Exploring Dataset Label Properties in Semi-Supervised Clustering

Marek Landert Johannes Nüesch Lukas Nüesch

[1]

Abstract

This work extends the DeepCluster framework by incorporating sparse label information through PCK-means must-link and cannot-link constraints. Using the MNIST dataset, we analyze how label sparsity, distribution, noise, granularity, and dynamic availability affect clustering performance. Our results show that even minimal supervision (1% labeled data or less) significantly enhances clustering performance, even under significant noise. However, balanced class labels are crucial for optimal performance, whereas coarse labels offer only marginal gains. These findings highlight the potential of semi-supervised clustering.

1. Introduction

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 extend it to use sparse label data. Namely, we add explicit soft constraints to k-means with PCK-means.

Through a series of experiments, we analyze how label characteristics influence the clustering outcomes.

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 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. Models and 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.

To directly add label information to k-means, we use PCK-means (Basu et al., 2004), which extends k-means by adding soft must-link and cannot-link constraints. Using these types of constraints provides a flexible way of introducing different types of label information. We use a slightly modified version of the implementation provided in (DatamoleAI). The main idea behind PCK-means is to modify the clustering objective function to include weighted constraint violations.

$$\begin{aligned}\mathcal{J}_{\text{pckm}} &= \frac{1}{2}\sum_{\mathbf{x}_i\in\mathcal{X}}\|\mathbf{x}_i-\boldsymbol{\mu}_{l_i}\|^2 \\ &+ \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{M}} w_{ij}\mathbf{1}[l_i\neq l_j] + \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{C}} \overline{w}_{ij}\mathbf{1}[l_i = l_j]\end{aligned} \tag{1}$$

The variables in the equation are defined as follows: $\mathcal{X}$ is the set of data points, $\boldsymbol{\mu}_{l_i}$ is the centroid of the cluster assigned to l_i, w_{ij} and $\overline{w}_{ij}$ are the weights for must-link and cannot-link constraints, respectively, $\mathcal{M}$ is the set of must-link pairs, and $\mathcal{C}$ is the set of cannot-link pairs.

To generate the must-link and cannot-link constraints for PCK-means, we take a subset of labels (e.g. 1%) and generate all resulting must-link and cannot-link constraints. To balance the number of must-link and cannot-link constraints, we only use a subset (10%) of the cannot-link constraints.

A CNN with three convolutional layers (8, 16, 32 channels)

[1] GitHub repo: `https://github.com/lnueesch/dl-project`

with max-pooling and batch normalization, followed by two fully connected layers (128, 64 neurons) with dropout, was chosen for fast training and to observe the effects of label variations.

4. Datasets

We conducted our experiments on the MNIST database of handwritten digits, using the training dataset, which consists of 60,000 images.

5. Baselines

Our baseline employs the original, fully unsupervised DeepCluster architecture with k-means, without any label integration. Our focus lies on examining the relative impact of different label information rather than achieving optimal absolute performance.

6. Experiments

We systematically evaluated the impact of various label properties on clustering performance through a series of experiments. Unless otherwise indicated, we fixed the number of (PC)k-means iterations to 5, the number of training epochs to 10, used 10% of all possible cannot-link constraints and all must-link constraints.

6.1. Overall Sparsity

We investigated the effect of varying the percentage of labeled samples within the dataset. The experiments included sparsity levels of 0.05%, 0.1%, 0.2%, 0.5%, and 1%.

6.2. Label Noise

We introduced label noise to evaluate the robustness to mislabeled data. A fixed percentage of the labeled samples (10%, 20%, 50%, 70%, 90%, 100%) was deliberately assigned incorrect class labels. The noise was applied uniformly across all labeled samples. This setup allowed us to test the model's ability to learn meaningful representations despite noisy supervision.

6.3. Class-Specific Label Distribution

We examined the scenario where labels are not equally spread over all classes and are only available for a subset of all possible classes (1, 2, 5, 8 classes).

6.4. Label Granularity

We examined the effect of label granularity by providing cannot-link constraints between superclasses, which include multiple digits. We compare superclasses which are constructed based on visual features (digits with vertical lines, digits with circles) to random selections of digits.

6.5. Dynamic Label Availability

To simulate real-world conditions where labeled data becomes available over time, we introduced labels dynamically during training. At regular intervals, additional labeled samples were provided to the model. This experiment evaluates the model's ability to adapt to a continuously evolving training set and how delayed labels affected the final clustering performance.

7. Results

In the following, the results of the experiments described above are presented and compared to the baseline (using k-means) and each other.

7.1. Equal label distribution without Noise

When introducing labeled data to the training of the DeepCluster architecture, the NMI of the clustering increases. Figure 3 shows the evolution of NMI with respect to the true labels at different overall sparsity levels. It can be seen that, the more data is labeled, the higher the NMI. Similar effects can be observed for other metrics like ARI and Silhouette score (see table 1). When comparing the T-SNE visualization of k-means (figure 1) and PCK-means clustering with 1% labeled data (figure 2), it can be qualitatively observed that the clustering performance with label support is closer to the ground truth. The silhouette score and T-SNE visualization both indicate improved cluster separation in the semi-supervised case.

