.69	12.65
	MambaTab	59.80	37.18	58.41	63.34	16.28
	Revisiting_ResNet(backbone)	82.89	54.56	62.65	75.16	2.15
	ODFA(ours)	**84.18**	**56.27**	**64.45**	**78.96**	–

准确率(ACC)性能分析:从实验结果表 2 可知,本文方法在 4 个有序表格数据集上的准确率超越了 PyTorchFrame 框架中的各基线方法(包括作为骨干模型的 Revisiting_ResNet[33]),平均性能高出 2.15 ~ 20.47 个百分点,这表明,在相同数据预处理及编码条件下,本文方法展现出明显的性能优势. 与基于检索机制的基线方法相比,本文方法平均性能高出 12.99 ~ 14.48 个百分点,展现出本文方法的较好性能. 与基于树模型的基线方法相比,本文方法平均性能高出 14.06 ~

15.61个百分点,进一步证明了本文方法在性能上超越了传统的树模型方法. 总体而言,本文方法在公开基准数据集上达到了最优 SOTA 的 ACC 性能,充分证明了本文方法的有效性和优越性.

表 3 有序表数据集上均方根误差 RMSE 性能比较(↓值越小性能越好)

Table 3 Comparison of root mean square error RMSE performance on ordinal tabular dataset (↓ smaller value better performance)

基线框架	基线方法	Wine_Quality	Abalone	Cmc	Eucalyptus	平均差异
树模型	XGBoost	0.602	1.418	0.706	0.717	-0.075
	CatBoost	0.606	1.424	0.707	0.684	-0.070
	LightGBM	0.612	1.400	**0.701**	0.741	-0.078
检索机制	SAINT	0.676	1.359	0.717	0.731	-0.085
	TabR	0.620	1.335	0.716	0.722	-0.063
	TabCGOK	0.611	1.321	0.715	0.703	-0.052
PyTorchFrame	TabTransformer	0.948	1.707	0.917	1.333	-0.441
	TabNet	0.645	1.342	0.827	1.998	-0.417
	FTtransformer	0.660	1.404	0.846	1.040	-0.202
	ExcelFormer	0.684	1.363	0.801	0.622	-0.082
	Trompt	0.664	1.323	0.770	0.585	-0.050
	MambaTab	0.968	1.486	0.941	1.202	-0.364
	Revisiting_ResNet(backbone)	0.610	1.286	0.767	0.537	-0.014
	ODFA(ours)	**0.587**	**1.283**	0.763	**0.510**	-

均方根误差(RMSE)性能分析:如前文所述,有序分类是介于分类与回归之间的任务,因此本文同时评估了分类指标准确率和回归指标均方根误差. 从实验结果表 3 可知,与 PyTorchFrame[31] 框架中的各基线方法相比,本文方法的 RMSE 性能优于各基线方法(包括作为骨干模型的 Revisiting_ResNet[33]),平均性能低出 0.014 ~ 0.441,由于 RMSE 值越低性能越好,这表明在相同的数据预处理及编码时,本文方法实现了较为优越的回归性能. 与基于树模型和基于检索机制的基线方法相比,本文方法 3 个数据集上表现优异,仅在 Cmc 数据集上的表现稍逊. 然而值得注意的是,本文方法在所有数据集上的表现均优于骨干模型 Revisiting_ResNet[33],这进一步验证了本文方法中两个改进模块的有效性. 总体而言,本文方法在 3 个数据集上达到了最优 SOTA 的 RMSE 性能,并且在所有选定数据集上均优于骨干模型,这有力地证明了本文方法的有效性.

