



基于 Transformer 的多维数据分类模型研究与应用

黄欣欣, 陈日城, 潘自坚, 郑德武, 吴启晗, 卜子洋*

(阳光学院人工智能学院, 福建 福州 350015)

摘要: 针对多维数据分类任务中变量维度多、变量间相关关系复杂、传统模型过度依赖人工特征工程等问题, 本文提出一种基于 Transformer 的跨维注意力分类模型 (Cross-Dimension Transformer, CD-Transformer)。该模型将表格型多维数据的每个变量维度视为一个特征 token, 通过标量映射层、维度标识嵌入层和 Transformer 编码器学习变量之间的非线性交互关系, 并采用均值池化与最大池化联合方式形成全局样本表征。为验证模型有效性, 本文选取 Wine 和 BreastCancer 两个公开多维分类数据集进行重复实验, 并与逻辑回归、支持向量机、随机森林和多层感知机等典型方法进行比较。实验结果表明, CD-Transformer 在两个数据集上的平均准确率分别达到 0.9444 和 0.9610, 整体优于普通 MLP 模型, 能够在保持较强非线性表达能力的同时提供一定的特征交互解释能力。

关键词: Transformer; 多维数据分类; 跨维注意力; 特征嵌入; 深度学习

收稿日期: 2026 年 5 月 10 日

中图分类号: TP183

通讯作者: * 卜子洋, 阳光学院人工智能学院

Research and Application of Multi-Dimensional Data Classification Model Based on Transformer

Xinxin Huang, Richeng Chen, Zijian Pan, Dewu Zheng, Qihan Wu, Ziyang Bu

(School of Artificial Intelligence, Yango University, Fuzhou, Fujian 350015)

Abstract: In the context of multi-dimensional data classification tasks, where there are numerous variable dimensions and complex correlations among variables, as well as the excessive reliance on manual feature engineering by traditional models, this paper proposes a cross-dimensional attention classification model based on Transformer (Cross-Dimension Transformer, CD-Transformer). This model treats each variable dimension of tabular multi-dimensional data as a feature token. Through scalar mapping layers, dimension identifier embedding layers, and Transformer encoders, it learns the nonlinear interaction relationships among variables, and uses a combined method of mean pooling and maximum pooling to form global sample representations. To verify the model's effectiveness, this paper conducts repeated experiments using two public multi-dimensional classification datasets, Wine and BreastCancer, and compares it with typical methods such as logistic regression, support vector machine, random forest, and multi-layer perceptron. The experimental results show that the average accuracy of CD-Transformer on the two datasets reaches 0.9444 and 0.9610 respectively, and is overall superior to the ordinary MLP model. It can maintain a strong nonlinear expression ability while providing certain feature interaction interpretation capabilities.

Key words: Transformer; Multidimensional data classification; Cross-dimensional attention; Feature embedding; Deep learning



1 引言

随着传感器采集、工业互联网和金融风控平台的快速发展,数据对象不再表现为单一变量或低维属性,而是呈现出维度多、结构复杂、来源异构和样本分布不均衡等特点。多维数据分类是从多组特征中学习类别判别规则的核心任务,在产品质量识别、异常风险检测 and 用户行为分析等场景中具有广泛应用价值。与传统低维分类问题相比,多维数据分类不仅需要判断单个变量与类别之间的关系,还需要捕捉变量之间的耦合关系、冗余关系和潜在交互结构。因此,分类模型的特征表达能力直接影响最终分类性能。

传统机器学习方法如逻辑回归、支持向量机、随机森林和梯度提升树在中小规模表格数据中长期表现稳定。其优势在于训练效率高、参数规模小、解释方式相对清晰。但在复杂非线性关系明显、特征维度较高或变量关系存在潜在层级结构时,传统模型往往需要依赖人工特征组合和领域经验。深度学习模型能够自动学习特征表示,但常见的多层感知机在输入层通常将所有变量直接拼接,容易忽略变量维度之间的显式关系;卷积神经网络更适合具有局部空间结构的数据;循环神经网络则主要面向具有明确时间顺序的序列数据。因此,如何构建适用于一般多维数据分类任务的统一表示模型,仍具有研究价值。

近年来,时间序列分类研究表明,深度学习模型已从早期卷积结构和循环结构逐渐扩展至注意力机制结构,模型对长距离依赖和全局交互的刻画能力显著增强^[1]。Transformer 的核心优势在于自注意力机制能够在不依赖固定局部窗口的情况下计算任意两个 token 之间的关联权重。该机制最初主要用于自然语言处理任务,但其“token 化输入—全局关系建模—上下文表示学习”的思想可以迁移到多维数据分类中。若将表格数据中的每一个变量维度视为一个 token,则 Transformer 可以用于学习变量维度之间的交互关系,从而为多维数据分类提供一种新的建模思路。

