Wasserstein $k$-means with sparse simplex projection

Takumi Fukunaga$^1$  Hiroyuki Kasai $^{2,3}$

$^1$Dept. of Computer Science and Engineering, School of Fundamental Science and Engineering, Waseda University, Japan

$^2$Dept. of Communications and Computer Engineering, School of Fundamental Science and Engineering, Waseda University, Japan

$^3$Dept. of Computer Science and Communications Engineering, Graduate School of Fundamental Science and Engineering, Waseda University, Japan

December 9, 2020
Introduction

Wasserstein $k$-means

- Propose
  - Improvement of computational cost of Wasserstein $k$-means
- Fast and efficient approaches
  - [Cuturi and Doucet, 2014, Bonneel et al., 2015, Anderes et al., 2016]
- Contribution of our methods
  - Sparsifying data on probability simplex and shrinking them by removing the zero elements
Introduction

Clustering Algorithm and its issue

- **$k$-means** [Lloyd, 1982]
  - High computational cost per iteration $\mathcal{O}(qk)$
  - Efficient approaches
    - [Arthur and Vassilvitskii, 2007, Kanungo et al., 2002]

- **Wasserstein $k$-means** [Ye et al., 2017]
  - Adopting Wasserstein distance and Wasserstein barycenter
  - High computational cost to calculate its distance
    $\mathcal{O}(n^3 \log n)$ [Cuturi, 2013, Rubner et al., 2000]

**assignment step**

$$s_i = \arg\min_{j=1,\ldots,k} d(x_i, c_j), \forall i \in [q]$$

**update step**

$$c_j = \text{mean}(\{x | s_i = j\}) \quad \text{or} \quad \text{barycenter}(\{x | s_i = j\}), \forall j \in [k]$$
Introduction

Optimal Transport

▶ Calculate the minimum transport cost [Peyre and Cuturi, 2019]
▶ When $n = m$, its cost is called Wasserstein distance of order $p$
▶ Using this distance, calculate Wasserstein barycenter [Benamou et al., 2015]

\[
T^* = \arg \min_{T \in \mathcal{U}_{mn}} \langle T, C \rangle \\
W_p(\mu, \nu) = \min_{T \in \mathcal{U}_{mn}} \langle T, C \rangle = \langle T^*, C \rangle \\
g(\mu) = \frac{1}{n} \sum_i W_p(\mu, \nu_i)
\]

▶ $\nu$ and $\mu$ of points
▶ $a$ and $b$ are in probability simplex
▶ $C$ is ground matrix
▶ Row and Column marginal constraint

\[
\mathcal{U}_{mn} = \{ T \in \mathbb{R}^{m \times n}_+ : T \mathbf{1}_n = a, T^T \mathbf{1}_m = b \}\]
Proposal

Motivation

- Reduce the size of data and centroid
- Adopt two following approaches

1. Sparsify datas
   - Make data sparser than the original ones
   - Maintain degradation of the clustering quality as small as possible

2. Shrink datas
   - No degradation
   - Key operator to reduce the computational complexities
Proposal

Basic idea - Sparse simplex projection

- Sparse simplex projection GSHP [Kyrillidis et al., 2013]

\[
\hat{\beta} = \text{Proj}^{(t)}(\beta) = \begin{cases} 
\hat{\beta}|_{S^*} = P_{\Delta_\kappa}(\beta|_{S^*}) \\
\hat{\beta}|_{(S^*)^c} = 0, 
\end{cases}
\]

- \( S \) is the subset of \( N = \{1, \ldots, n\} \)
- \( a|_S \) extracts the elements of \( S \) in \( a \)
- \( (P_{\Delta_\kappa}(\beta|_{S^*}))_v = [(\beta|_{S^*})_v + \tau]^+ \), \( \tau := \frac{1}{\kappa}(1 + \sum_{S^*}|\beta|_{S^*}) \)
- \( S^* = \text{supp}(P_{\Delta_\kappa}) \)
  - \( \text{supp}(a) = \{i : a_i \neq 0\} \)
Proposal

Basic idea - Shrinking datas

- The zero elements don’t have effect on transport matrix
  - Removing the zero elements
  - Define shrinking operator to vector and matrix

\[
\tilde{v}_i = \text{shrink}(\hat{v}_i) = (\hat{v}_i)|_{S_{\text{samp}}} \in \mathbb{R}^{|S_{\text{samp}}|}
\]
\[
\tilde{c}_i = \text{shrink}(\hat{c}_i) = (\hat{c}_i)|_{S_{\text{cent}}} \in \mathbb{R}^{|S_{\text{cent}}|}
\]
\[
\tilde{C} = \text{Shrink}(C_{\nu c}) = C_{\text{supp}(\hat{v}_i), \text{supp}(\hat{c}_i)} \in \mathbb{R}^{|S_{\text{samp}}| \times |S_{\text{cent}}|}
\]

\[
\text{supp}(a) = \{2,3,5,6\}
\]

\[
\begin{pmatrix}
0 & 1 & 4 & 9 & 16 & 25 & 36 \\
1 & 0 & 1 & 4 & 9 & 16 & 25 \\
4 & 1 & 0 & 1 & 4 & 9 & 16 \\
9 & 4 & 1 & 0 & 1 & 4 & 9 \\
16 & 9 & 4 & 1 & 0 & 1 & 4 \\
25 & 16 & 9 & 4 & 1 & 0 & 1 \\
36 & 25 & 16 & 9 & 4 & 1 & 0 \\
\end{pmatrix}
\]

\[
\begin{pmatrix}
1 & 4 & 16 & 25 \\
1 & 0 & 4 & 9 \\
4 & 1 & 1 & 4 \\
9 & 4 & 0 & 1 \\
\end{pmatrix}
\]
Proposal

Basic procedure

1. Update sparsity ratio $\gamma(t)$
2. Project $\nu_i$ into $\hat{\nu}_i$ and shrink $\hat{\nu}_i$ into $\tilde{\nu}_i$
3. Project $c_j$ into $\hat{c}_j$ and shrink $\hat{c}_j$ into $\tilde{c}_i$
4. Shrink ground cost matrix $C$ into $\tilde{C}$
5. Find closest centroids and update centroids
6. Unless cluster centroids stop changing, repeat step 1

▶ Control parameter of sparse ratio $\gamma(t)$

$$\gamma(t) := \begin{cases} 
\gamma_{\text{min}} & \text{(FIX)} \\
1 - \frac{(1 - \gamma_{\text{min}})}{T_{\text{max}}} t & \text{(DEC)} \\
\gamma_{\text{min}} + \frac{(1 - \gamma_{\text{min}})}{T_{\text{max}}} t & \text{(INC)},
\end{cases}$$

▶ $T_{\text{max}}$ is maximum iteration of $k$-means
Experiment

Settings

- Use of Algorithm
  - Wasserstein barycenter [Cuturi and Doucet, 2014]
  - \(k\)-means with litekmeans
  - linprog of Mosek to solve LP [Andersen et al., 2000]

- Datasets
  - COIL-100 [Nene et al., 1996]
  - the USPS handwritten dataset
Experiment

2-D histogram evaluation

Figure: Performance results of 2-D histogram data on the USPS dataset.
**Experiment**

**Convergence Performance**

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

Comparison on different sparsity

Figure: Performance comparison on different ratios on the USPS dataset.
Experiment

Conclusion

- We propose a faster Wasserstein $k$-means algorithm
- Our experiments demonstrate the effectiveness of our method
  - Reducing the computational complexity of Wasserstein distance
  - Keeping accuracy before sparsifying and shrinking
References


References II


References III


References IV

Thank you for listening.