3.5 消融实验

本文开发的有序距离优化模块(OD)和特征模糊增强模块(FA)旨在增强模型对有序表格数据的预测能力. 为了评估这两个模块的影响,本文在 Abalone 和 Eucalyptus 数据集上进行了消融实验. 根据实验结果表 4 可知,实验(2)性能优于实验(1)性能,表明有序距离优化特征空间能够有效改善模型的预测性能;实验(3)性能优于实验(1),证明了利用特征 x 与标签 y 之间的近似映射关系,通过添加噪音实现特征模糊增强,可以丰富特征多样性,从而提升模型鲁棒性和泛化能力,最终提高模型性能;实验(4)与实验(2)、(3)的对比,说明在骨干模型上同时引入有序距离优化模块(OD)和特征模糊增强模块(FA)能够获得最大增益;此外,实验(4)与实验(1)的对比也充分证明了本文方法的有效性. 综上所述,消融实验结果验证了各模块的有效性,并且表明同时集成这两模块可获得最大性性能改善.

表 4 ODFA 消融实验

Table 4 Ablation study for ODFA

实验	消融模块			Abalone		Eucalyptus	
	骨干	OD	FA	ACC↑	RMSE↓	ACC↑	RMSE↓
(1)	√			54.56%	1.286	75.16%	0.537
(2)	√	√		54.57%	1.285	75.65%	0.526
(3)	√		√	56.20%	1.285	78.60%	0.527
(4)	√	√	√	**56.27%**	**1.282**	**78.96%**	**0.510**

3.6 模块即插即用效果分析

本文开发的有序距离优化模块和特征模糊增强模块主要对特征进行操作,而不依赖具体的特征提取器(骨干模型)结构. 为验证这两个模块的即插即用效果,本实验选取 FTtransformer[33] 和 MambaTab[36] 两种基线方法,在 Abalone、Cmc、Eucalyptus 3 个数据集上进行 RMSE 性能评估. 表 5 实验结果表明,在引入本文提出的改进模块后,两种基线方法的整体性能均有所提升,仅在 Abalone 数据集上表现欠佳. 这些实验结果充分证实了本文所提的改进模块具备良好的普适性和即插即用性能,并能够有效地提升不同模型的预测能力.

表 5 模块的即插即用普适性实验(RMSE 性能↓)

Table 5 Plug-and-play universality experiments on modules (RMSE performance↓)

	Abalone	Cmc	Eucalyptus
FTtransformer	**1.404**	0.846	1.040
FTtransformer + ODFA 模块	1.408	**0.844**	**0.991**
MambaTab	1.486	0.941	1.202
MambaTab + ODFA 模块	**1.3288**	**0.763**	**0.675**

3.7 有序距离优化效果可视化

有序距离优化模块是本文方法的核心创新之一,旨在通过有序距离优化类间特征空间分布,提升模型的判别能力. 为了直观展示该模块的效果,本实验选择在 Eucalyptus 数据集上分别对基础骨干模型(backbone)和结合有序距离优化的模型(backbone + OD)进行特征分布可视化分析. 具体地,本实验选择解码器前一层的特征表示作为分析对象. 图 4 展示了两种方案的可视化结果. 从图中可观察到,在仅使用基础骨干

模型时(左图),不同类别的特征分布存在较大重叠,类间边界较为模糊. 而在引入有序距离优化模块后(右图),各类别的特征分布边界更为清晰,类间的可分性显著增强. 这一结果

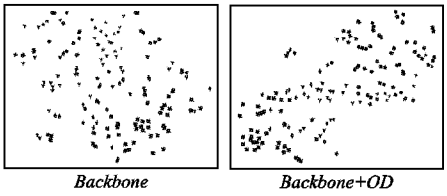

图 4　有序距离优化模块的特征空间优化效果可视化

Fig. 4　Visualization of the feature space optimization effect of the ordinal distance optimization module

表明,有序距离优化模块能够有效优化类间特征分布,减少类别混淆,从而提升表模型的整体性能.