基于上述分析,本文提出一种 CD-Transformer 模型。该模型的核心思想是:对每个数值特征进行标量映射,使不同量纲特征进入统一向量空间;其次引入维度标识嵌入,用于保留变量身份信息;

然后通过 Transformer 编码器学习跨维度特征关系;最后使用均值池化与最大池化联合策略形成样本级表示,并输入分类层输出类别概率。

2 相关研究

多维数据分类可以视为传统分类任务在高维特征空间中的扩展。早期研究多依赖基于距离、核函数或树模型的方法,通过构造判别边界完成类别划分。对于具有明显序列属性的数据,研究者进一步引入多元时间序列分类框架,从变量关系和时间依赖两个维度刻画数据结构。张旭等提出基于不确定性的多元时间序列分类算法,通过图神经网络和变分贝叶斯思想增强模型对变量关系及分类不确定性的表达能力^[2]。该类研究说明,多维分类任务的关键不只是提高分类器容量,还在于合理表达变量之间的结构关系。

在模型鲁棒性方面,时间序列分类模型容易受到噪声扰动、异常采样和对抗样本影响。王璐瑶等针对时间序列分类模型设计了集成对抗训练防御方法,说明分类模型在追求准确率之外,还需要关注稳定性和泛化性^[3]。这对多维数据分类同样具有启发意义,因为实际应用中的多维数据往往存在测量误差、缺失值、异常点和特征漂移。若模型仅在训练集上获得较高准确率,而缺乏对特征扰动的适应能力,其工程应用价值会受到限制。

除深度学习模型外,基于 Shapelets 的时间序列分类方法也具有代表性。王威娜和李明莉提出基于 Shapelets 的多元时间序列分类方法,通过发现具有判别意义的局部子序列提升分类能力^[4]。这类方法强调局部判别模式的提取,而 Transformer 更强调全局关系建模。两类方法的差异说明,多维数据分类中既存在局部关键变量,也存在跨变量组合关系。因此,在模型设计中同时考虑局部表达和全局交互具有合理性。

从预测建模角度看,深度学习已广泛应用于时间序列建模。毛远宏等对基于深度学习的时间序列预测方法进行了综述,指出卷积、循环和注意力结构在不同数据场景中各有优势^[5]。梁宏涛等进一步总结了深度学习在时序预测中的应用进展,并指出 Transformer 模型在长依赖捕捉方面具有突出潜力^[6]。虽然预测任务与分类任务目标不同,但二者在特征表示学习方面具有共同基础,即都



需要从复杂多维输入中抽取有效表征。

Transformer 相关研究也逐渐扩展至缺失补全、信号识别和多参数检测等任务。谷小兵等围绕多元时间序列缺失值插补构建了基于 Transformer 的插补方法,说明自注意力结构能够为多变量数据完整性建模提供支持^[7]。郑庆河等提出结合小波去噪卷积与稀疏 Transformer 的调制识别方法,体现了 Transformer 与领域特征提取模块融合的可行性^[8]。此外,刘辉等从空间自注意力角度研究严重缺失多元时间序列插补问题,为跨变量关系学习提供了进一步参考;庄学彬等使用并行 CNN-Transformer 神经网络进行多参数欺骗检测,也表明 Transformer 结构可以与传统特征提取模块结合用于复杂分类判别场景^[9]。综上,现有研究已证明注意力机制在多维数据建模中具有潜力,但针对一般表格型多维分类任务的跨维 Transformer 结构仍有进一步研究空间^[10]。

3 问题定义与建模思路

设多维分类数据集为 $D = \{ (x_i, y_i) \}_{i=1}^N$, 其中 $x_i \in \mathbb{R}^d$ 表示第 i 个样本的 d 维特征向量, $y_i \in \{1, 2, \dots, C\}$ 表示对应类别标签。多维数据分类任务是学习一个映射函数 $f_\theta: \mathbb{R}^d \rightarrow \mathbb{R}^C$, 使模型输出的类别概率分布 $p_i = \text{softmax}(f_\theta(x_i))$ 尽可能接近真实标签分布。与普通向量输入分类不同,本文关注的是变量维度之间的交互关系,即模型不仅要学习每个特征本身的判别作用,还要学习特征组合对类别的影响。

