Exemplar-based Subspace Clustering for Class-imbalanced Data

Chong You, Chi Li, Daniel P. Robinson, René Vidal

Johns Hopkins University, Baltimore, USA

JOHNS HOPKINS MATHEMATICAL INSTITUTE for DATA SCIENCE

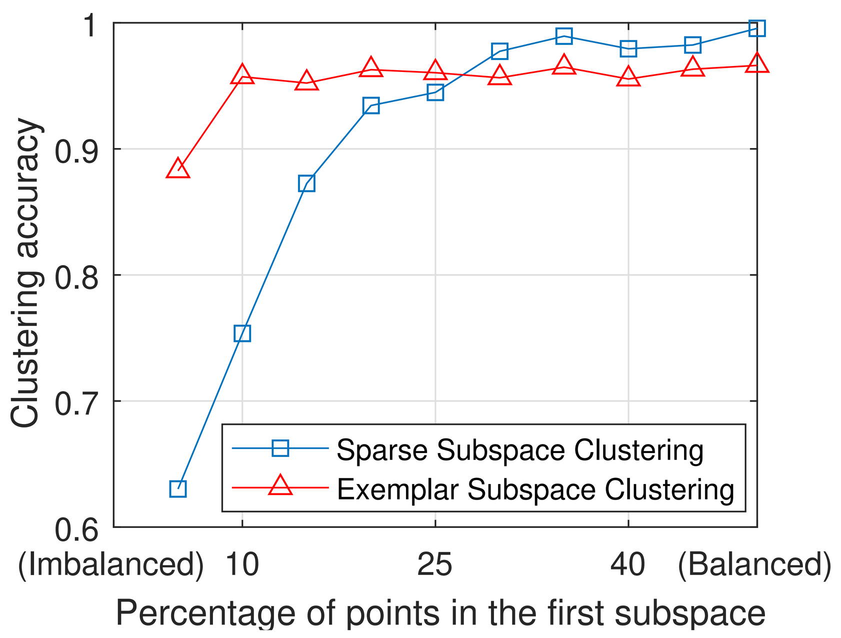
ECCV 2018 European Conference on Computer Vision 8 – 14 September 2018 | Munich, Germany

Introduction

- Vision datasets often contain multiple classes, each lying in a low-dimensional subspace
- **Subspace clustering** can discover these subspaces in an unsupervised manner

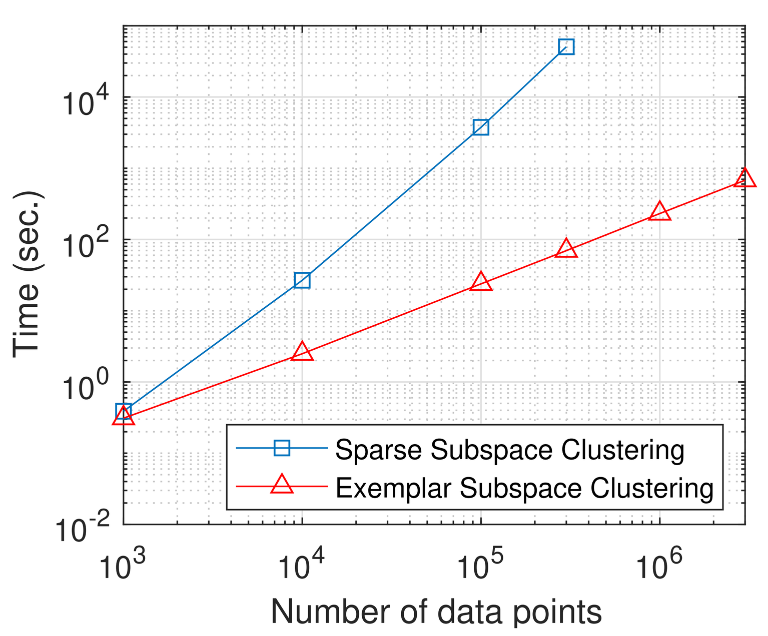
Challenges

- Many vision datasets are imbalanced and the performance of classical methods degrades with the imbalance level



- Subsampling methods do not apply since the dataset is unlabeled before clustering