3.8　模型稳定性分析

为了消除随机性带来的偏差,本文在实验中采用了 15 轮随机种子. 表 2 和表 3 展示了 15 次实验结果的均值,而图 5 则呈现了这 15 次实验结果的标准差. 从图 5 中可以看出,在

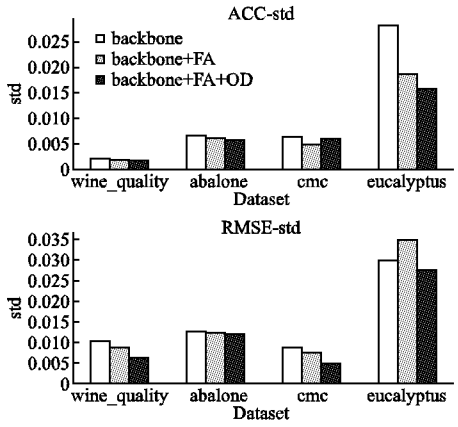

图 5　改进模块对模型 ACC 和 RMSE 性能稳定性影响

Fig. 5　Impact of improvement module on model ACC and RMSE performance stability

对比骨干模型(backbone)、骨干模型 + 特征增强模块(backbone + FA)和骨干模型 + 特征增强模块 + 有序距离优化模块(backbone + FA + OD)3 种模型时,8 组条形图中的 6 组表现出标准差(std)逐渐降低的趋势. 仅在 Cmc 数据集的 ACC-std 和 Eucalyptus 数据集的 RMSE-std 上未出现如此趋势,但第 3 种方案(backbone + FA + OD)的标准差仍低于第一种方案(backbone). 这些现象表明,特征增强模块中的高斯噪声操作有助于提升模型性能稳定性,增强模型的鲁棒性和泛化能力. 此外,高质量且具有判别性的特征进一步促进了模型的稳定性. 通过这一分析,验证了本文提出的改进模块在提升模型泛化能力方面的有效性.

3.9　单双支骨干网络对比实验

本文通过增加高斯噪音以增强特征,为探讨在特征提取过程中采用共享权重还是分别独立权重的设计,本文进行了单双支骨干网络对比实验. 单支表示原特征 x 和加噪音后的特征 x_g 均通过骨干网络 Revisiting_ResNet[33] 提取特征,共享参数;双支表示原特征 x 和加噪音后的特征 x_g 分别通过两个结构相同但不共享参数的骨干网络 Revisiting_ResNet[33] 提取特征. 表 6 中实验结果表明,双支骨干网络在选取的 Wine_Quality、Cmc、Eucalyptus 3 个数据集上的准确率表现优于单

表 6　单双支骨干网络的准确性 ACC(%)性能比较

Table 6　Comparison of accuracy ACC (%) performance of single and double branch backbone

	Wine_Quality	Abalone	Cmc	Eucalyptus
单支	83.62	**56.99**	63.22	75.97
双支	**84.18**	56.27	**64.45**	**78.96**

支骨干网络. 原因分析认为,单支骨干网络共享特征提取的参数,可能由于不同输入特征的差异导致梯度更新出现冲突,从而影响优化过程. 而双支骨干网络通过分开处理原特征和加噪音的特征,有效避免了这一问题,从而帮助模型收敛到更优的解,提升了模型预测性能.

3.10　时间开销对比

本文在实验数据集上,对骨干模型与本文提出的 ODFA 方法的时间开销进行了对比分析,具体结果见表 7. 从 ACC 和 RMSE 指标的训练、验证及测试的总时间来看,ODFA 模型的时间开销约为骨干基线模型的两倍. 这一差异主要源于 ODFA 模型采用了双支子网络结构,而骨干基线模型则为单支子网络结构. ODFA 在训练过程中需要同时计算类内特征表示与类间有序关系,通过额外的计算模块优化有序信息建模,因此相比于传统的单分支网络,计算量显著增加. 此外,ODFA 还涉及有有序距离的特征空间优化、特征模糊增强等,进一步增加了计算复杂度,从而导致更高的时间开销. 尽管 ODFA 方法的计算成本有所上升,但其在分类精度和有序信息利用能力方面表现更优,能够更精准地捕捉表格数据的有序特性. 因此,在实际应用中,可以根据计算资源和任务需求,在模型性能与计算效率之间进行权衡.