在传统 MLP 中,输入向量 x_i 通常一次性映射为隐藏层向量 $h = \sigma(Wx_i + b)$ 。这种方式将所有变量混合在同一线性变换中,虽然能够表达非线性关系,但难以显式解释变量间的依赖结构。本文将每个特征 x_{ij} 视为一个独立 token,通过共享标量映射函数将其转化为 $e_{ij} \in \mathbb{R}^m$ 。为避免所有标量 token 在语义上不可区分,模型进一步为第 j 个变量引入维度标识向量 $v_j \in \mathbb{R}^m$,使输入 token 表示为 $z_{ij} = e_{ij} + v_j$ 。这样,Transformer 编码器输入序列 $Z_i = [z_{i1}, z_{i2}, \dots, z_{id}]$ 既包含数值信息,也包含变量身份信息。

本文使用交叉熵损失函数作为优化目标: $L = -(1/N) \sum_{i=1}^N \sum_{c=1}^C I(y_i = c) \log p_{ic}$ 。模型训练的关键是通过最小化损失函数学习参数

θ , 包括标量映射层参数、维度嵌入、Transformer 编码器参数和分类头参数。由于输入特征来自不同量纲,实验前对所有连续变量执行 Z-score 标准化,以减小量纲差异对训练过程的影响。

4 CD-Transformer 模型设计

CD-Transformer 模型结构如图 1 所示。模型由五个主要模块组成:数据标准化模块、标量映射模块、维度标识嵌入模块、跨维 Transformer 编码模块和分类输出模块。与直接将多维向量输入 MLP 不同,该结构把变量维度展开为 token 序列,使自注意力机制能够在变量维度之间建立连接。若某些变量对类别判别存在组合效应,模型可通过注意力权重和深层表示学习对其进行捕捉。

标量映射层用于将单个连续变量映射为 m 维向量。对于第 i 个样本的第 j 维特征 x_{ij} ,有 $e_{ij} = W_s x_{ij} + b_s$, 其中 $W_s \in \mathbb{R}^{m \times 1}$, $b_s \in \mathbb{R}^m$ 。该层采用共享参数,使所有变量通过相同函数进入统一表示空间。共享标量映射可以降低模型参数量,并避免在小样本数据上为每个变量单独学习过多参数。

维度标识嵌入用于弥补共享标量映射带来的变量身份缺失问题。具体而言,为每个特征维度设置可学习向量 v_j ,并构造 $z_{ij} = e_{ij} + v_j$ 。该设计类似位置编码,但其含义不是时间位置,而是变量身份。对于表格数据而言,第 1 维和第 2 维并不存在天然顺序关系,但它们具有不同业务含义,因此维度标识嵌入比单纯位置编码更符合表格型多维数据特点。

跨维自注意力编码器是模型的核心。给定输入矩阵 $Z_i \in \mathbb{R}^{d \times m}$,通过线性映射得到 $Q = Z_i W_Q$, $K = Z_i W_K$, $V = Z_i W_V$,自注意力计算为 $\text{Attention}(Q, K, V) = \text{softmax}(QK^T / \sqrt{m})V$ 。多头注意力机制进一步将表示空间划分为多个子空间,分别学习不同类型的变量关系,然后进行拼接和线性变换。通过堆叠编码层,模型能够获得包含上下文变量关系的特征表示 $H_i \in \mathbb{R}^{d \times m}$ 。

在样本级聚合阶段,本文采用 Mean-Max 联合池化。均值池化反映所有变量 token 的整体分布,最大池化突出最强判别响应。二者拼接后得到 $g_i = [\text{mean}(H_i); \text{max}(H_i)]$,再输入前馈分类层得到 logits。相较于只使用单个 [CLS] 标记,Mean-

Max 池化在中小规模表格数据中更稳定，因为它直接聚合所有变量 token 的信息，降低了 CLS 标记训练不足带来的波动。

模型训练采用 AdamW 优化器，学习率设置为 0.001，权重衰减为 $1e-4$ 。由于本文实验数据集规模较小，为降低过拟合风险，在分类头中加入 Dropout，并采用重复随机划分实验评估模型稳定性。本文实现中嵌入维度设置为 32，注意力头数为 4，Transformer 编码器层数为 1。这一轻量化设置可以在保证表达能力的同时控制参数规模，更

适合中小样本多维分类场景。

5 实验设计与结果分析

本文采用两个公开多维分类数据集进行实验。Wine 数据集包含 178 个样本、13 个连续特征和 3 个类别；BreastCancer 数据集包含 569 个样本、30 个连续特征和 2 个类别。两个数据集均为典型小样本表格型多维数据，适合检验模型在有限样本条件下的分类性能。实验中将数据按 7:3 划分为训练集和测试集，采用随机种子 42、2024 和 2026 重复三次，并报告平均值与标准差。