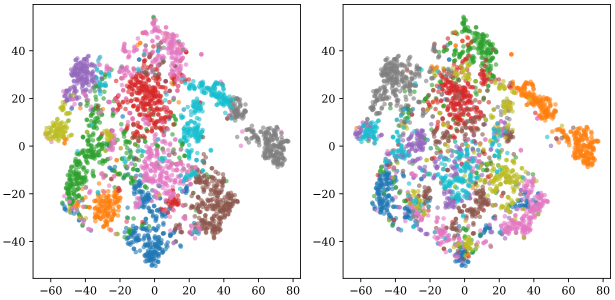

Figure 1. T-SNE visualization of baseline k-means clusters (left) and true labels (right) at epoch 10.

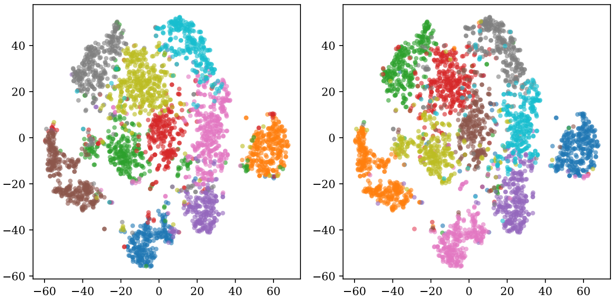

Figure 2. T-SNE visualization of PCK-means clusters (left) and true labels (right) at epoch 10 and 1% labeled, highlighting the alignment between clustering results and ground truth.

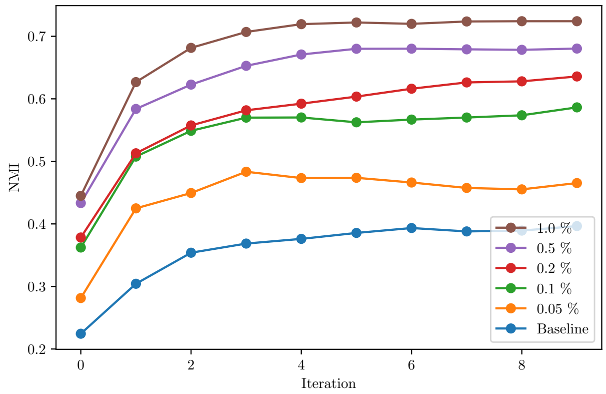

Figure 3. Evolution of NMI with respect to true labels at different sparsity levels.

Table 1. Overall sparsity results.

LABELED (%)	NMI	ARI	SILHOUETTE
0.0	0.40	0.26	0.23
0.05	0.47	0.29	0.28
0.1	0.59	0.48	0.30
0.2	0.64	0.53	0.33
0.5	0.68	0.61	0.35
1	0.72	0.69	0.37

7.2. Equal label distribution with Noise

Figure 4 shows the NMI scores at different noise levels while keeping the label sparsity at 1%. Notice that as long as the noise ratio is lower than 90%, semi-supervised clustering outperforms the baseline.

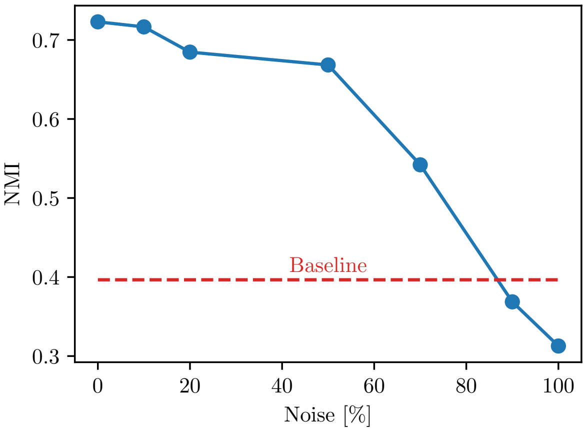

Figure 4. Evolution of NMI with respect noise levels. Using 1% of labels.

7.3. Class-Specific Label Distribution

Figure 5 shows the NMI scores for increasing numbers of labeled classes when keeping the label sparsity at 1%. Noticeably, the baseline gets outperformed if the labels come from more than one class.

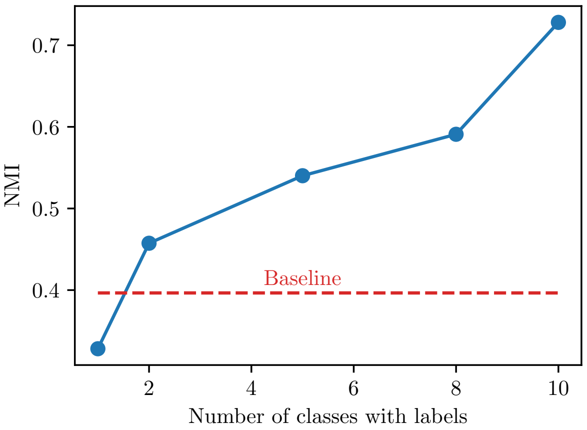

Figure 5. Evolution of NMI with varying number of classes with labels. Using 1% of labels

7.4. Label Granularity

Creating superclasses of multiple digits is done by adding cannot-links between the superclasses. Must-links cannot be used. Notably, our model performs worse than the baseline only using cannot-links (Table 2). Handcrafted groups do not show improvements.

