Sparse-Dense Subspace Clustering

Shuai Yang, Wenqi Zhu, Yuesheng Zhu

Institute of Big Data Technologies, Peking University, China

ABSTRACT

Subspace clustering refers to the problem of clustering high-dimensional data into a union of low-dimensional subspaces. Current subspace clustering approaches are usually based on a two-stage framework. In the first stage, an affinity matrix is generated from data. In the second one, spectral clustering is applied on the affinity matrix. However, the affinity matrix produced by two-stage methods cannot fully reveal the similarity between data points from the same subspace, resulting in inaccurate clustering. Besides, most approaches fail to solve large-scale clustering problems due to poor efficiency. In this paper, we first propose a new scalable sparse method called Iterative Maximum Correlation (IMC) to learn the affinity matrix from data. Then we develop Piecewise Correlation Estimation (PCE) to densify the intra-subspace similarity produced by IMC. Finally we extend our work into a Sparse-Dense Subspace Clustering (SDSC) framework with a dense stage to optimize the affinity matrix for two-stage methods. We show that IMC is efficient for large-scale tasks, and PCE ensures better performance for IMC. We show the universality of our SDSC framework for current two-stage methods as well. Experiments on benchmark data sets demonstrate the effectiveness of our approaches.

Iterative Maximum Correlation (IMC): A Better Solution for Sparse Coefficients Matrix Generation

Algorithm 1 Iterative Maximum Correlation (IMC)

Input: Data set X, IMC iteration number Γ.

Initialize coefficients matrix C as a $N \times N$ zero matrix, index of current data point $i = 1$.
while $i \leq N$ **do**
  Initialize current iteration $\gamma = 0$, residual $\boldsymbol{\psi}_0 = \boldsymbol{x}_i$.
  while $\gamma < \Gamma$ **do**
    $c_{ij^*} = \left|\rho_{\boldsymbol{\psi}_\gamma \boldsymbol{x}_{j^*}}\right|$, where $j^* = \arg\max_{j \neq i} \left|\rho_{\boldsymbol{\psi}_\gamma \boldsymbol{x}_j}\right|$, and $\rho_{\boldsymbol{\psi}_\gamma \boldsymbol{x}_j}$ is calculated by (3).
    Update residual $\boldsymbol{\psi}_{\gamma+1} = \boldsymbol{\psi}_\gamma - (\boldsymbol{\psi}_\gamma \cdot \boldsymbol{x}_j)\boldsymbol{x}_j$.
    $\gamma \leftarrow \gamma + 1$.
  end while
  $i \leftarrow i + 1$
end while

Output: The coefficients matrix C.

Piecewise Correlation Estimation (PCE): Densify the Coefficients Matrix

Algorithm 2 Piecewise Correlation Estimation (PCE)

Input: θ_1, θ_2, W, data set X.

Compute distance matrix $D = I - W$
for each pair of data points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ **do**
  for each intermediate point $\boldsymbol{x}_k \in X \backslash \{\boldsymbol{x}_i, \boldsymbol{x}_j\}$ **do**
    $$d^*_{ij} = \begin{cases} \min(d_{ik}, d_{kj}), & \textbf{if TUR condition 1)} \\ \frac{1}{2}(d_{ik} + d_{kj}), & \textbf{if TUR condition 2)} \\ \max(d_{ik}, d_{kj}), & \textbf{if TUR condition 3)} \\ d_{ij}, & \textbf{else} \end{cases}$$
  end for
end for
Compute the new affinity matrix $W^* = I - D^*$

Output: A new affinity matrix $W^* \in \mathbb{R}^{N \times N}$.

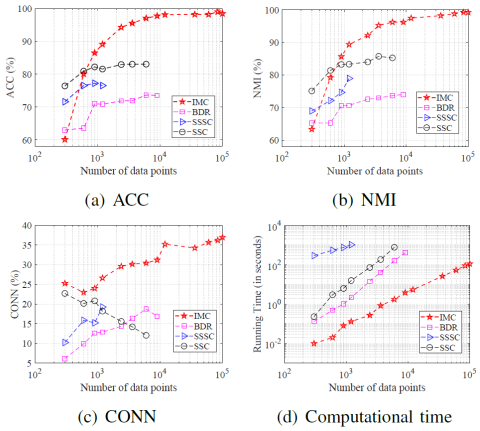
Sparse-Dense Subspace Clustering (SDSC): A Universal Framework

Algorithm 3 Sparse-Dense Subspace Clustering (SDSC)

Input: Data set X.

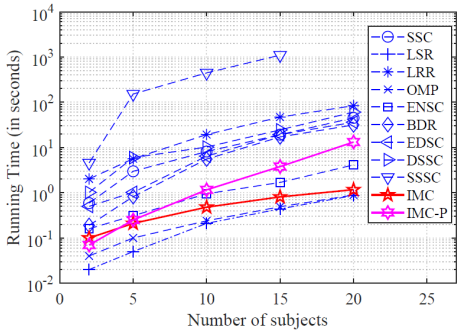
Compute a affinity matrix W from data by different data representation methods.
Optimize similarity in W to get W^* by a dense method.
Apply spectral clustering on W^*.

Output: Clustering results.

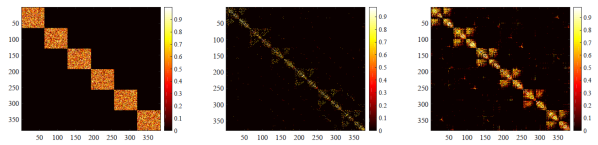


(a) ACC (b) NMI (c) CONN (d) Computational time

Performance on synthetic data



Computational time on the Extended Yale B



(a) The ground truth (b) W by IMC (c) Densified W^*

Data Visualization: the Densification by PCE