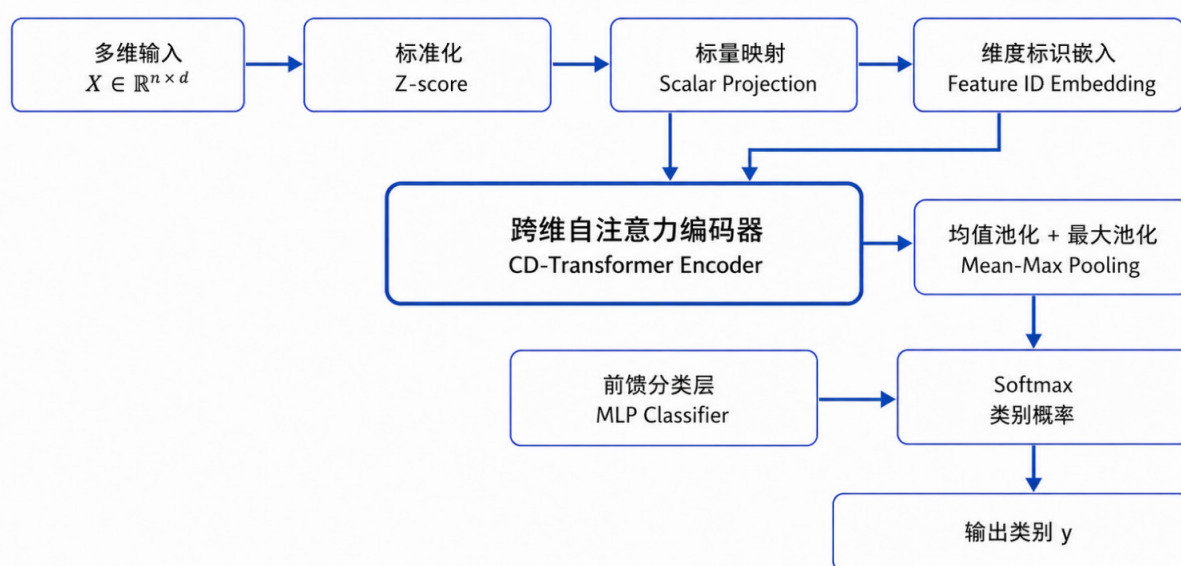


图 1 CD-Transformer 多维数据分类模型结构

对比模型包括逻辑回归（LR）、支持向量机（SVM）、随机森林（RF）、多层感知机（MLP）和本文提出的 CD-Transformer。LR 和 SVM 代表传统线性 / 核方法，RF 代表集成树模型，MLP 代表基础深度神经网络。评价指标包括 Accuracy、Precision、Recall 和 F1-score，均采用宏平均方式计算，以避免类别数量差异对结果造成偏差。

从表 3 可以看出，CD-Transformer 在 Wine 数据集上获得 0.9444 的平均准确率，在 BreastCancer 数据集上获得 0.9610 的平均准确率。与普通 MLP 相比，CD-Transformer 在两个数据集上均取得更高 Accuracy 和 F1-score，说明跨维注意力结构相较简单全连接网络能够更有效地表达变量交互关系。但在两个数据集上，SVM 和 LR 仍取得最高或接近最高的准确率，这说明对于样本规模较小、特征已经

较充分且线性可分程度较高的数据，传统模型仍具有较强竞争力。本文结果不简单宣称 Transformer 在所有表格任务中优于传统方法，而是表明其作为一种跨维关系建模框架具有可扩展价值。

图 2 展示了不同模型在两个数据集上的准确率对比。可以看到，CD-Transformer 明显优于 MLP，且在 BreastCancer 数据集上超过随机森林。图 3 给出了 CD-Transformer 训练损失曲线，两个数据集均表现出稳定下降趋势，说明模型训练过程没有出现明显震荡。图 4 展示了 BreastCancer 数据集上的混淆矩阵，模型对两类样本均具有较好区分能力。图 5 展示了前 18 个变量之间的注意力热力图，可作为观察变量交互强弱的辅助工具。虽然注意力权重不能直接等同于因果解释，但可以为后续特征筛选和专家分析提供线索。



表 1 实验数据集基本信息

数据集	样本数	特征数	类别数	任务说明
Wine	178	13	3	葡萄酒理化指标分类
BreastCancer	569	30	2	乳腺肿瘤良恶性分类

表 2 实验参数设置

参数项	设置
数据划分	训练集 70%，测试集 30%
重复次数	3 次随机划分
标准化	Z-score
嵌入维度	32
注意力头数	4
编码器层数	1
优化器	AdamW
学习率	0.001
权重衰减	1e-4
评价指标	Accuracy、Precision、Recall、F1-score

