基于过程诱导二分轨迹的鲁棒区间标定方法研究

2026 年 1 月 11 日

1 研究背景与科学问题 (Introduction)

1.1 研究背景

在高精度工业制造与科学实验中，表格数据（Tabular Data）的回归任务无处不在。现有的深度表格学习模型（如 FT-Transformer, SAINT）主要致力于提升**点估计（Point Estimation）**的精度。然而，在诸如光模块电压调节、医疗剂量控制等高风险场景中，模型必须具备**量化不确定性（Uncertainty Quantification）**的能力，即在输出预测值的同时，给出一个高置信度的安全工作区间（Safety Interval）。

1.2 现有研究不足 (Research Gap)

当前的区间预测方法主要面临以下挑战：

1. **异方差性（Heteroscedasticity）难题：**工业数据往往存在严重的异方差性（即不同样本的噪声水平不同）。传统的均方误差（MSE）损失函数假设同方差高斯分布，无法自适应地调整区间宽度。

2. **缺乏物理可解释性：**现有的分位数回归（Quantile Regression）或贝叶斯神经网络（BNN）虽然能输出区间，但其推理过程通常是“黑盒”的，缺乏与物理调节过程的逻辑映射。

3. **确定性分类的局限：**虽然近期出现的序列化回归方法（如 Ord2Seq）在精度上取得了突破，但其本质仍是确定性的分类模型，缺乏对偶然不确定性（Aleatoric Uncertainty）的显式建模机制。

1.3 研究目标

本研究旨在提出一种名为 **TabSeq-Trace** 的生成式推理框架。核心思想是将回归任务重构为**“基于物理过程的逻辑推理”**，通过模拟二分查找的决策轨迹，在无区间标注监督的情况下，利用模型推理的“犹豫度”自适应地生成鲁棒的置信区间。

2 拟采用研究方法 (Proposed Methodology)

本研究将提出一套完整的生成式推理框架 **TabSeq-Trace**。该框架摒弃了传统的直接数值回归范式，转而采用“轨迹构造—逻辑推理—不确定性注入”的三阶段处理流程。

2.1 阶段一：物理同构轨迹构造 (Isomorphic Trace Construction)

为了模拟硬件 DAC（数模转换器）的物理调节过程，我们将连续的回归任务离散化为二分决策序列预测任务。

2.1.1 空间离散化与二分树映射

我们将连续的物理输出空间 $\mathcal{Y} = [0, V_{max}]$ 映射到一棵深度为 D 的完全二叉树 $\mathcal{T}$ 上。

- **物理对齐：** 树包含 $N = 2^D$ 个叶子节点，每个叶子节点对应硬件的最小分辨率区间（Bin）。
- **路径编码：** 树的每一层代表一次二分决策（左子树编码 0，右子树编码 1）。

2.1.2 标签序列化 (Label Serialization)

对于任意给定的真实标签 $y \in \mathcal{Y}$，我们回溯其在树上的唯一路径，将其转化为长度为 D 的二进制决策序列：

$$y_{seq} = [c_1, c_2, \ldots, c_D], \quad c_i \in \{0, 1\} \tag{1}$$

该序列经过右移操作后（$[s, c_1, \ldots, c_{D-1}]$），作为后续推理网络的自回归输入目标。

2.2 阶段二：深度感知推理网络 (Depth-Aware Reasoning Network)

本阶段构建模型主体，负责根据输入的表格特征重构上述决策轨迹。

2.2.1 异构特征编码 (Heterogeneous Encoder)

针对工业表格数据的异构性，我们拟采用 **FT-Transformer** 作为编码基座：

- **Feature Tokenizer：** 将数值特征（如温度）和类别特征（如厂商 ID）统一映射为高维嵌入向量 $e^{(j)}$。
- **Contextual Encoding：** 通过 Self-Attention 机制捕捉特征间的高阶交互，生成上下文特征矩阵 $X_{ctx} \in \mathbb{R}^{F \times d}$。

2.2.2 深度感知上下文注意力 (Depth-Aware Contextual Attention, DACA)