表 7　本文 ODFA 方法与骨干模型时间开销比较

Table 7　Comparison of ODFA method and backbone model time overhead

	ACC 时间开销对比(秒)				RMSE 时间开销对比(秒)			
	Wine_Quality	Abalone	Cmc	Eucalyptus	Wine_Quality	Abalone	Cmc	Eucalyptus
骨干模型	76.75	66.37	40.8	30.92	109.19	70.81	45.02	32.34
ODFA(ours)	156.64	132.31	84.25	79.39	125.69	152.24	74.37	60.29

3.11　泛化局限性

Microsoft[5] 数据集是一个用于查询结果相关性预测的大型表格数据集,包含 1,200,192 个样本,特征维度为 136 维,类别共有 5 个等级,呈现明显的有序性. 为进一步验证本文方

[5] https://www.microsoft.com/en-us/research/project/mslr/

法在大规模、高维表格数据上的适用性,本文在 Microsoft[5] 数据集上对多个基线模型和本文方法进行了准确率(ACC)对比实验,结果如图 6 所示. 实验结果表明,本文方法 ODFA 在多个基线模型的比较中取得了次优性能,表现仅次于骨干模型 Revisiting_ResNet,且二者的准确率几乎相当. 这一现象可能表明,当数据集规模足够大时,各类别的样本分布已较为充分,类别间的特征模式已被模型较好地捕获. 在此情况下,本文方法所采用的高斯噪声扰动增强策略对数据多样性的提升作用变得有限,可能导致特征模糊增强模块的贡献减弱. 这反

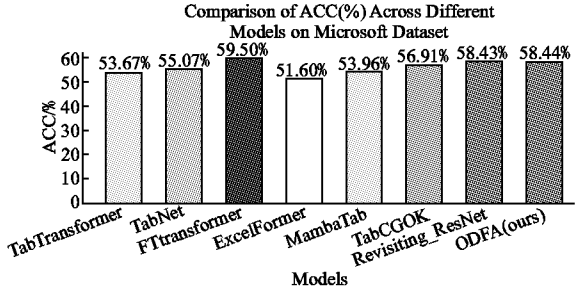

图 6 各方法在 Microsoft 数据集上 ACC(%)性能比较

Fig. 6 Comparison of ACC(%) performance of methods on Microsoft dataset

映出,当前方法在极大规模数据集上的增益效果可能受到一定限制,尤其是在数据已经较为完备的情况下,其增强策略的有效性值得进一步研究和优化. 未来的工作可以探索更适应大规模数据场景的增强策略,例如结合自监督学习或对抗训练方法,以更有效地提升模型对有序关系的建模能力.

4 总 结

为进一步利用表格有序关系提升预测性能,本文提出了一种有序距离优化与特征模糊增强的表格有序分类方法. 该方法主要包含有序距离优化模块和特征模糊增强模块两个核心模块:有序距离优化模块基于有序关系的非等距性和包含性,由类间距离构造有序距离,并将其作为类间特征分布优化权重;而特征模糊增强模块则基于有序关系种特征与标签的近似映射特性,通过高斯噪音扰动生成新特征,以实现特征模糊增强. 本文方法在 4 个公开有序表格数据上取得较为优越的性能,充分验证了本文方法有效性. 此外,通过消融实验、模型稳定性分析、单双支骨干网络对比实验等实验分析,证明了各模块对模型性能和泛化性的促进作用,同时也说明了本文方法策略的合理性. 然而,本文方法在时间开销和大规模数据集上的表现尚未达到预期,未来的研究将进一步探索更多的表格有序分类方法,深入分析特征与标签联合的有序关系特性,以便设计能够生成更高质量、更具判别性的特征表示的模型. 此外,未来还将考虑优化计算效率,并探讨如何更好地适应大规模有序表格数据的处理需求. 期望能够在提高表格有序分类任务的性能和泛化性的同时,减少模型在处理大规模数据时的时间开销,从而推动该领域的发展.