表 3 不同模型分类性能对比（均值 ± 标准差）

数据集	模型	Accuracy	Precision	Recall	F1-score
BreastCancer	CD-Transformer	0.9610 ± 0.0034	0.9578 ± 0.0047	0.9594 ± 0.0041	0.9585 ± 0.0035
BreastCancer	LR	0.9747 ± 0.0122	0.9744 ± 0.0116	0.9714 ± 0.0145	0.9728 ± 0.0131
BreastCancer	MLP	0.8967 ± 0.0531	0.8882 ± 0.0552	0.8976 ± 0.0501	0.8917 ± 0.0543
BreastCancer	RF	0.9474 ± 0.0101	0.9442 ± 0.0101	0.9433 ± 0.0117	0.9437 ± 0.0109
BreastCancer	SVM	0.9766 ± 0.0101	0.9751 ± 0.0109	0.9750 ± 0.0109	0.9750 ± 0.0108
Wine	CD-Transformer	0.9444 ± 0.0185	0.9479 ± 0.0213	0.9494 ± 0.0143	0.9466 ± 0.0183
Wine	LR	0.9877 ± 0.0107	0.9883 ± 0.0101	0.9894 ± 0.0092	0.9886 ± 0.0099
Wine	MLP	0.8765 ± 0.0875	0.8919 ± 0.0675	0.8836 ± 0.0714	0.8728 ± 0.0922
Wine	RF	0.9753 ± 0.0283	0.9744 ± 0.0283	0.9780 ± 0.0257	0.9758 ± 0.0272
Wine	SVM	0.9815 ± 0.0185	0.9853 ± 0.0145	0.9778 ± 0.0222	0.9806 ± 0.0195

