Selective Sampling-based Scalable Sparse Subspace Clustering

Shin Matsushima — 東京大学 THE UNIVERSITY OF TOKYO

Maria Brbić — Stanford University

paper codes poster

OVERVIEW

Sparse Subspace Clustering (SSC)

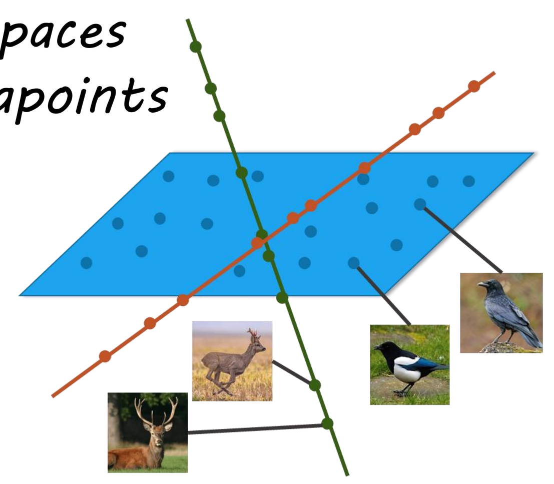
- ✓ high performance clustering for high dimensional data with strong theoretical guarantees
- ☹ quadratic complexity w.r.t. number of data points

Selective Sampling-based Scalable Sparse Subspace Clustering (S^5C)

- ✓ Theoretical Guarantee → Theoretical Scalability
- ✓ Low computational cost → Experimental Scalability
- ✓ Good clustering performance

SUBSPACE CLUSTERING

L subspaces
N datapoints

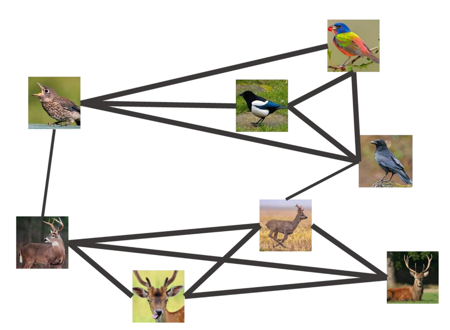


Assumption: high-dimensional data points lie in the union of low-dimensional subspaces.

Goal: identify subspaces and assign data points to the subspaces

-0.1 +0.3 +0.8 ≒

Representation of



Algorithm:

1. Representation learning
2. Derive an affinity graph
3. Spectral clustering

Application:
Clustering images
Clustering documents...

KEY IDEA

Represent data point as a linear combination among a small number of subsamples { , , }

$$\min_c \left\| \left(\cancel{c_1} + c_2 + \cancel{c_3} + c_5 + \cancel{c_6} + c_7 + \cancel{c_8} \right) - \right\|^2$$

→Perform selective sampling for subsamples so that all data points are represented well

S^5C ALGORITHM

1. Randomly sample a data point
2. Solve $\min_c \left\| \left(c_2 + c_5 + c_7 \right) - \right\|^2$ among current collection of subsamples { , , }
3. Find selective sample that helps representation of best
4. Add to current collection of subsamples { , , }
5. Repeat 1.-4. T times starting from { }
6. Solve $\min_c \left\| \left(c_2 + c_3 + c_5 + c_7 \right) - \cdot \right\|^2$ w.r.t. final subsamples for all data points

According to the absolute value of the gradient

→ O(NT) algorithm!

Iteratively select samples which seems to help better representations!

1. Initialize N-by-L matrix V as a random orthogonal matrix
2. V ← (2I – L) V
3. Orthogonalize V by QR decomposition
4. Repeat 2.-3. until convergence
5. Apply K-means for V

Graph Laplacian L has O(N) nonzero elements → O(N) algorithm!

RELATED WORK

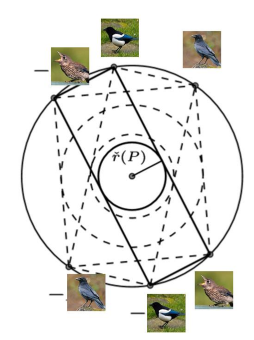
	SSC	SSC-OMP	SSC-ORGEN	SSSC	S^5C
Theoretical Scalability	✗	✗	✗	✗	✓
Experimental Scalability	✗	✓	✓	✓	✓

THEORETICAL SCALABILITY

S^5C Algorithm performs perfect clustering in O(N) running time in high probability under some generative model

Required number of subsample (≒T) is O(1) w.r.t. N

Semi Random model



Data points are uniformly randomly generated from unit ball in each subspace

Subspace detection property (SDP)

1. $c_{i'}$ for i-th data is nonzero only when i and i' shares the same subspace
2. c is not all zeros

-0.1 +0.3 +0.8 ≒

CLUSTERING PERFOEMANCE

Clustering error (%)

	Nystrom	AKK	SSC	SSC-OMP	SSC-ORGEN	SSSC	S^5C
Yale B	76.8	85.7	33.8	35.9	37.4	59.6	39.3
Hopkins 155	21.8	20.6	4.1	23.0	20.5	21.1.	14.6
COIL-100	54.5	53.1	42.5	57.9	89.7	67.8	45.9
Letter-rec	73.3	71.7	/	95.2	68.6	68.4	67.7
CIFAR-10	76.6	75.6	/	/	82.4	82.4	75.1
MNIST	45.7	44.6	/	/	28.7	48.7	40.4
Devanagari	73.5	72.8	/	/	58.6	84.9	67.2

SPARSE SUBSPACE CLUSTERING

Solve for all data points → O(N²) algorithm! → not scalable ☹

$$\min_c \left\| \left(c_1 + c_2 + c_3 + c_5 + c_6 + c_7 + c_8 \right) - \right\|^2$$

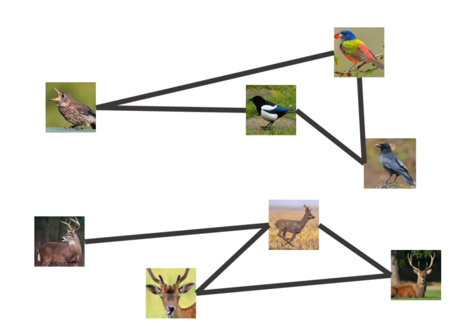
+ L1 regularizer on c

→ represent a data point as a linear combination among all data points

-0.1 +0.3 +0.8 ≒

→ only some data points in the same subspace remain

SDP property



→perfect clustering is theoretically proven

EXPERIMENTAL SCALABILITY