References:

[1] Gorishniy Y, Rubachev I, Kartashev N, et al. Tabr: unlocking the power of retrieval-augmented tabular deep learning[J]. arXiv preprint arXiv. 2307. 14338, 2023.

[2] Cardoso J S, Cruz R P M, Albuquerque T. Unimodal distributions for ordinal regression[J]. IEEE Transactions on Artificial Intelligence, 2025, doi: 10. 1109/TAI. 2025. 3549740.

[3] Cao W, Mirjalili V, Raschka S. Rank consistent ordinal regression for neural networks with application to age estimation[J]. Pattern Recognition Letters, 2020, 140: 325-331, doi: 10. 1016/j. patrec. 2020. 11. 008.

[4] Guo K, Lovell B C. Domain-aware triplet loss in domain generalization[J]. Computer Vision and Image Understanding, 2024, 243: 103979, doi: 10. 1016/j. cviu. 2024. 103979.

[5] Xu J, Liu X, Zhang X, et al. X2-softmax: margin adaptive loss function for face recognition[J]. Expert Systems with Applications, 2024, 249: 123791, doi: 10. 1016/j. eswa. 2024. 123791.

[6] Wang F, Cheng J, Liu W, et al. Additive margin softmax for face verification[J]. IEEE Signal Processing Letters, 2018, 25(7): 926-930.

[7] Wen Y, Zhang K, Li Z, et al. A discriminative feature learning approach for deep face recognition[C]//European Conference on Computer Vision, 2016: 499-515.

[8] Jelodar H, Wang Y, Yuan C, et al. Latent dirichlet allocation (LDA) and topic modeling: models, applications, a survey[J]. Multimedia Tools and Applications, 2019, 78(11): 15169-15211.

[9] Zhang S, Yang L, Mi M B, et al. Improving deep regression with ordinal entropy[J]. arXiv preprint arXiv: 2301. 08915, 2023.

[10] Fuchs T S, Keshet J. Thor: threshold-based ranking loss for ordinal regression[J]. arXiv preprint arXiv: 2205. 04864, 2022

[11] Gutiérrez P A, PerezOrtiz M, Sanchez Monedero J, et al. Ordinal regression methods: survey and experimental study[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 28(1): 127-146.

[12] Taylor J E, Rousselet G A, Scheepers C, et al. Rating norms should be calculated from cumulative link mixed effects models[J]. Behavior Research Methods, 2023, 55(5): 2175-2196.

[13] Zhu F, Chen X, Chen S, et al. Relative margin induced support vector ordinal regression[J]. Expert Systems with Applications, 2023, 231: 120766, doi: 10. 1016/j. eswa. 2023. 120766.

[14] Huhn J C, Hullermeier E. Is an ordinal class structure useful in classifier learning? [J]. International Journal of Data Mining, Modelling and Management, 2008, 1(1): 45-67.

[15] JIN L H, CHEN S B, ZHANG X Q, et al. Research on key technology of power transportation and distribution management system [J]. Computer Engineering, 2009, 35(5): 257-258.

[16] LI P C, ZHANG Q T, HU Y. Smart contract vulnerability detection method based on graph convolutional network with dual attention mechanism[J]. Netinfo Security, 2024, 24(11): 1624-1631.

[17] LI W Q, XIE Z P. A study of visually parallel relationships in web tables based on graph models[J]. Journal of Chinese Computer Systems, 2014, 35(7): 1567-1572.

[18] Chen T, Guestrin C. Xgboost: a scalable tree boosting system [C]//Proceedings of the 22nd ACMSIGKDD International Conference on Knowledge Discovery and Data Mining, 2016: 785-794.

[19] Ke G, Meng Q, Finley T, et al. Lightgbm: a highly efficient gradient boosting decision tree[C]//31st Annual Conference on Neural Information Processing Systems(NIPS), 2017, 3146-3154.