- Many vision datasets are large scale and classical methods are not able to deal with large-scale data



- Scalable methods do exist, but they still have quadratic complexity

Contributions

Key idea: select a set of exemplars $\mathcal{X}_0^* \subseteq \mathcal{X} := \{\mathbf{x}_j\}_{j=1}^N$ and use them to cluster $\mathcal{X}$

- Our method handles class-imbalanced data if the selected set of exemplars is balanced
- Our method handles large-scale data if the selected set of exemplars is small

Prior method

Each point $\mathbf{x}_j \in \mathcal{X}$ is expressed in terms of a few points in $\mathcal{X}$, i.e.

$$\min_{\mathbf{c}_j \in \mathbb{R}^N} \|\mathbf{c}_j\|_1 + \frac{\lambda}{2} \cdot \|\mathbf{x}_j - \sum_{i \neq j : \mathbf{x}_i \in \mathcal{X}} c_{ij}\mathbf{x}_i\|_2^2$$

Proposed solution

Each point $\mathbf{x}_j$ in $\mathcal{X}$ is expressed in terms of a few exemplars in $\mathcal{X}_0^*$, i.e.

$$\min_{\mathbf{c}_j \in \mathbb{R}^N} \|\mathbf{c}_j\|_1 + \frac{\lambda}{2} \cdot \|\mathbf{x}_j - \sum_{i : \mathbf{x}_i \in \mathcal{X}_0^*} c_{ij}\mathbf{x}_i\|_2^2$$

Exemplar Subspace Clustering

- **Step 1:** Select a set of exemplars $\mathcal{X}_0^*$ that minimizes the self-representation cost function

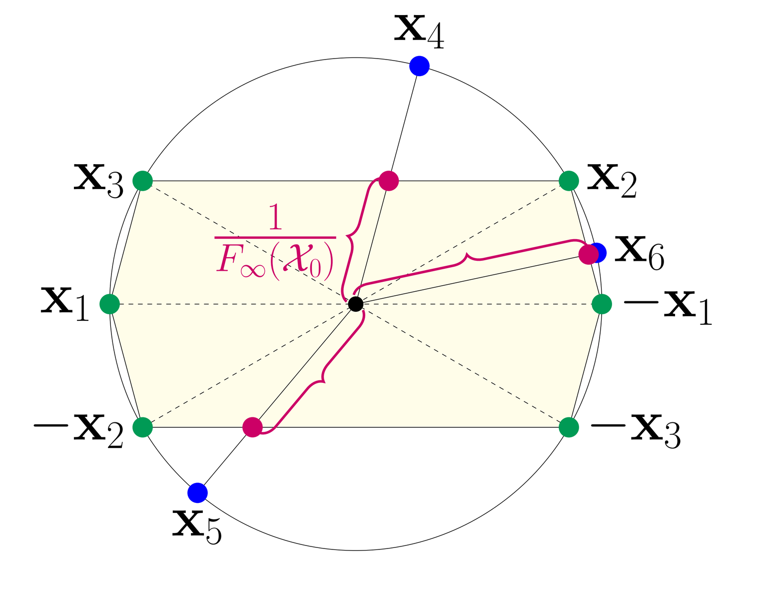
$$F_\lambda(\mathcal{X}_0) := \max_{\mathbf{x}_j \in \mathcal{X}} \min_{\mathbf{c}_j \in \mathbb{R}^N} \|\mathbf{c}_j\|_1 + \frac{\lambda}{2}\|\mathbf{x}_j - \sum_{i : \mathbf{x}_i \in \mathcal{X}_0} c_{ij}\mathbf{x}_i\|_2^2, \quad \text{where} \quad \lambda \in (1, \infty] \tag{1}$$

– $F_\lambda(\mathcal{X}_0)$ measures how well the data $\mathcal{X}$ is covered by the exemplars $\mathcal{X}_0$ (see figure on right)

- **Step 2:** For each $\mathbf{x}_j \in \mathcal{X}$, compute $\mathbf{c}_j^*$ by solving the following optimization problem

$$\mathbf{c}_j^* = \arg\min_{\mathbf{c}_j \in \mathbb{R}^N} \|\mathbf{c}_j\|_1 + \frac{\lambda}{2} \cdot \|\mathbf{x}_j - \sum_{i : \mathbf{x}_i \in \mathcal{X}_0^*} c_{ij}\mathbf{x}_i\|_2^2$$

