Bachelor Thesis Proposal: Benchmarking Tabular Data Augmentation Techniques for Deep Clustering

Mamdouh Aljoud

February 2026

1 Introduction

Clustering tabular data is important in many applications where labels are unavailable. However, classical clustering methods often struggle on raw tabular features due to mixed feature types, noise, and nonlinear relationships. Deep clustering addresses this by learning a latent representation that is easier to cluster.

Recently, contrastive learning has been used to learn robust representations by pulling together different augmented "views" of the same sample and pushing other samples apart. While this works well in vision, tabular data augmentation is less standardized: some augmentations may help representation learning, while others may distort the data and harm clustering.

This thesis studies how different tabular augmentation techniques affect the performance of deep clustering models.

2 Problem Statement

Which tabular augmentation techniques are most effective for learning cluster-friendly representations with contrastive learning, and how do they impact deep clustering performance?

3 Tasks

- Conduct a literature review on tabular data augmentation and contrastive/self-supervised learning for tabular data [1, 2]
- Implement a contrastive learning benchmark for tabular data using simple augmentations (e.g., feature masking, noise, feature swapping)
- Evaluate clustering quality using known metrics (e.g., NMI, ARI)
- Investigate the limitations and failure cases of augmentation techniques

4 Requirements

- Basic knowledge of machine learning and deep learning
- Python programming skills (preferably PyTorch)
- Interest in unsupervised learning

5 Contact

If you are interested, please send an email to `aljoud@dbs.ifi.lmu.de` with your CV and transcript, and briefly describe your interest in this topic.

References

[1] Dara Bahri, Heinrich Jiang, Yi Tay, and Donald Metzler. SCARF: Self-Supervised Contrastive Learning using Random Feature Corruption, March 2022.

[2] Zaitian Wang, Pengfei Wang, Kunpeng Liu, Pengyang Wang, Yanjie Fu, Chang-Tien Lu, Charu C. Aggarwal, Jian Pei, and Yuanchun Zhou. A Comprehensive Survey on Data Augmentation, October 2025.