

Problem of interest

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

$$\mathcal{W}_p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \min_{\mathbf{T} \in \mathcal{U}_{mn}} \langle \mathbf{T}, \mathbf{C} \rangle = \langle \mathbf{T}, \mathbf{C} \rangle$$

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

- $\boldsymbol{\mu}, \boldsymbol{\nu}$ are empirical probability distributions.
- \mathbf{C} is the ground cost matrix.
- Takes high computational costs to solve it
 - $\mathcal{O}(n^3 \log(n))$
- 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

$$s_i = \arg \min_{j=1, \dots, k} d(\mathbf{x}_i, \mathbf{c}_j), \forall i \in [q]$$

$$\mathbf{c}_j = \text{barycenter}(\{\mathbf{x} | s_i = j\}), \forall j \in [k]$$

- 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 costs.
- The optimum solution gives the Wasserstein distance.
- Wasserstein barycenter is defined as

$$g(\boldsymbol{\mu}) = \frac{1}{n} \sum_i W_p(\boldsymbol{\mu}, \boldsymbol{\nu}_i)$$

- 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.

Sparse Simplex Projection

- Sparse simplex projection (GSHP) [5]

$$\hat{\boldsymbol{\beta}} = \text{Proj}^{\gamma(t)}(\boldsymbol{\beta}) = \begin{cases} \hat{\boldsymbol{\beta}}_{|\mathcal{S}^*} & = \mathcal{P}_{\Delta_{\kappa}}(\boldsymbol{\beta}_{|\mathcal{S}^*}) \\ \hat{\boldsymbol{\beta}}_{|(\mathcal{S}^*)^c} & = 0 \end{cases}$$

- \mathcal{S} is the subset of $\mathcal{N} = \{1, 2, \dots, n\}$.
- $\mathbf{a}_{|\mathcal{S}}$ extracts the elements of \mathcal{S} in \mathbf{a} .
- $\kappa = \lfloor n \cdot \gamma(t) \rfloor$.
- $\mathcal{S}^* = \text{supp}(\mathcal{P}_{\kappa}(\boldsymbol{\beta}))$.
- The v -th element of $\mathcal{P}_{\Delta_{\kappa}}(\boldsymbol{\beta}_{|\mathcal{S}^*})$ is defined as

$$(\mathcal{P}_{\Delta_{\kappa}}(\boldsymbol{\beta}_{|\mathcal{S}^*}))_v = [(\boldsymbol{\beta}_{|\mathcal{S}^*})_v + \tau]_+$$

where τ is

$$\tau := \frac{1}{\kappa} (1 + \sum_{|\mathcal{S}^*|} \boldsymbol{\beta}_{|\mathcal{S}^*}).$$

- Computational complexities is $\mathcal{O}(n \min(\kappa, \log(n)))$

Shrinking operations according to zero elements

- **Vector shrinking** operator

$$\tilde{\boldsymbol{\nu}}_i = \text{shrink}(\hat{\boldsymbol{\nu}}_i) = (\hat{\boldsymbol{\nu}}_i)_{|\mathcal{S}_{\text{samp}}} \in \mathbb{R}^{|\mathcal{S}_{\text{samp}}|}$$

$$\tilde{\mathbf{c}}_i = \text{shrink}(\hat{\mathbf{c}}_i) = (\hat{\mathbf{c}}_i)_{|\mathcal{S}_{\text{cent}}} \in \mathbb{R}^{|\mathcal{S}_{\text{cent}}|}$$

- Vector shrinking operator removes zero elements from the projected sample $\hat{\boldsymbol{\nu}}_i$ and centroid $\hat{\mathbf{c}}_i$ and generates $\tilde{\boldsymbol{\nu}}_i$ and $\tilde{\mathbf{c}}_i$ respectively.

- **Matrix shrinking** operator

$$\tilde{\mathbf{C}} = \text{Shrink}(\mathbf{C} \boldsymbol{\nu} \mathbf{c})$$

$$= \mathbf{C}_{\text{supp}(\tilde{\boldsymbol{\nu}}_i), \text{supp}(\tilde{\mathbf{c}}_i)} \in \mathbb{R}^{|\mathcal{S}_{\text{samp}}| \times |\mathcal{S}_{\text{cent}}|}.$$

- Shrink 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.