[20] Prokhorenkova L, Gusev G, Vorobev A, et al. CatBoost: unbiased boosting with categorical features[C]//32nd Conference on Neural Information Processing Systems(NIPS),2018:6638-6648.

[21] Klambauer G, Unterthiner T, Mayr A, et al. Selfnormalizing neural networks[C]//31st Annual Conference on Neural Information Processing Systems(NIPS),2017:972-981.

[22] Bender R, Grouven U. Using binary logistic regression models for ordinal data with non-proportional odds[J]. Journal of Clinical Epidemiology,1998,51(10):809-816.

[23] Ke Guolin, Xu Zhenhui, Zhang Jia, et al. Deepgbm: a deep learning framework distilled by gbdt for online prediction tasks[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining,2019:384-394.

[24] Arik S Ö, Pfister T. Tabnet: attentive interpretable tabular learning [C]//Proceedings of the AAAI Conference on Artificial Intelligence,2021:6679-6687.

[25] Huang X, Khetan A, Cvitkovic M, et al. Tabtransformer: tabular data modeling using contextual embeddings[J]. arXiv preprint arXiv:2012.06678,2020.

[26] Hollmann N, Müller S, Eggensperger K, et al. Tabpfn: a transformer that solves small tabular classification problems in a second[J]. arXiv:2207.01848,2022.

[27] Shi X, Cao W, Raschka S. Deep neural networks for rank-consistent ordinal regression based on conditional probabilities[J]. Pattern Analysis and Applications,2023,26(3):941-955.

[28] Sánchez Monedero J, Gutiérrez P A, Tiňo P, et al. Exploitation of pairwise class distances for ordinal classification[J]. Neural Computation,2013,25(9):2450-2485.

[29] Kim K J, Ahn H. A corporate credit rating model using multi-class support vector machines with an ordinal pairwise partitioning approach[J]. Computers & Operations Research, 2012, 39(8): 1800-1811.

[30] Luo Z, Atawulla A, Yang F, et al. TabCGOK: intra-class groups retrieval and inter-class ordinal knowledge augmented network for ordinal tabular data prediction[C]//European Conference on Artificial Intelligence(ECAI),2024:2242-2249.

[31] Hu W, Yuan Y, Zhang Z, et al. PyTorch frame: amodular framework for multi-modal tabular learning[J]. arXiv preprint arXiv: 2404.00776,2024.

[32] Somepalli G, Goldblum M, Schwarzschild A, et al. Saint: improved neural networks for tabular data via row attention and contrastive pre-training[J]. arXiv preprint arXiv:2106.01342,2021.

[33] Gorishniy Y, Rubachev I, Khrulkov V, et al. Revisiting deep learning models for tabular data[C]//35th Annual Conference on Neural Information Processing Systems(NeurIPS),2021:18932-18943.

[34] Chen J, Yan J, Chen Q, et al. Excelformer: a neural network surpassing gbdts on tabular data[J]. arXiv preprint arXiv: 2301.02819,2023.

[35] Chen K Y, Chiang P H, Chou H R, et al. Trompt: towards a better deep neural network for tabular data[J]. arXiv preprint arXiv: 2305.18446,2023.

[36] Ahamed M A, Cheng Q. MambaTab: a plug-and-play model for learning tabular data[C]//IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR),2024: 369-375.

附中文参考文献:

[15] 金丽华,陈圣波,张旭晴,等. 电力输配电管理系统关键技术研究[J]. 计算机工程,2009,35(5):257-258.

[16] 李鹏超,张全涛,胡　源. 基于双注意力机制图神经网络的智能合约漏洞检测方法[J]. 信息网络安全,2024,24(11):1624-1631.

[17] 李雯琴,谢志鹏. 基于图模型的 Web 表格中视觉并列关系的研究[J]. 小型微型计算机系统,2014,35(7):1567-1572.