Wasserstein k-means with sparse simplex projection

Takumi Fukunaga (Waseda University, Japan) and Hiroyuki Kasai (Waseda University, Japan)

Problem of interest

• Focuses on Wasserstein k-means
• Necessary to solve optimal transport as the subproblem of this problem

\[ \gamma_k(\mu, \nu) = \min_{\Pi \in \Upsilon_{1n}} (T, C) = (T, C) \]

\[ \Upsilon_{1n} = \{ T \in R^{n \times n} : T_{1n} = a, T^T 1_n = b \} \]

• \( \mu, \nu \) are empirical probability distributions.
• C is the ground cost matrix.
• Takes high computational costs to solve it
• \( O(n^2 \log(n)) \) complexity
• Various applications
  - e.g., machine learning, color transfer.

Contributions

• Propose Sparse simplex projection Wasserstein k-means (SSPW k-means).
• Numerical evaluations demonstrate the effectiveness in two followings points
  - Reducing the complexities of Wasserstein Distance
  - Maintaining the clustering quality before sparsifying and shrinking

Clustering algorithm [1,2]

• One of popular algorithms is k-means method.
• Consists of two following steps
  - \( s_i = \arg \min_{j=1, \ldots, k} d(x_i, c_j) \), \( \forall i \in [q] \)
  - \( c_j = \text{barycenter}\{x_i | s_i = j\}, \forall j \in [k] \)
• Former assignment step causes high computational cost.
• Call it Wasserstein k-means when adopting Wasserstein distance and barycenter

Optimal Transport [3,4]

• Minimizes the total transport cost.
• The optimum solution gives the Wasserstein distance.
• Wasserstein barycenter is defined as

\[ g(\mu) = \frac{1}{n} \sum_{i=1}^{n} W_{ij}(\mu, \nu) \]

• From the formula and the domain, optimal transport is solved by linear programming (LP).
• Linear programming is difficult to solve because of the high computational complexities.

Simplex Projection

• Sparse simplex projection (GSHP) [5]

\[ \beta = \text{Proj}_T(\beta) = \left\{ \begin{array}{ll} \beta_T & \text{if } \beta_T > 0 \\ 0 & \text{otherwise} \end{array} \right. \]

• \( S \) is the subset of \( V = \{ 1, 2, \ldots, n \} \).
• \( \gamma \) extracts the elements of \( S \) in \( a \).
• \( \gamma = \sum_{i \in S} \gamma_i \)
• \( n^2 \) is the number of elements of \( \beta \).
• Computational complexities is \( O(n \min(\kappa, \log(n))) \)

Shrinking operations according to zero elements

• Vector shrinking operator

\[ \hat{\nu}_j = \text{shrink}(\nu_j) = \left( \frac{\nu_j}{\|\nu_j\|}\right)_{\nu_j \in R^{|S_{\text{zero}}|}} \]

• Vector shrinking operator removes zero elements from the projected sample \( \hat{\nu}_j \) and centroid \( \hat{c}_j \) and generates \( \nu_j \) and \( c_j \) respectively.

• Matrix shrinking operator

\[ \hat{C} = \text{Shrink}(C_{\text{bary}}) \]

\[ = C_{\text{bary}}(\hat{\nu}_j)_{\nu_j \in R^{|S_{\text{zero}}|}} \]

• Shrinks the elements of the ground cost matrix, of which correspond to the removed vectors.
• Produce no degradations because zero elements don’t have effect on transport matrix.

Sparse simplex projection Wasserstein k-means (SSPW k-means) [Alg.1]

• Three \( \gamma(t) \) control algorithms:

\[ \begin{align*}
\gamma_{\text{min}}(1 + \frac{1}{T_{\text{max}}}) \quad & (\text{INC}) \\
\gamma_{\text{max}} - \gamma_{\text{min}} \quad & (\text{DEC}) \\
\gamma_{\text{min}} \quad & (\text{FIX})
\end{align*} \]

• Denoted as ‘INC’ (increase), ‘DEC’ (decrease), and ‘FIX’ (fix).

Control parameter of sparse ratios

Figure 1: Example of shrinking operation.

Figure 2: Clustering performance results of 2-D histogram data (USPS dataset).

Figure 3: Left: Convergence performance with different projection data using DEC algorithm of \( \gamma_{\text{min}} = 0.5 \). Right: Convergence performance comparison of different algorithm of \( \gamma(t) \) of \( \gamma_{\text{min}} = 0.5 \).

References


Numerical evaluations

A. Clustering performance

B. Convergence performance

C. Comparison on different sparsity

Figure 4: Performance comparison on different ratios (USPS dataset).