Exploring Dataset Label Properties in Semi-Supervised Clustering

Marek Landert Johannes Nüesch Lukas Nüesch

1. Motivation

In many real-world scenarios, datasets often contain sparse labels or lack labels for certain classes entirely. Understanding how different types of label sparsity influence clustering performance is crucial for designing robust semi-supervised learning systems.

This project aims to investigate the impact of sparse labels on clustering performance by building on the DeepCluster architecture. While DeepCluster operates fully unsupervised using k-means, we will extend it to use sparse label data. Namely, we modify the convolutional neural network's (CNN) loss function and add explicit soft constraints to k-means with PCKMeans.

Through a series of experiments, we aim to understand how dataset properties and label characteristics influence the clustering outcomes. These insights could guide future approaches for semi-supervised and constrained clustering.

2. Related work

Deep Neural Networks and clustering were first combined in (Tian et al., 2014). The DeepCluster architecture on which our work is based on is presented in (Caron et al., 2019). There are other approaches that make use of DNNs for clustering and also propose extensions for semi-supervised deep clustering (Ren et al., 2019), (Wang et al., 2023), (Xu et al., 2023). (Ren et al., 2024) present an overview covering different Deep Clustering methods. There has been work on semi-supervised clustering in the presence of specific label properties. For example (Sachdeva et al., 2023) propose a robust method to handle label noise. (Willetts et al., 2020) and (Guo et al., 2022) propose approaches to correctly detect and classify unseen classes.

3. Methods

DeepCluster is an unsupervised learning framework that jointly trains a CNN and performs clustering in an iterative process. It alternates between using a clustering algorithm (e.g., k-means) to group CNN features into clusters and employing the resulting cluster assignments as pseudo-labels to update the CNN weights through supervised learning.

Besides modifying the loss function of the CNN we can also directly add label information to k-means. To that end, we will use PCKMeans (Berry et al., 2004), which extends k-means by adding soft must-link and cannot-link connections.

4. Datasets

We plan to conduct our experiments on MNIST. If we find MNIST to be in insufficient or inappropriate for our experiments we will consider also testing on CIFAR-10/CIFAR-100 or even Synthetic Datasets.

5. Baselines

Our baseline will be the original fully unsupervised DeepCluster architecture using k-means without any label support.

6. Experiments

We will explore various label properties to systematically evaluate their effect on clustering performance:

- Overall Sparsity: Varying the percentage of labeled images in the dataset (e.g., 1%, 5%, 10%, 20%).
- Sparsity Patterns: Varying the class-wise label distribution. (e.g. half of the classes have no labels)
- Noise: Introducing incorrect labels to evaluate the model's ability to handle mislabeled data.
- Label Granularity: Testing the effect of varying label granularity (e.g., coarse-grained vs. fine-grained class labels).
- Dynamic Data: Introducing labels dynamically during training to simulate real-world scenarios where labeled data becomes available over time (temporal sparsity).

7. Expected Results

We hope to gain insights into the impact of various types of label sparsity on clustering performance and recommendations for handling sparse labels in semi-supervised clustering tasks.

References

Berry, M. W., Dayal, U., Kamath, C., and Skillicorn, D. *Proceedings of the 2004 SIAM International Conference on Data Mining (SDM)*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2004. doi: 10.1137/1.9781611972740.

Caron, M., Bojanowski, P., Joulin, A., and Douze, M. Deep clustering for unsupervised learning of visual features, 2019.

Guo, L.-Z., Zhang, Y.-G., Wu, Z.-F., Shao, J.-J., and Li, Y.-F. Robust semi-supervised learning when not all classes have labels. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 3305–3317. Curran Associates, Inc., 2022.

Ren, Y., Hu, K., Dai, X., Pan, L., Hoi, S. C., and Xu, Z. Semi-supervised deep embedded clustering. *Neurocomputing*, 325:121–130, 2019. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2018.10.016.

Ren, Y., Pu, J., Yang, Z., Xu, J., Li, G., Pu, X., Yu, P. S., and He, L. Deep clustering: A comprehensive survey. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–21, 2024. doi: 10.1109/TNNLS.2024.3403155.

Sachdeva, R., Cordeiro, F. R., Belagiannis, V., Reid, I., and Carneiro, G. Scanmix: Learning from severe label noise via semantic clustering and semi-supervised learning. *Pattern Recognition*, 134:109121, 2023. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2022.109121.

Tian, F., Gao, B., Cui, Q., Chen, E., and Liu, T.-Y. Learning deep representations for graph clustering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 28(1), Jun. 2014. doi: 10.1609/aaai.v28i1.8916.

Wang, Y., Zou, J., Wang, K., Liu, C., and Yuan, X. Semi-supervised deep embedded clustering with pairwise constraints and subset allocation. *Neural Networks*, 164: 310–322, 2023. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2023.04.016.

Willetts, M., Roberts, S., and Holmes, C. Semi-unsupervised learning: Clustering and classifying using ultra-sparse labels. In *2020 IEEE International Conference on Big Data (Big Data)*, pp. 5286–5295, 2020. doi: 10.1109/BigData50022.2020.9378265.

Xu, X., Hou, H., and Ding, S. Semi-supervised deep density clustering. *Applied Soft Computing*, 148:110903, 2023. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc.2023.110903.