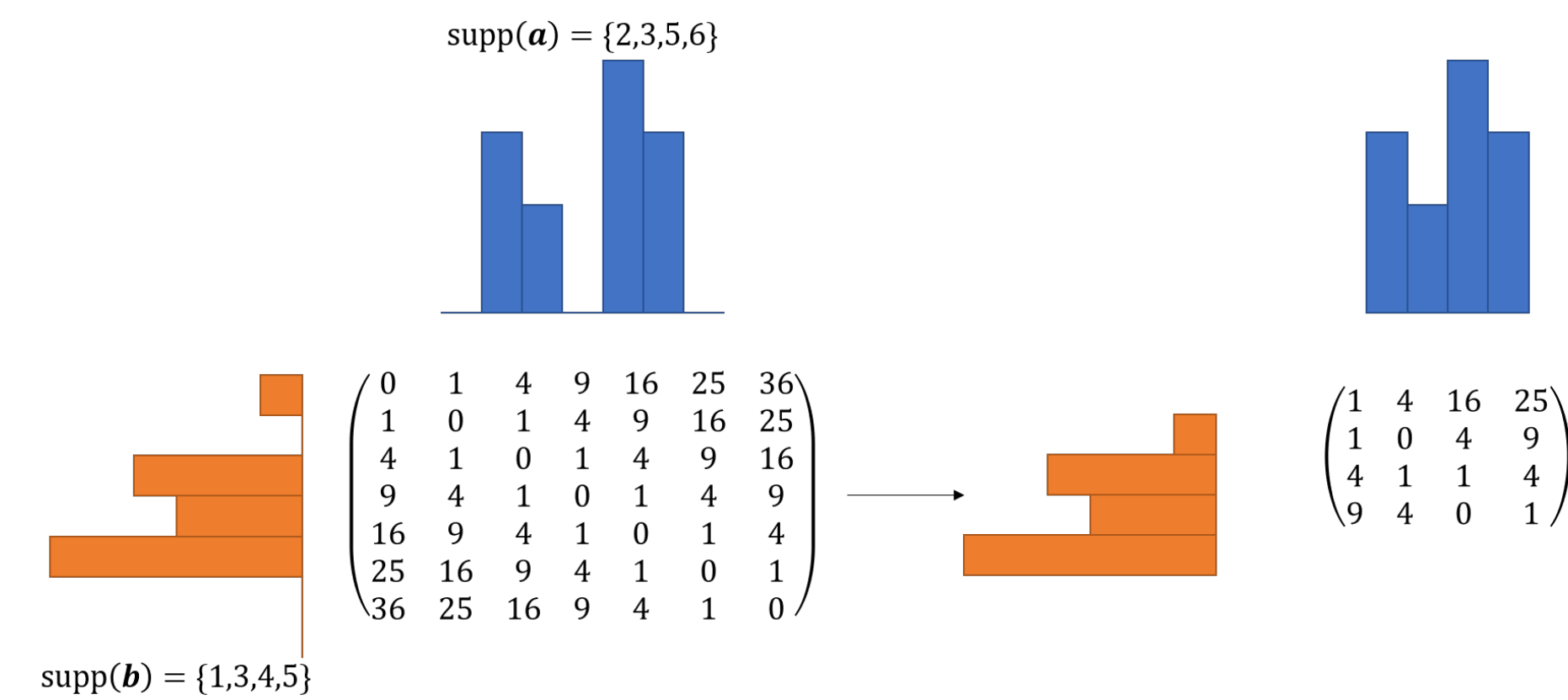


Figure 1: Example of shrinking operation.

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

Require: data $\{\boldsymbol{\nu}_1, \dots, \boldsymbol{\nu}_q\}$, cluster number $k \in \mathbb{N}$, ground cost matrix $\mathbf{C} \in \mathbb{R}^{n \times n}$, maximum iteration number T_{\max} , γ_{\min} .

- 1: Initialize centroids $\{\tilde{\mathbf{c}}_1, \dots, \tilde{\mathbf{c}}_k\}$, set $t = 1$.
 - 2: **repeat**
 - 3: Update sparsity ratio $\gamma(t)$.
 - 4: Project $\boldsymbol{\nu}_i$ to $\hat{\boldsymbol{\nu}}_i$ on sparse simplex Δ_p :
 $\hat{\boldsymbol{\nu}}_i = \text{Proj}^{\gamma(t)}(\boldsymbol{\nu}_i) \forall i \in [q]$.
 - 5: Shrink $\hat{\boldsymbol{\nu}}_i$ into $\tilde{\boldsymbol{\nu}}_i$: $\tilde{\boldsymbol{\nu}}_i = \text{shrink}(\hat{\boldsymbol{\nu}}_i)$.
 - 6: Project \mathbf{c}_j into $\hat{\mathbf{c}}_j$ on sparse simplex Δ_p :
 $\hat{\mathbf{c}}_j = \text{Proj}^{\gamma}(\mathbf{c}_j) \forall j \in [k]$.
 - 7: Shrink $\hat{\mathbf{c}}_j$ into $\tilde{\mathbf{c}}_j$: $\tilde{\mathbf{c}}_j = \text{shrink}(\hat{\mathbf{c}}_j)$.
 - 8: Shrink ground cost matrix \mathbf{C} into $\tilde{\mathbf{C}}$: $\tilde{\mathbf{C}} = \text{Shrink}(\mathbf{C})$.
 - 9: Find closest centroids (assignment step):
 $s_i = \arg \min_{j=1, \dots, k} W_p(\tilde{\boldsymbol{\nu}}_i, \tilde{\mathbf{c}}_j), \forall i \in [q]$.
 - 10: Update centroids (update step):
 $\mathbf{c}_j = \text{barycenter}(\{\boldsymbol{\nu} | s_i = j\}), \forall j \in [k]$.
 - 11: **until** cluster centroids stop changing. StateUpdate the iteration number t as $t = t + 1$.
- Ensure:** cluster centers $\{\mathbf{c}_1, \dots, \mathbf{c}_k\}$.