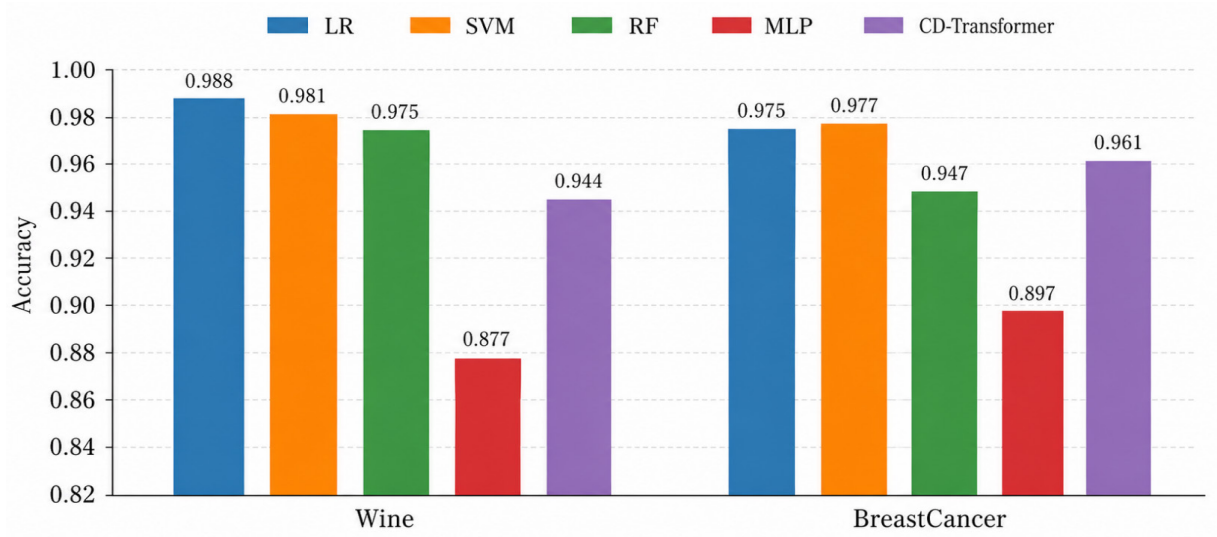


图 2 不同模型在多维数据集上的分类准确率对比

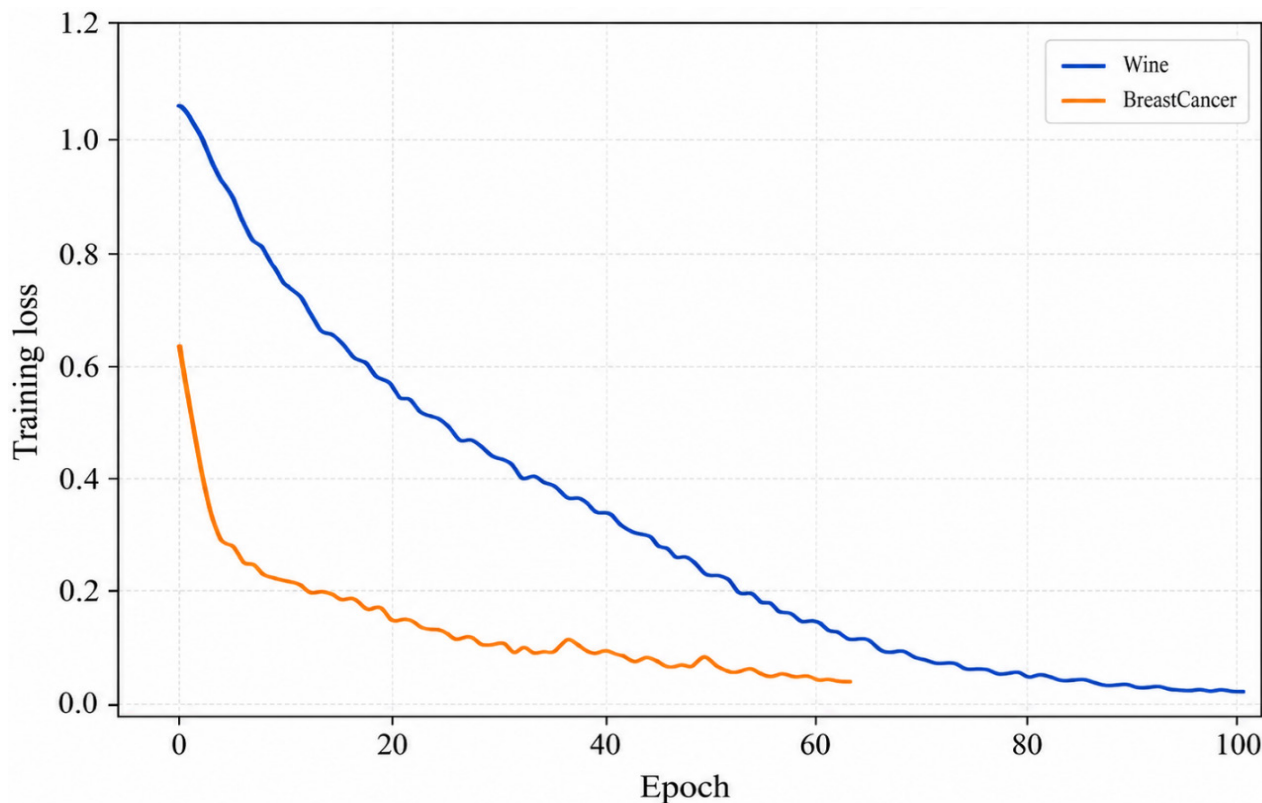


图 3 CD-Transformer 训练损失收敛曲线

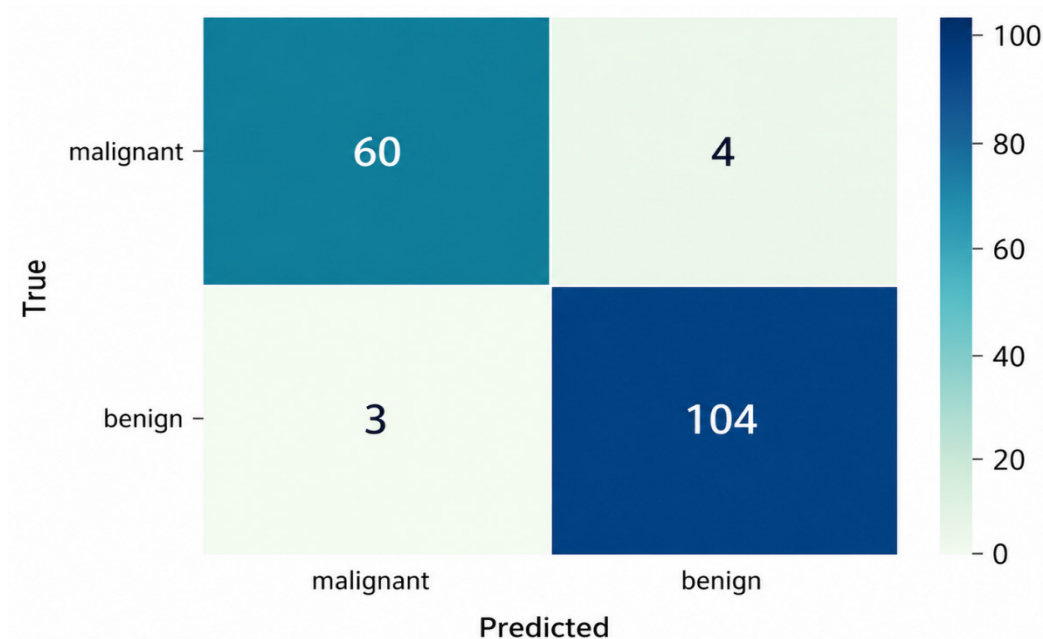


图 4 BreastCancer 数据集混淆矩阵

6 消融实验与讨论

为进一步分析模型组成部分对分类性能的影响，本文在 BreastCancer 数据集第三次随机划分上

开展消融实验。消融实验不以追求最高单次准确率为目标，而用于观察维度标识嵌入、跨维编码器和池化方式对模型行为的影响。结果见表 4。

