Master Thesis Topic: AutoML for Clustering in Active Learning

Winter Semester 2025/26

1 Introduction and Background

Expert-labeled data is often scarce and expensive to obtain. As such, there are multiple strategies to minimize the amount of labeled data required to train a good classifier. Active Learning [7, 5] is one such strategy. It works by automatically selecting samples that are assumed to provide the biggest gain in quality if labeled according to some measure and requesting them to be labeled by an oracle (e.g., a human annotator).

Among these Active Learning methods, TypiClust [3] employs the deep clustering method SCAN [2] to aid in this selection. From each cluster obtained from SCAN, the most typical sample is picked to be labeled. Clustering, as an unsupervised learning approach, can inherently produce many valid label assignments, depending on the chosen parameterization. Subsequently, these do not necessarily align with the distribution of the ground truth labels, which could lead to problems with the sample selection. Thus, it is important to pick suitable hyperparameters to prevent this.

An intuitive way to improve hyperparameter selection is Automated Machine Learning (AutoML) [4], which automatically tunes the configuration of algorithms to maximize their performance according to a metric. While there has been some research into AutoML with clustering [6, 1], the issue of the dependency on the chosen metric is a problem for its application. However, for Active Learning, some labels are available, whether from an initial starting budget or prior rounds through the oracle. As such, an AutoML process using the labels as a ground truth is possible and may yield a better performance for the overall choice of samples to label, as the clustering obtained this way should more closely align with the desired segmentation and a fully labeled dataset. Ultimately, the goal is to apply AutoML, utilizing the ground truth information to tune the clustering aspect of TypiClust and to investigate the impact of proper alignment between the clustering and the underlying ground truth on the overall Active Learning performance.

2 Research Question

To which extent does applying AutoML to the clustering step in TypiClust improve the choice of samples for the Oracle?

3 Tasks & Goals

- **Literature Review:** Understanding the primary literature regarding Active Learning, Deep Clustering (in the context of SCAN), and AutoML for Clustering
- **Data Preparation:** Selection of a suitable dataset for the experiments based on the criteria of non-trivial classification, clustering suitability, and runtime
- **Implementation:** Adjust existing implementation of TypiClust (`https://github.com/avihu111/TypiClust`) to include SMAC3 (`https://github.com/automl/SMAC3`) to tune the three main hyperparameters of SCAN: entropy weight, confidence threshold and number of neighbors
- **Introductory Analysis:** Analysis of the (3D) hyperparameter space concerning the clustering performance on starting labels and full-dataset, as well as the accuracy of the final Active Learning classifier
- **Evaluation:** Performance both in regard to clustering quality as well as the active learning accuracy under consideration of the default configuration
- **Discussion:** Analysis of experiments, particularly in regards to trade-offs and design decisions made

4 Expected Outcomes

- Combining the AL approach Typiclust with AutoML for the clustering step;
- Running experiments and comparing the method with the original Typiclust;
- In-depth analysis of influencing components and hyperparameters.
- A well-documented thesis with reproducible code.

5 Requirements

- Study in the field of computer science

- Prior programming experience in Python
- Beneficial: understanding of active learning
- GPU access: for the purposes of this thesis, access to university GPUs will be provided

6 Contact

If you are interested, please send your CV and transcripts to jahn@dbs.ifi.lmu.de

References

[1] Matheus Camilo da Silva, Gabriel Marques Tavares, Eric Medvet, and Sylvio Barbon Junior. Problem-oriented automl in clustering. *CoRR*, abs/2409.16218, 2024.

[2] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. SCAN: learning to classify images without labels. In *ECCV (10)*, volume 12355 of *Lecture Notes in Computer Science*, pages 268–285. Springer, 2020.

[3] Guy Hacohen, Avihu Dekel, and Daphna Weinshall. Active learning on a budget: Opposite strategies suit high and low budgets. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 8175–8195. PMLR, 2022.

[4] Xin He, Kaiyong Zhao, and Xiaowen Chu. Automl: A survey of the state-of-the-art. *Knowl. Based Syst.*, 212:106622, 2021.

[5] Dongyuan Li, Zhen Wang, Yankai Chen, Renhe Jiang, Weiping Ding, and Manabu Okumura. A survey on deep active learning: Recent advances and new frontiers. *IEEE Trans. Neural Networks Learn. Syst.*, 36(4):5879–5899, 2025.

[6] Yannis Poulakis, Christos Doulkeridis, and Dimosthenis Kyriazis. A survey on automl methods and systems for clustering. *ACM Trans. Knowl. Discov. Data*, 18(5):120:1–120:30, 2024.

[7] Burr Settles. Active learning literature survey. 2009.