Table 2. NMI scores using different granularities

Labeled (%)	NMI
Baseline	0.40
2 groups(vertical lines)	0.30
2 groups(circles)	0.28
2 groups random	0.30
4 groups random	0.27
5 groups random	0.32

7.5. Dynamic Label Availability

The evolution of NMI, when increasing the percentage of labeled data during training, is shown in Figure 6. With scheduled dynamic labels, the NMI lies between the baseline and the 1% labeled PCK-means version, and does not reach it, even though in the end, the percentage of labeled data is the same. Our chosen label sparsity schedule starts at 0% for the first two epochs, increases to 0.1% at epoch 3, doubles approximately every two epochs, and reaches 1% by epoch 8, remaining constant thereafter.

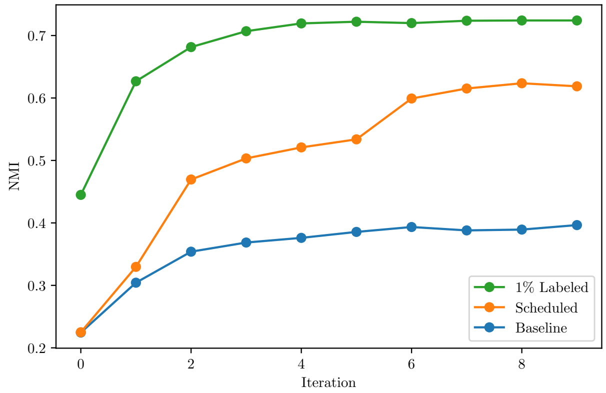

Figure 6. Evolution of NMI with while increasing labels every epoch

7.6. Comparison to Existing Work

The results of (Ren et al., 2019) provide an opportunity to compare our clustering performance to existing work, also using pair-wise constraints on the MNIST dataset. While we are using a similar number of constraints (60′000 total) for this comparison, note that the sampling of the constraints is done differently.

8. Discussion

Our results show that adding even very little supervision can significantly improve clustering performance. This is shown by greatly improved NMI, ARI and silhouette scores with less than 1% of the data.

Table 3. Comparison of clustering performance on MNIST to results of (Ren et al., 2019)

Method	NMI	ARI
k-means	0.50	0.37
PCK-means	0.50	0.38
AE+KM	0.72	0.65
AE+KM-cst	0.73	0.67
DEC	0.82	0.77
IDEC	0.78	0.73
SDEC	0.83	0.79
DeepCluster (Ours)	0.40	0.26
DeepCluster+PCK-means (Ours)	0.73	0.69

Even when using noisy labels, i.e. some mislabeled samples, the performance still improves over the baseline when the label noise is smaller than 90%. While this might seem surprising at first, this makes sense since 90% noise is equivalent to just providing uniformly random labels. This indicates that this approach is very robust to heavy noise.

Our experiments also show that it is beneficial to have access to balanced class labels. If certain classes lack labels entirely, this affects clustering negatively even if the number of labels stays constant. Nevertheless, even just labels from 2 out of 10 classes provide improvements over the baseline.

When only provided with coarse labels, we can't see any improvements. Rather our model performs worse. One possible reason is that cannot-link constraints carry less information since they only penalize grouping certain points together but do not actively force any points to a cluster. The algorithm may therefore reassign points in inefficient ways, incurring penalties but not gaining any meaningful structure in return.

Furthermore, our clustering approach shows versatile learning capabilities as performances benefit, even if label availability is increased only later during training.

Overall, our results show that semi-supervision for deep clustering approaches has great potential even when label information is sparse, unbalanced, or noisy.

9. Summary

In this work, we extended the DeepCluster framework by integrating PCK-means constraints to incorporate sparse label information. Through extensive experiments on the MNIST dataset, we demonstrated that even minimal supervision (less than 1% labeled data) significantly enhances clustering performance across various label sparsity, noise, and distribution scenarios. Our results highlight the robustness of semi-supervised clustering to noise and its sensitivity to balanced and granular label distributions. These findings highlight the potential of sparse supervision to improve clustering with limited and imperfect labels.

References

Basu, S., Banerjee, A., and Mooney, R. J. *Active Semi-Supervision for Pairwise Constrained Clustering*, pp. 333–344. Society for Industrial and Applied Mathematics, 2004. doi: 10.1137/1.9781611972740.31. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611972740.31.

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

DatamoleAI. active-semi-supervised-clustering. https://github.com/datamole-ai/active-semi-supervised-clustering. Accessed: 2025-01-14.

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.