Exemplar-based Subspace Clustering for Class-imbalanced Data

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

Johns Hopkins University, Baltimore, USA

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

Contributions

Key idea: select a set of exemplars $\mathcal{X}_6^*$ that minimizes the self-representation cost function

$$ F_6(\mathcal{X}_6) := \max_{x_i \in \mathcal{X}_6} \min_{c_i \in \mathbb{R}^d} \|c_i\|_1 + \frac{1}{2} \|x_i - \sum_{j \in \mathcal{X}_6} c_j x_j\|_2^2, $$

where $\lambda \in (1, \infty)$ (1)

- $F_6(\mathcal{X}_6)$ measures how well the data $\mathcal{X}$ is covered by the exemplars $\mathcal{X}_6$ (see figure on right)

Step 1: Select a set of exemplars $\mathcal{X}_6^*$ that minimizes the self-representation cost function

Step 2: For each $x_i \in \mathcal{X}$, compute $c_i^*$ by solving the following optimization problem

$$ c_i^* = \arg \min_{c_i \in \mathbb{R}^d} \|c_i\|_1 + \frac{1}{2} \|x_i - \sum_{j \in \mathcal{X}_6} c_j x_j\|_2^2 $$

Step 3: Compute nearest neighbor affinity $A$ from $\{c_i^*\}^N_{i=1}$ and apply spectral clustering

Theorem 1: i) $\mathcal{X}_6^*$ contains at least dim($\mathcal{S}$) points from each subspace $\mathcal{S}$, ii) the affinity $A$ has no wrong connections. The result holds even if data $\mathcal{X}$ is class imbalanced

A Farthest First Search (FFS) Algorithm

- Since minimizing (1) is NP-hard in general, we propose to compute $\mathcal{X}_6(0)$ by iteratively selecting the worst represented point (see figure on right)
  1. Select $x_i \in \mathcal{X}$ at random and set $\mathcal{X}_6(0) \leftarrow \{x_i\}$
  2. for $i = 1, \ldots, k \leq 1$ do
  3. \hspace{0.5cm} $\mathcal{X}_6(i) = \mathcal{X}_6(i-1) \cup \arg \max_{x_i \in \mathcal{X}} \min_{c_i \in \mathbb{R}^d} \|c_i\|_1 + \frac{1}{2} \|x_i - \sum_{j \in \mathcal{X}_6(i-1)} c_j x_j\|_2^2$
  4. end for

Theorem 2: $\mathcal{X}_6(i)$ 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

- 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

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