Control parameter of sparse ratios

- Three $\gamma(t)$ control algorithms:

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

- Denoted as 'FIX' (fixed), 'DEC' (decrease), and 'INC' (increase).
- $\gamma_{\min} \in \mathbb{R}$ is the minimum value.
- $T_{\max} \in \mathbb{N}$ is the maximum number of the iterations.

References

- [1] Lloyd, S., Least squares quantization in pcm, IEEE Trans. Inf. Theory, 28(2):129-137, 1982.
- [2] Ye, Y., Wu, P., Wang, J. Z., and Li, J., Fast discrete distribution clustering using Wasserstein barycenter with sparse support, IEEE Trans. Signal Process, 65(9):2317-2332, 2017.
- [3] Peyre, G. and Cuturi, M., Computational optimal transport, Foundations and Trends in Machine Learning, 11(5-6):355-607, 2019.
- [4] Benamou, J. D., Carlier, G., Cuturi, M., Nenna, L., and Peyr'e, G., Iterative bregman projections for regularized transportation problems, SIAM Journal on Scientific Computing, 37(2):1111-A1138, 2015.
- [5] Kyrillidis, A., Becker, S., Cevher, V., and Koch, C., Sparse projections onto the simplex, In ICML, 2013.

Numerical evaluations

A. Clustering performance



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

B. Convergence performance

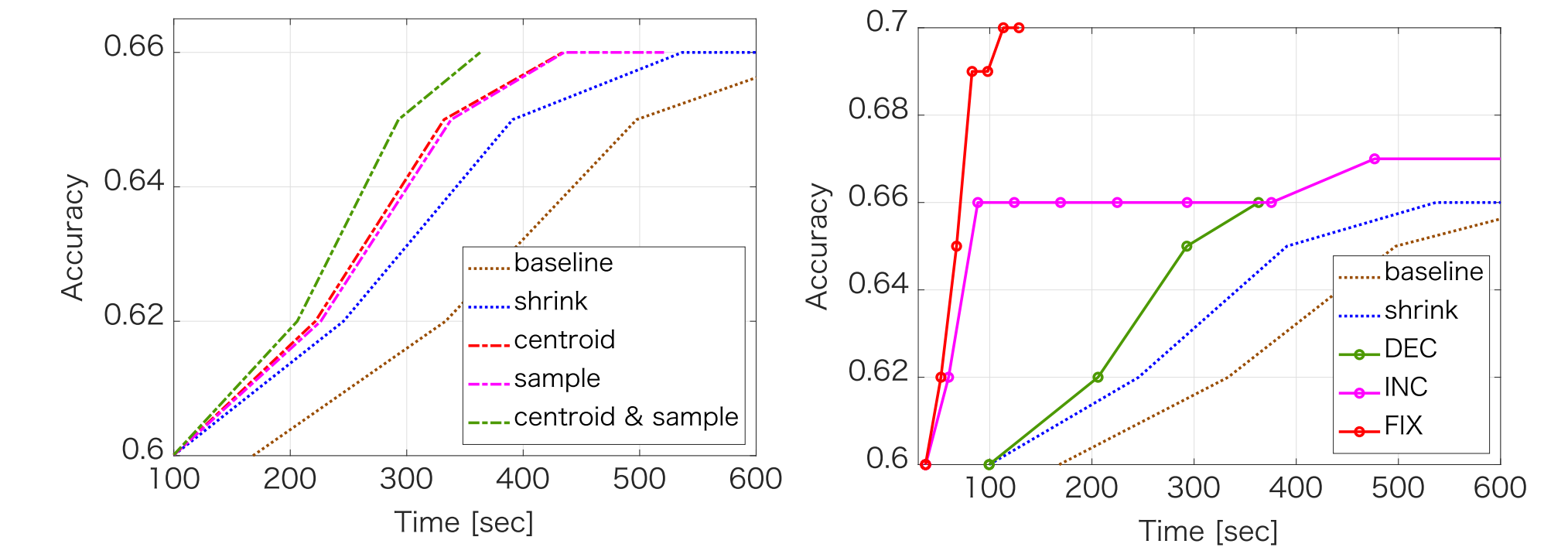


Figure 3: Left: Convergence performance with different projection data using DEC algorithm of $\gamma_{\min} = 0.5$. Right: Convergence performance comparison of different algorithms of $\gamma(t)$ of $\gamma_{\min} = 0.5$.

