2024年7月23日

第一阶段任务：

分析并总结现有ML修复模型的具体原理和机制，以及它们在下游ML聚类应用中的性能。

|  |  |  |
| --- | --- | --- |
|  | 理论部分 | 实验部分 |
| 1 | 分析ML清洗算法的相关参考文献，确定每种算法具体采取了哪种ML类型，以及具体的过程和原理是什么。 | 整理和补充实现这些算法的代码 |
| 2 | 对每种错误类型，从理论角度分析它们在ML算法的哪个步骤中被引入或者被放大，并探讨可能的原因。 | 整理面向聚类任务相关的数据集，为实验做准备 |
| 3 | 针对6种不同的聚类算法，进一步从理论角度分析这些ML算法是如何作用和影响下游任务的。 | 针对这些数据集对相关ML清洗算法进行评测，比较它们在不同聚类算法上的性能，并于理论结果比较 |

|  |  |
| --- | --- |
| 清洗算法列表 | 文献链接 |
| Scare | https://dl.acm.org/doi/abs/10.1145/2463676.2463706 |
| Baran | https://dl.acm.org/doi/abs/10.14778/3407790.3407801 |
| Holoclean | https://arxiv.org/abs/1702.00820 |
| Metadata-Driven | https://dl.acm.org/doi/abs/10.1145/3221269.3223028 |
| RAHA | https://dl.acm.org/doi/abs/10.1145/3299869.3324956 |
| ED2 | https://dl.acm.org/doi/abs/10.1145/3357384.3358129 |
| Picket | https://link.springer.com/article/10.1007/s00778-021-00699-w |
| ActiveClean | https://dl.acm.org/doi/abs/10.14778/2994509.2994514 |
| Boostclean | https://arxiv.org/abs/1711.01299 |
| CPClean | https://arxiv.org/abs/2005.05117 |

聚类算法

Gaussian Mixture (GMM)

K-Means

Affinity Propagation (AP)

Hierarchical Clustering (HC)

OPTICS

BIRCH