Intra-view and Inter-view Correlation Guided Multi-view Novel Class Discovery

Supplementary Material

1. Datasets

We use eight multi-view datasets in our experiments, each described in detail below.

1. **BRCA**[1]: This dataset is designed for classifying PAM50 subtypes in breast invasive carcinoma, incorporating three distinct omics data types: messenger RNA (mRNA), Copy Number Variation (CNV), and Reverse-Phase Protein Array (RPPA). It consists of 511 samples, divided into four categories: Luminal A, Luminal B, Triple-Negative Breast Cancer, and HER2-positive.
2. **KIPAN**[2]: This dataset is designed for kidney cancer subtype classification. It integrates three types of omics data: DNA methylation, miRNA expression, and mRNA expression. It contains 707 samples, divided into three kidney cancer subtypes: KICH (Kidney Chromophobe), KIRC (Kidney Renal Clear Cell Carcinoma), and KIRP (Kidney Renal Papillary Cell Carcinoma).
3. **uci-digit**[3]: This dataset is designed for handwritten digit recognition. It contains images of digits from 0 to 9, with 2,000 samples divided into 10 categories.
4. **Cora**[4]: A widely used benchmark for machine learning and network analysis, this dataset comprises 2,708 computer science publications. It provides two main views: a citation network formed by paper citations and content-related words for each paper. The publications are divided into seven research area categories, such as Case-Based, Genetic Algorithms, and Neural Networks.
5. **Wiki**[5]: This dataset is used for text classification, containing feature representations of 2,866 Wikipedia articles divided into 10 categories.
6. **CCV**[6]: This dataset is used for video classification tasks and contains consumer videos from YouTube across categories such as sports, music, and movies. It includes 6,773 video clips.
7. **STL10**[7]: This dataset is used for image recognition tasks and includes 10 object classes.
8. **YTB10**[8]: This dataset is used for face recognition tasks, containing 38,654 face images extracted from YouTube videos.

[1] https://www.cancer.gov/tcga
[2] https://www.linkedomics.org/data_download/TCGA-KIPAN/
[3] https://archive.ics.uci.edu/dataset/72/multiple+features
[4] https://linqs.org/datasets/
[5] https://archive.ics.uci.edu/dataset/72/multiple+features
[6] http://www.ee.columbia.edu/ln/dvmm/CCV/
[7] https://cs.stanford.edu/~acoates/stl10/
[8] https://www.cs.tau.ac.il/~wolf/ytfaces/

Table 1. The influence of pseudo-labels on our model.

Datasets	Random	Sinkhorn-knopp	k-means	Ours
BRCA	64.24	63.03	60.61	**98.79**
KIPAN	54.60	54.13	90.64	**92.51**
uci-digit	29.50	27.60	93.40	**95.30**
Cora	28.33	28.47	29.47	**76.36**
Wiki	22.74	23.46	31.21	**65.42**
CCV	12.86	12.68	30.45	**34.20**
STL10	21.65	21.86	98.98	**99.02**
YTB10	33.86	33.86	79.08	**94.55**

2. The influence of pseudo-labels

In our paper, we mentioned that the use of pseudo-label guidance during the clustering process can lead to unstable model performance. To validate this hypothesis, we replaced our algorithm with pseudo-label induced clustering. Specifically, we added a constraint term for pseudo-labels similar to the known class label constraint term in Eq. (3). There are three methods for generating pseudo-labels: random labeling, using the Sinkhorn-Knopp algorithm, and labeling with the k-means algorithm. The results are shown in the Table 1, which demonstrates that different methods of pseudo-label guided clustering significantly impact model performance. Therefore, it is essential to use algorithms that do not rely on pseudo-labels during the process of discovering new classes.