C. Comparison on different sparsity

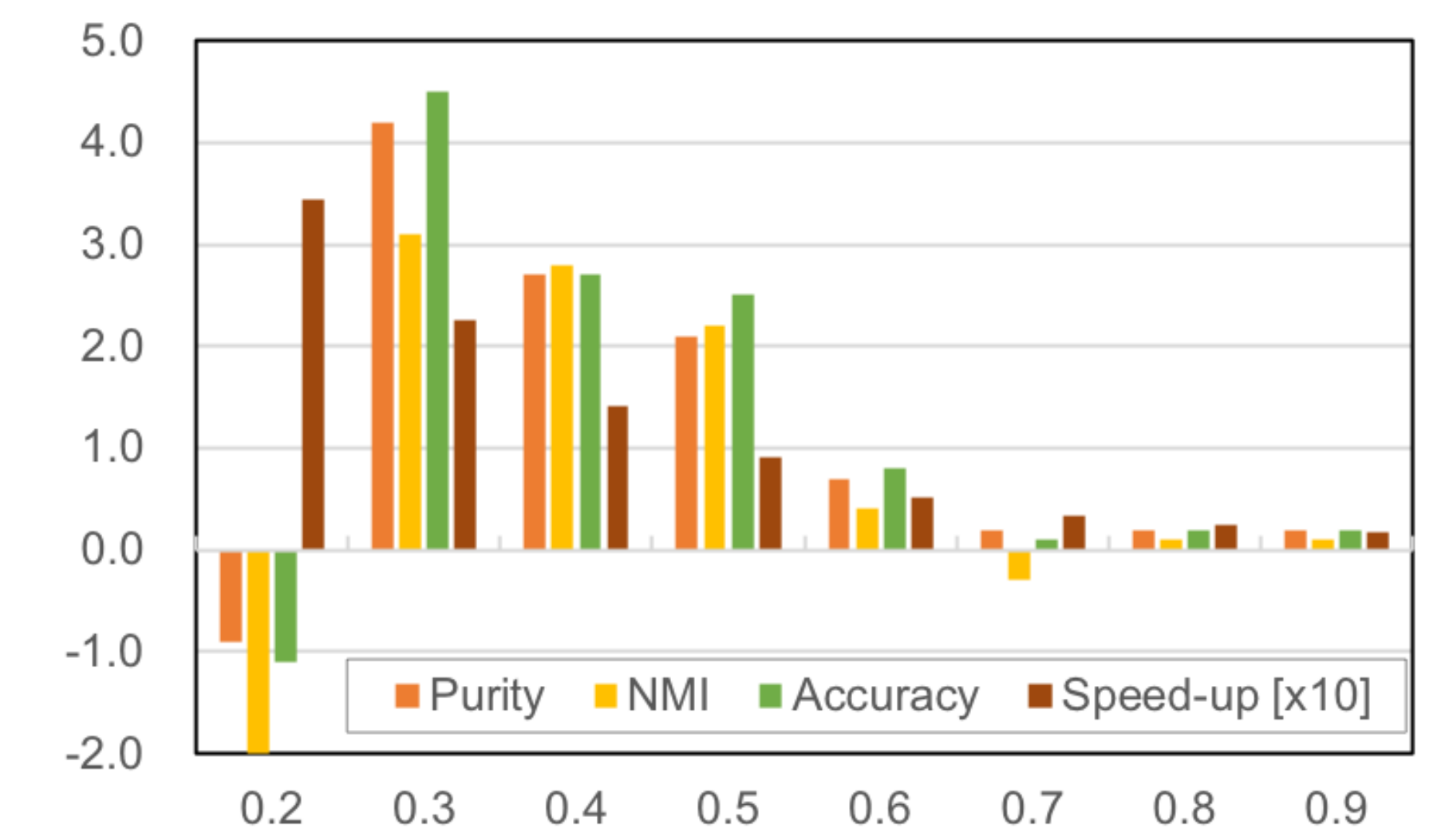


Figure 4: Performance comparison on different ratios (USPS dataset).