刘思阳

liusy@lamda.nju.edu.cn · 个人主页

谷歌学术 · OpenReview · GitHub

教育背景

南京大学 2024.09 – 2027.06（预计）

学术型硕士 人工智能学院 | LAMDA 研究所| 导师：叶翰嘉

研究方向：表格数据基座模型. 表格数据（Tabular Data）是结构化数据的典型形式，以行列结构组织，广泛应用于推荐系统、*AI4Science*、医疗诊断、风险评估等关键场景，是机器学习应用最广泛的数据类型之一. 但因数据异质性限制，表格数据难以进行联合训练，导致深度学习模型的性能一直难以超过传统树模型. 本人的研究围绕提升深度学习在表格数据预测任务的效果展开，以表格数据基座模型为核心，聚焦于以下三个子方向：表格数据基座模型，表格模型测评，大模型辅助表格基座模型预测.

南京大学 学士 人工智能学院 | GPA 4.45 / 5 2020.09 – 2024.06

科研成果

1) 表格模型测评与理解：系统性开展表格深度模型的评测、分析与综述工作.

- TALENT: A Tabular Analytics and Learning Toolbox | ***Si-Yang Liu***, et al. | **JMLR, CCF-A, 第一作者**
  ★ 已接收. 统一评测 *30+* 表格深度模型的开源工具包，*GitHub Star > 800.*
- A Closer Look at Deep Learning Methods on Tabular Datasets | Han-Jia Ye, ***Si-Yang Liu***[*], Hao-Run Cai[*], et al. | **JMLR Major Revision，CCF-A，学生第一作者**
  ☆ Major Revision. 在 *300+* 数据集上统一评估 *30+* 种模型，揭示数据异质性对性能的关键影响.
- Representation Learning for Tabular Data: A Comprehensive Survey | Jun-Peng Jiang, ***Si-Yang Liu***, et al. | **TPAMI，CCF-A，第二作者**
  ★ 已接收. 首个按泛化能力对表格表示学习进行层次化分类的综述.

2) 表格数据基座模型：聚焦于构建、理解和复用具备跨表格迁移能力的表格预训练模型.

- TabPFN Unleashed: A Scalable and Effective Solution to Tabular Classification Problems | ***Si-Yang Liu***, Han-Jia Ye | **ICML 2025，CCF-A，第一作者**
  ★ 已接收. 提出表格基座模型高效复用方案 BETA，显著提升表格基座模型在复杂任务上的性能.
- A Closer Look at TabPFN v2: Strength, Limitation, and Extension | Han-Jia Ye, ***Si-Yang Liu***[S], Wei-Lun Chao. | **NeurIPS 2025，CCF-A，学生第一作者**
  ★ 已接收. 首个系统性分析表格基座模型的工作，揭示其机制并提出复用策略和改进方案.
- SwiftPFN: Revisiting Row-Wise Attention–Only Tabular Foundation Models with Adaptive Early Exit | ***S-Yang Liu***, Han-Jia Ye. | **第一作者**
  ☆ Under Review. 提出轻量化表格基座模型，在保证性能的同时简化结构并提升推理效率.
- From Universal Prediction to Universal Workflows: A Survey of Foundation Models for Structured Data | ***Si-Yang Liu***, Han-Jia Ye. | **第一作者**
  ☆ Under Review. 统一对表格与时序数据基座模型进行分类的综述，并将视角拓展到通用工作流.

3) 大语言模型辅助表格基座模型预测：研究如何借助大语言模型进一步增强表格基座模型能力.

- Lookahead Automated Feature Engineering for Tabular Prediction via Kaggle-Guided Knowledge Transfer | ***Si-Yang Liu***, et al. | **第一作者**
  ☆ Under Review. 输入层面增强：结合 *Kaggle* 知识与 *LLM* 进行前瞻性自动化表格数据特征工程
- Make Still Further Progress: Chain of Thoughts for Tabular Data Leaderboard | ***Si-Yang Liu***[*], Qile Zhou[*], Han-Jia Ye. | **ICML 2025 FMSD Workshop，共一第一**
  ★ 已接收. 输出层面增强：利用 *LLM* 实现表格数据实例级动态推理与模型智能集成.
- TWINS: Synergizing Tabular Data and Text Signals via Joint Representation Learning | Jun-Peng Jiang, ***Si-Yang Liu***, et al. | ☆ 第二作者 Under Review.

4) 表格基座模型的应用与拓展：探索表格基座模型在新领域的应用.

- CausalFM: A Scalable Data-Driven Foundation Model for Causal Discovery | Zi-Rong Li, ***Si-Yang Liu***, Tian-Zuo Wang, Han-Jia Ye. | ☆ 第二作者 Under Review.
- On the Effectiveness of Tabular Representation for EEG Foundation Model | Hao-Run Cai, ***Si-Yang Liu***, Han-Jia Ye, Hai-Rong Zheng. | ☆ 第二作者 Under Review.

5) 其他工作.

- R^3-VAE: Reference Vector-Guided Rating Residual Quantization VAE for Generative Recommendation | Qiang Wan, Ze Yang, Dawei Yang, Ying Fan, Xin Yan, ***Si-Yang Liu***. | ☆ Under Review

工作影响力和代码开源

- 主导开发的开源工具包 **TALENT** 获 **GitHub Star 超 800**，被国内外多个研究组采纳为表格学习的实验基准.
- 相关工作被 Michael I. Jordan (美国三院院士、中国科学院外籍院士)、Francis Bach (法国科学院院士)等国际知名学者的工作使用或拓展 (如 KE-TALENT).
- 提出的 Embedding 提取代码集成至官方 **TabPFN-Extensions** 工具库，成为 contributor 之一.
- Google Scholar统计引用数 200 +.

实习经历

字节跳动–今日头条推荐–基础技术 | 算法实习生 2025 年 7 月 –

背景：Semantic ID（SID）是生成式模型与去 ID 化 CTR 模型的重要组件；负责 SID 相关底层表征建设与评估.

- SID 残差编码：基于快手 OneRec 思路完成 RQ-kmeans 残差编码实现与超参数优化，助力头条精排 CTR 去 ID 模型和 OneRec 生成式模型.
- SID 平行编码：探索码本大小、层数与优化策略，8192^4 码本在离线测试集上冲突率约 4.1%，方案已接入发文流.
- SID 特征回溯与压缩：搭建多级 SID 合并为 Dense 特征的离线回溯流程，将 SID 离线验证存储成本压缩到原方案的 **1/4** 左右，并系统评估 10+ 版本 SID 与不同码本使用策略在 CTR 小模型上的效果.
- 构建头条文本 OneRec 多任务预训练数据 packing 框架，并探究不同码本的语义对齐能力.

★ 硕士阶段获奖情况

南京大学优秀研究生标兵 2024–2025 年度
比亚迪奖学金 2024–2025 年度
南京大学学业奖学金 2024–2025 年度
南京大学学业奖学金 2023–2024 年度

其他

- **会议审稿人**: AAAI'25, ICLR'25, IJCAI'25, ICML'26, ICLR'26, PAKDD'26
- **课程助教**: 机器学习导论 (2024 秋 & 2025 秋), 面向本科生