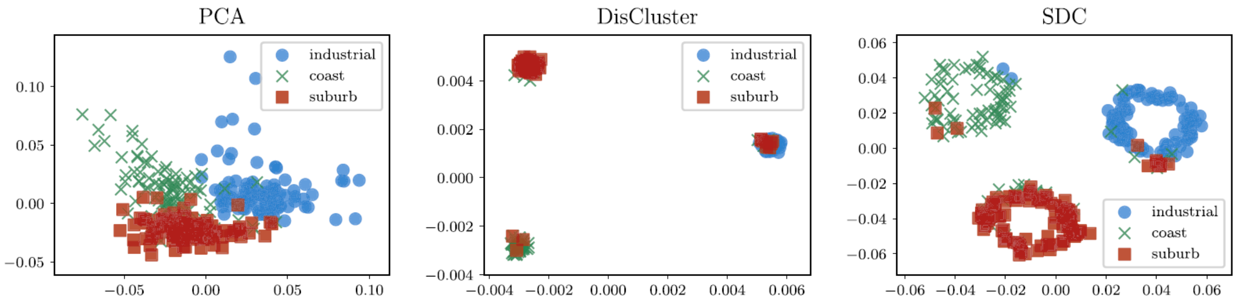
Discriminative clustering using regularized subspace learning

Passalis and Tefas, 2019, *Pattern Recognition*

- **Topic**
  - Discriminative clustering (clustering + subspace learning for classification)
- **Problems**
  - No guarantee the discovered clusters correspond to the class of the data
- **Motivation**
  - Motivation example: DisCluster overfits the data



  - Learning a regularized representation that can increase cluster separability
  - Not being overly confident on the given class labels
- **Proposed method: SDC (Similarity-based Discriminative Clustering)**
  - Given dataset $\mathcal{X} = \{x_1, \cdots, x_N\}$, $x \in \mathbb{R}^d$, learning projection function
    $f_W: \mathbb{R}^d \to \mathbb{R}^m \ (m < d)$, $f_W(x) = W^T x \ (d \times m)^T \times (d \times 1)$
  - $P(N \times N)$: similarity matrix of the projected data
    $$P_{ij} = \exp\left(-\left\|f_W(x_i) - f_W(x_j)\right\|_2^2 / \sigma\right)$$
  - $T(N \times N)$: target matrix
    $$T_{ij} = \begin{cases} a_{intra} (< 1) & x_i \text{ and } x_j \text{ belong to the same cluster} \\ a_{inter} (> 0) & \text{otherwise} \end{cases}$$
  - $M(N \times N)$: weight matrix
    $$M_{ij} = \begin{cases} 1 & x_i \text{ and } x_j \text{ belong to the same cluster} \\ 1/(C-1) & \text{otherwise} \end{cases}$$

- Error function
  $$J_S(\mathcal{X}, W) = \frac{1}{2\|M\|_1} \sum_{i=1}^{N} \sum_{j=1}^{N} M_{ij} \left(P_{ij} - T_{ij}\right)^2$$
  where $\|M\|_1 = \sum_{i=1}^{N} \sum_{j=1}^{N} \left|M_{ij}\right|$
- Orthogonality regularization term (Passalis and Tefas, 2018)
  $$J_P(W) = \frac{1}{2m^2} \|W^T W - I_{m \times m}\|_F^2$$
- Optimization problem (continuous and differentiable)
  $$\underset{W}{\operatorname{argmin}} (2 - \alpha) J_S(\mathcal{X}, W) + \alpha_p J_P(W)$$

Algorithm 1 Similarity-based Discriminative Clustering Algorithm.