- **Step 3:** Compute nearest neighbor affinity $\mathbf{A}$ from $\{\mathbf{c}_j^*\}_{j=1}^N$ and apply spectral clustering
- **Theorem 1:** i) $\mathcal{X}_0^*$ contains at least $\dim(\mathcal{S})$ points from each subspace $\mathcal{S}$, ii) the affinity $\mathbf{A}$ has no wrong connections. The result holds even if data $\mathcal{X}$ is class imbalanced



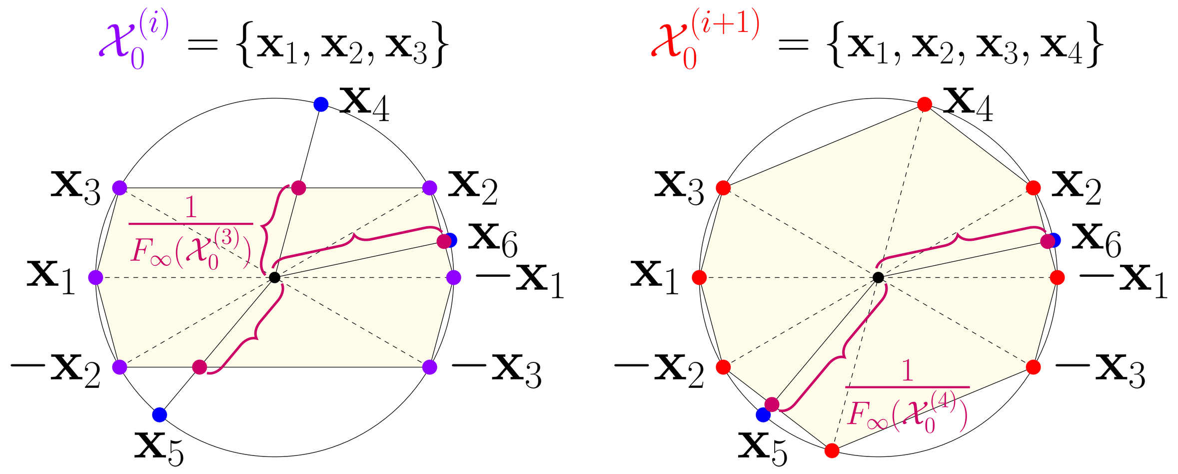
Let $\mathcal{X} = \{\mathbf{x}_j\}_{j=1}^6$ and $\mathcal{X}_0 = \{\mathbf{x}_j\}_{j=1}^3 \subseteq \mathcal{X}$. The shaded area represents $\text{conv}(\pm\mathcal{X}_0)$. $F_\infty(\mathcal{X}_0)$ is the reciprocal of the length of the ray $\{t\mathbf{x}_4 : t \geq 0\}$ inside $\text{conv}(\pm\mathcal{X}_0)$

A Farthest First Search (FFS) Algorithm

- Since minimizing (1) is NP-hard in general, we propose to compute $\mathcal{X}_0^{(k)}$ by iteratively selecting the worst represented point (see figure on right)

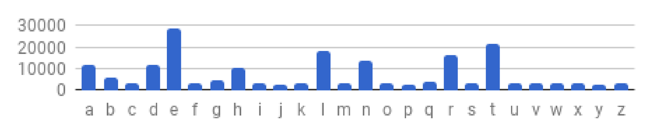
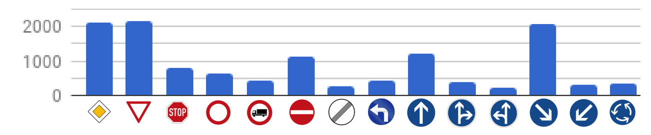
1: Select $\mathbf{x} \in \mathcal{X}$ at random and set $\mathcal{X}_0^{(1)} \leftarrow \{\mathbf{x}\}$
2: **for** $i = 1, \cdots, k-1$ **do**
3: $\mathcal{X}_0^{(i+1)} = \mathcal{X}_0^{(i)} \cup \arg\max_{\mathbf{x}_j \in \mathcal{X}} \min_{\mathbf{c}_j \in \mathbb{R}^N} \|\mathbf{c}_j\|_1 + \frac{\lambda}{2}\|\mathbf{x}_j - \sum_{i : \mathbf{x}_i \in \mathcal{X}_0^{(i)}} c_{ij}\mathbf{x}_i\|_2^2$
4: **end for**