针对标准 Transformer 无法感知推理阶段差异的问题，我们提出 DACA 机制以模拟“由粗到细”的物理推理逻辑。在解码器的第 t 个时间步，我们通过位置编码生成动态门控 G_t：

$$G_t = \sigma(\text{MLP}(E_{pos}(t))) \tag{2}$$

并利用该门控动态调整 Cross-Attention 的 Key 矩阵：

$$K_t = X_{ctx} \odot G_t \tag{3}$$

此机制迫使模型在浅层决策（$t \leq 3$）时聚焦于厂商等静态属性，在深层决策（$t \geq 8$）时聚焦于温度等动态环境属性，实现特征层面的物理过程解耦。

2.3 阶段三：自适应不确定性学习 (Adaptive Uncertainty Learning)

这是本框架实现“区间预测”的核心。我们将结合多热监督信号与自适应掩码机制，使模型学会量化数据的不确定性。

2.3.1 多热监督信号 (Multi-hot Supervision)

为了使模型感知决策的“辖域”，我们在每一步生成的监督目标不是单一的 0/1，而是一个全量程的多热向量 $y^t_{mht} \in \{0,1\}^N$。

$$y^{t,j}_{mht} = \begin{cases} 1 & \text{若叶子节点} j \text{ 位于第} t \text{ 步选择的正确子树内} \\ 0 & \text{其他情况} \end{cases} \tag{4}$$

这迫使模型学习当前合法的数值范围，而不仅仅是二分方向。

2.3.2 自适应置信度掩码 (Adaptive Confidence Masking)

为了捕捉数据的异方差性，我们拟构建动态掩码函数 $\alpha(x,t)$ 来加权损失函数：

$$\alpha(x,t) = \alpha_{depth}(t) \cdot \alpha_{instance}(x) \tag{5}$$

- **深度衰减** $\alpha_{depth}(t)$：随深度增加而降低，反映对深层细微偏差的容忍。
- **实例感知** $\alpha_{instance}(x)$：由轻量级网络 $\phi(X_{ctx})$ 实时预测。对于高噪样本（烂芯片），α 自动降低。

损失函数定义：

$$\mathcal{L} = \sum_{t=1}^{D} M_t(\alpha) \odot \text{BCE}(\hat{p}_t, y^t_{mht}) \tag{6}$$

当 α 降低时，模型在非目标分支上的惩罚减小，从而被允许保留“犹豫”（概率质量），导致预测分布的熵增加，区间变宽。

2.4 阶段四：全息推理与区间生成 (Holographic Inference)

在推理阶段，我们将计算全量程的生存概率累积，而非单一路径搜索：

$$P(V_i) = \prod_{t=1}^{D} P(\text{node}^{(i)}_t | \text{path}_{<t}) \tag{7}$$

基于生成的离散概率密度函数 (PDF)，我们计算累积分布函数 (CDF) 并截取 $1-\epsilon$ 置信区间 $[L,U]$，作为最终的安全工作边界。

3 实验验证计划 (Experimental Plan)

3.1 数据集

- **工业数据集：**某光模块产线的电压标定数据（包含多厂商、多工况，具有显著异方差性）。
- **公开基准：**拟选用常用的回归基准（如 California Housing, YearPredictionMSD）以验证方法的泛化性。

3.2 评价指标

- **点估计精度：** MAE, RMSE。
- **区间质量：** PICP（区间覆盖率，目标 $\geq 90\%$）与 MPIW（平均区间宽度，越窄越好）。
- **可解释性：** DACA 注意力热力图分析。

4 预期创新点与贡献 (Expected Contributions)

1. **范式创新：**提出 TabSeq-Trace 框架，首次将表格回归任务完整重构为“轨迹构造—深度推理—全息标定”的生成式过程。
2. **机制创新：**提出的 DACA 机制解决了序列模型在表格推理中的特征解耦问题；自适应掩码机制解决了单点监督下的异方差不确定性量化难题。
3. **工程价值：**预期提供一套物理可解释、鲁棒的标定算法，显著提升工业控制系统的安全性。