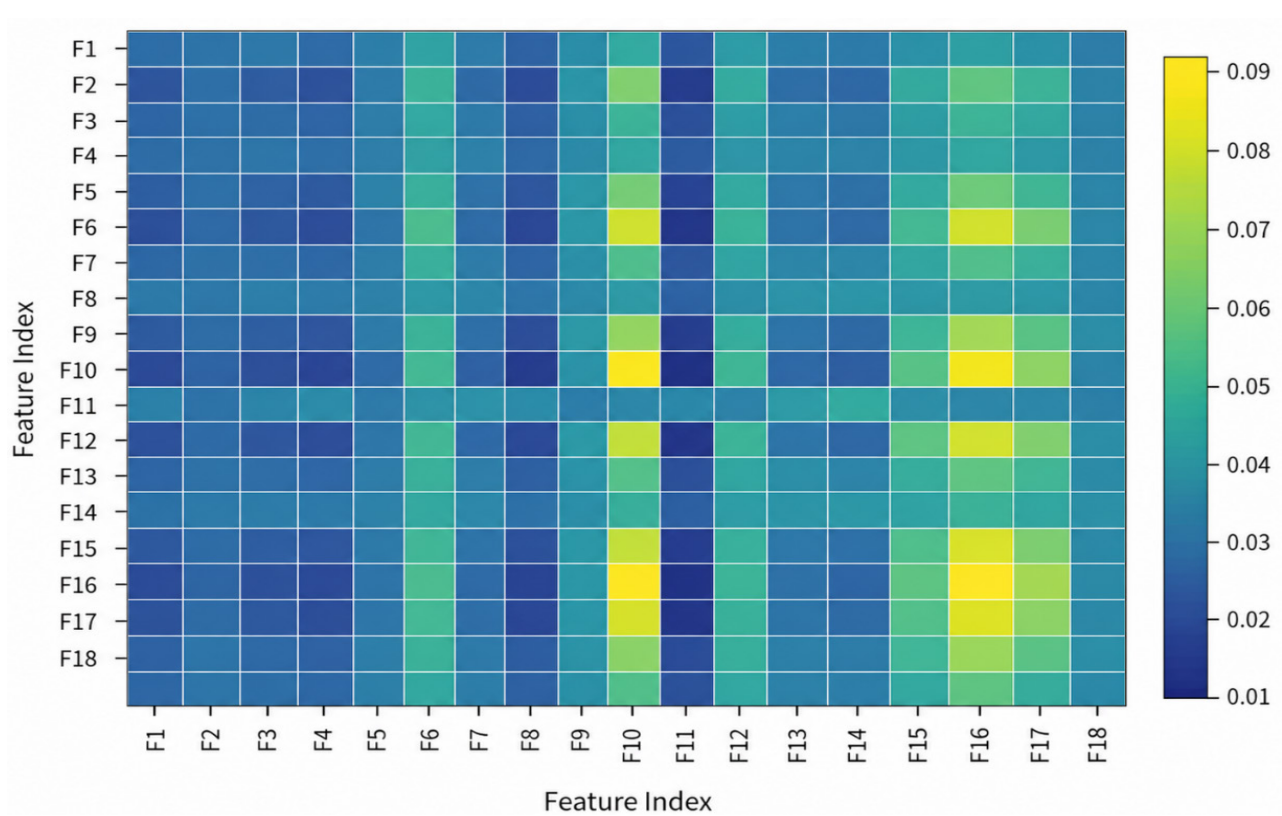


图 5 BreastCancer 前 18 个特征的跨维注意力热力图

表 4 BreastCancer 数据集消融实验结果

模型变体	Accuracy	Precision	Recall	F1-score
去除Transformer编码器	0.9708	0.9675	0.9704	0.9689
去除维度标识嵌入	0.8713	0.8856	0.8407	0.8553
仅均值池化	0.9591	0.9551	0.9579	0.9564
完整CD-Transformer	0.9649	0.9655	0.9594	0.9623

