2024 年 7 月 17 日

目前研究面临的几个问题：

1. 目前尚未找到一个综合的、通用的数据清洗方法，在不同错误类型下和面向不同 ML 应用下都表现良好的数据清洗模型。我的观点是短期之内也无法找到，即使能找到，模型也将非常复杂。
2. 设计数据清洗方法需要具体面向数据的错误类型和下游的 ML 应用（如分类、聚类、回归等），这使得我们设计算法过程中要有针对性。面向具体错误类型和具体应用的清洗方法设计可以使研究更加深入。
3. 针对 ML 模型是否必要的问题，两篇论文中都做了详细的回答。目前的情况是使用 ML 的清洗方法有很多缺点，比如运行时间长、针对性不强、可扩展性差。在很多数据集上的准确性提高的不多但是要付出极大的时间代价（100 倍）。如果用简单的模型能在短时间内达到与机器学习相近的准确度，那么 ML 模型就不是必须的，比如面向分类应用。
4. REIN 的论文中提到将清洗数据用于 ML 的回归和聚类应用时，清洗算法的鲁棒性一般，此时 ML 是有必要的，但是分类应用使用 ML 意义不大。所以一个研究思路是把他们细化，评估 ML 方法在数据清洗工程中的必要性。
5. 两篇论文都着重提到了用户规则和约束条件对错误检测器的性能至关重要，二者可以采用自动化工具进行整合，而且这个问题是错误检测和修复的基础，接下来研究要着重注意这个方面，可以分析影响程度并考虑具体的整合策略和评估方法。

可能的研究思路：

1. 论文中对于研究步骤的观点我非常支持，数据清洗分成错误检测和错误修复两个步骤，而且前者是后者的基础。目前的研究成果是错误检测方法比较全面和细致，但是错误修复方面存在大量的不足，非常有待改进，属于研究重点。我的想法是可以针对具体的错误检测方法提出相对应的修复方法，二者具有连带关系，但是具体怎么做还没有想好。有一些方法错误检测很准确，但错误修复不理想，我们可以提出一个基于现有方法的修复策略，例如 XXXX + repair.
2. 两篇论文的实验部分已经非常详细的分析了现有的这些先进清洗方法的缺点和不足。因此我们下一个研究重点应该就是如何在现有的方法上改进并提出新的算法。我个人的观点是，某些算法的思路和见解很好，我们在原有基础上调整和改进比提出一个全新模型更具有可操作性，比如提出 XXXX clean model 2.0 等。当然，由于算法的本质差异，此处的改进策略需要将 ML 和非 ML 区分开。
3. 浙大的论文对已有的评估模型进行了修改和补充，很具有启发性，这使我思考现有评估模型参数的局限性问题。我们下一步需要从实际问题出发而不是只停留在理论角度来建立参数。
4. 除此之外，关于模型鲁棒性和可扩展性的检测指标还是没有详细的定义，论文中通过改变数据集类型和数据集规模的方式来进行定性评估，我们可以考虑定量模型以及形式化的定义，让它们成为实验验证中可观测到的一个提升目标。
5. 浙大的论文中特别提到了大语言模型 LLM 的应用价值，由于其较强的推理能力和广泛的知识库，可以基于 LLM 提出一种新的错误修复方案并进行评估，这也可以作为我们的一个研究方

向。