Input: A set of points $\mathcal{X}$, the batch size N_{batch}, the number of optimization steps N_{iters}, gradient descent iterations $N_{sgditers}$, and clusters N_C

Output: The clustering solution $\mathcal{S}_1$

```
procedure SIMILARITY-BASED DISCRIMINATIVE CLUSTERING
    Calculate the initial clustering solution S_0 using k-means
    (i.e., solve the problem defined in (6))
    Set S = S_0
    for i from 1 to N_iters do
        for i from 1 to N_sgditers do
            Sample a batch of data x
            Construct the target similarity matrix for the selected
            samples x using (8) and the current solution S.
            Perform one optimization iteration using (5)
        Project the data samples into the new low-
        dimensional space defined by f_W(·)
        Calculate the updated clustering solution S_1 using k-
        means on the
low-dimensional representation (i.e., solve the problem defined
in (12))
        S = S_1
    return the final clustering solution S
```

- Gradients (v-th column vector $W_{\cdot v}$ of matrix W)

$$\frac{\partial J_S(\mathcal{X}, W)}{\partial W_{uv}} = \frac{1}{\|M\|_1} \sum_{i=1}^{N} \sum_{j=1}^{N} M_{ij}(P_{ij} - T_{ij}) \frac{\partial P_{ij}}{\partial W_{uv}}$$

$$\frac{\partial J_P(W)}{\partial W_{\cdot v}} = \frac{2}{m^2} \sum (W_{\cdot v}^T W_{\cdot k} - \delta_{vk}) W_{\cdot k}$$

$$\delta_{vk} = I(v = k)$$

➢ Experiments

- 4 dataset
- The number of clusters was set to the number of classes
- $m = 50$
- Sample results

Table 2
Spectral clustering evaluation.

Method	Dataset	Rand	NMI	Homogeneity	Completeness	FMI
Original	Yale	0.025 ± 0.003	0.247 ± 0.007	0.245 ± 0.008	0.248 ± 0.007	0.051 ± 0.002
PCA	Yale	0.027 ± 0.003	0.251 ± 0.006	0.249 ± 0.006	0.253 ± 0.006	0.053 ± 0.007
DisCluster	Yale	0.039 ± 0.005	0.284 ± 0.009	0.281 ± 0.010	0.288 ± 0.008	0.067 ± 0.016
SDC	Yale	**0.108 ± 0.005**	**0.411 ± 0.006**	**0.408 ± 0.006**	**0.414 ± 0.006**	**0.133 ± 0.005**
Original	15-scene	0.189 ± 0.016	0.353 ± 0.018	0.351 ± 0.018	0.354 ± 0.018	0.244 ± 0.009
PCA	15-scene	0.193 ± 0.014	0.355 ± 0.017	0.354 ± 0.017	0.357 ± 0.017	0.247 ± 0.007
DisCluster	15-scene	0.212 ± 0.013	0.381 ± 0.013	0.380 ± 0.013	0.382 ± 0.013	0.265 ± 0.007
SDC	15-scene	**0.241 ± 0.016**	**0.420 ± 0.017**	**0.418 ± 0.017**	**0.422 ± 0.017**	**0.292 ± 0.001**
Original	Corel	0.096 ± 0.005	0.431 ± 0.003	0.435 ± 0.004	0.426 ± 0.003	0.110 ± 0.003
PCA	Corel	0.099 ± 0.004	0.436 ± 0.002	0.441 ± 0.002	0.430 ± 0.003	0.113 ± 0.003
DisCluster	Corel	0.103 ± 0.002	0.439 ± 0.003	0.444 ± 0.003	0.434 ± 0.003	0.117 ± 0.006
SDC	Corel	**0.116 ± 0.005**	**0.452 ± 0.002**	**0.456 ± 0.003**	**0.448 ± 0.002**	**0.129 ± 0.002**
Original	KTH	0.403 ± 0.000	0.532 ± 0.000	0.524 ± 0.000	0.540 ± 0.000	0.507 ± 0.000
PCA	KTH	0.400 ± 0.002	0.528 ± 0.002	0.521 ± 0.002	0.535 ± 0.002	0.504 ± 0.000
DisCluster	KTH	0.400 ± 0.001	0.541 ± 0.001	**0.537 ± 0.001**	0.545 ± 0.001	0.502 ± 0.000
SDC	KTH	**0.411 ± 0.003**	**0.543 ± 0.003**	0.536 ± 0.003	**0.551 ± 0.003**	**0.513 ± 0.001**

➢ References

- Passalis and Tefas, 2018, Dimensionality reduction using similarity-induced embeddings (Orthogonality regularization term)
- **To be added more**