- **Theorem 2:** $\mathcal{X}_0^{(k)}$ found by FFS satisfies the statement of Theorem 1



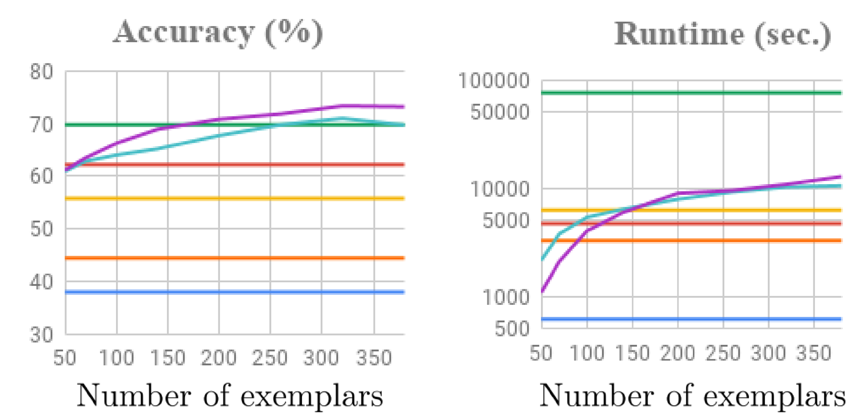
Experiments on Street Sign and Letter Image Databases

- We use two datasets that are class imbalanced and large scale
  - EMNIST: handwritten letter database containing 190,998 images
  - GTSRB: street sign database containing 12,390 images

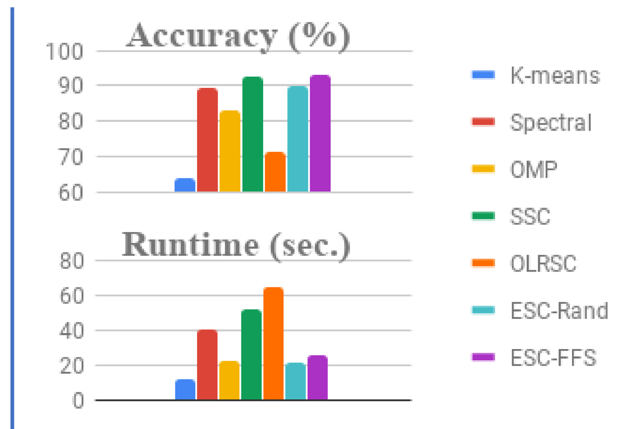



Number of points in each class of EMNIST (left) and GTSRB (right)

- On EMNIST, we vary the number of exemplars $k \in [50, 380]$
  - Accuracy: ESC-FFS outperforms SSC when $k > 200$
  - Runtime: ESC-FFS is $\sim$ 10 times faster than SSC, and is similar to ESC-Rand which uses random exemplar selection
- On GTSRB, we fix the number of exemplars to be 160
  - ESC-FFS has highest accuracy and moderate runtime



Results on EMNIST



Results on GTSRB

[1] E. Elhamifar, R. Vidal, Sparse Subspace Clustering: Algorithm, Theory, and Applications, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.
[2] C. You, D. Robinson, R. Vidal, Scalable Sparse Subspace Clustering by Orthogonal Matching Pursuit, *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
[3] C. You, C.-G. Li, D. Robinson, R. Vidal, Oracle Based Active Set Algorithm for Scalable Elastic Net Subspace Clustering, *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
[4] J. Shen, H. Xu, P. Li, Online low-rank subspace clustering by basis dictionary pursuit, *International Conference on Machine Learning*, 2016.

This work was supported by NSF under grant #1618637 and by IARPA under grant #127228.