从模型表现看，CD-Transformer 的优势主要体现在三个方面。第一，它能够将每个变量作为独立 token 处理，从结构上显式建模变量间关系。第二，维度标识嵌入能够保留变量身份，避免所有数值特征在共享映射后语义混淆。第三，注意力热力图提供了可视化观察入口，使研究者能够初步判断哪些变量之间存在较强交互。对于医疗、工业和风控等场景，这种特征交互可解释性具有现实意义。

从消融结果看，去除维度标识嵌入后模型性能下降明显，说明在表格型多维数据中，变量身份信息非常关键。仅使用均值池化时，模型性能略低于完整模型，说明最大池化可以补充局部强

响应信息。不过，在 BreastCancer 这种小样本且可分性较强的数据集上，去除 Transformer 编码器的轻量模型在单次划分中也能取得较好表现。这一现象说明深度结构并非越复杂越好，模型选择应结合样本规模、特征噪声和任务难度进行判断。

本文模型仍存在局限。首先，实验数据集规模有限，尚不能完全代表工业大规模异构数据场景。其次，当前模型主要处理连续数值特征，对类别型变量、缺失值和混合模态数据的处理仍需扩展。再次，注意力热力图只能提供相关性层面的解释，不能直接证明变量之间存在因果关系。最后，在小样本场景中，Transformer 模型可能存在参数冗余问题，需要结合正则化、预训练或轻量化结



构进一步优化。

7 结论与展望

本文围绕多维数据分类问题,提出了一种基于 Transformer 的 CD-Transformer 模型。该模型将表格型多维特征转化为特征 token 序列,通过标量映射、维度标识嵌入和跨维自注意力机制学习变量间关系,并使用 Mean-Max 池化形成样本级表示。公开数据集实验表明,CD-Transformer 相较普通 MLP 具有更好的分类性能和更稳定的训练表现,在多维变量交互建模方面具有一定优势。

同时,实验也表明,在小样本、特征质量较高且线性可分程度较强的任务中,传统 SVM 和逻辑回归仍具有很强竞争力。因此,Transformer 并不是所有表格分类问题的绝对最优选择。其价值更体现在变量关系复杂、非线性交互明显、样本规模适中或需要特征交互解释的应用场景。未来研究可从三个方向继续展开:第一,引入类别型变量嵌入与缺失值掩码机制,提升模型对真实业务数据的适配能力;第二,结合自监督预训练方法,提高小样本条件下的泛化能力;第三,探索轻量化注意力机制,降低模型计算复杂度,使其更适合边缘端或工业现场部署。

参考文献:

[1] 任利强,贾舒宜,王海鹏,王子玲.基于深度学习的时间序列分类研究综述[J].电子与信息学报,2024,46(8):3094-3116.

[2] 张旭,张亮,金博,张红哲.基于不确定性的多元时间序列分类算法研究[J].自动化学报,2023,49(4):790-804.

[3] 王璐瑶,曹渊,刘博涵,曾恩,刘坤,夏元清.时间序列分类模型的集成对抗训练防御方法[J].自动化学报,2025,51(1):144-160.

[4] 王威娜,李明莉.基于 Shapelets 的多元时间序列分类方法[J].科学技术与工程,2025,25(1):252-261.

[5] 毛远宏,孙琛琛,徐鲁豫,刘曦,柴波,贺鹏超.基于深度学习的时间序列预测方法综述[J].微电子学与计算机,2023,40(4):8-17.

[6] 梁宏涛,刘硕,杜军威,胡强,于旭.深度学习应用于时序预测研究综述[J].计算机科学与探索,2023,17(6):1285-1300.

[7] 谷小兵,牛少彰,王茂森,安洪旭,史成洁.基于 Transformer 的时间序列插补技术研究[J].图像与信号处理,2024,13(2):151-162.

[8] 郑庆河,刘方霖,余礼苏,姜蔚蔚,黄崇文,桂冠.一种结合小波去噪卷积与稀疏 Transformer 的调制识别方法[J].电子与信息学报,2025,47(7):2361-2374.

[9] 刘辉,等.融合空间自注意力感知的严重缺失多元时间序列插补算法[J/OL].电子与信息学报,2025.

[10] 庄学彬,等.基于门控机制的并行 CNN-Transformer 神经网络的多参数欺骗检测方法[J/OL].电子与信息学报,2025.

作者简介:黄欣欣(2003-),女,汉族,福建泉州人,阳光学院本科在读,主要研究方向为数据分析、机器学习、图像识别;卜子洋(2001-),男,汉族,福建泉州人,阳光学院人工智能学院双创导师,中国国际科技促进会全生命周期绿色科技分会技术专家,主要研究方向为人工智能、机器学习、大数据分析等。