Core-sets for Fair and Diverse Data Summarization

Sepideh Mahabadi and Stojan Trajanovski

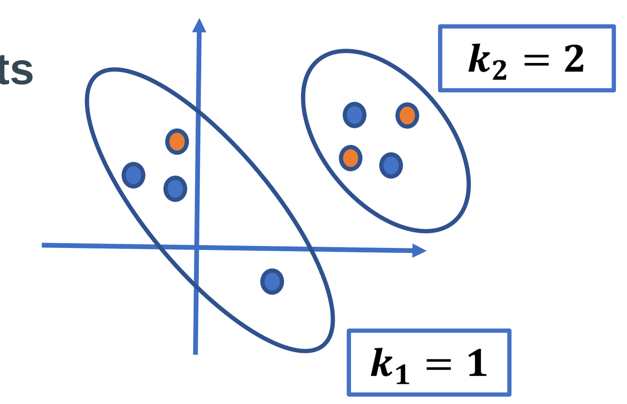
Constrained / Fair Diversity Maximization

Input:
sets of vectors $P_1, \cdots, P_m$, $P = \cup_i P_i$
and $k_1, \cdots, k_m \le d$, $k = \sum_i k_i$

Goal: pick k_i points $S_i \subset P_i$ s.t. the diversity of the picked points $S = \cup_i S_i$ is maximized

Diversity measures for a subset S of points

- MIN-PAIRWISE DIST $= \min_{p,q \in S} dist(p,q)$
- SUM-PAIRWISE DIST $= \sum_{p,q \in S} dist(p,q)$
- SUM-NN DIST $= \sum_{p \in S} \min_{q \in S \setminus \{p\}} dist(p,q)$



Applications in Summarization

Modeling recency in user's feed generation

- Each message has a timestamp being posted
- Show a "diverse" summary to the user
- Goal: show more recent messages and less of old messages
- Divide the messages in a month into four groups based on the week they have been posted
- Set k_i to be higher for more recent weeks

Recommendation System

- Different Movie Genres

Core-sets for Diversity Maximization

Input:
A point set P_i along with k_i
Goal: a summarization algorithm $\mathcal{A}$

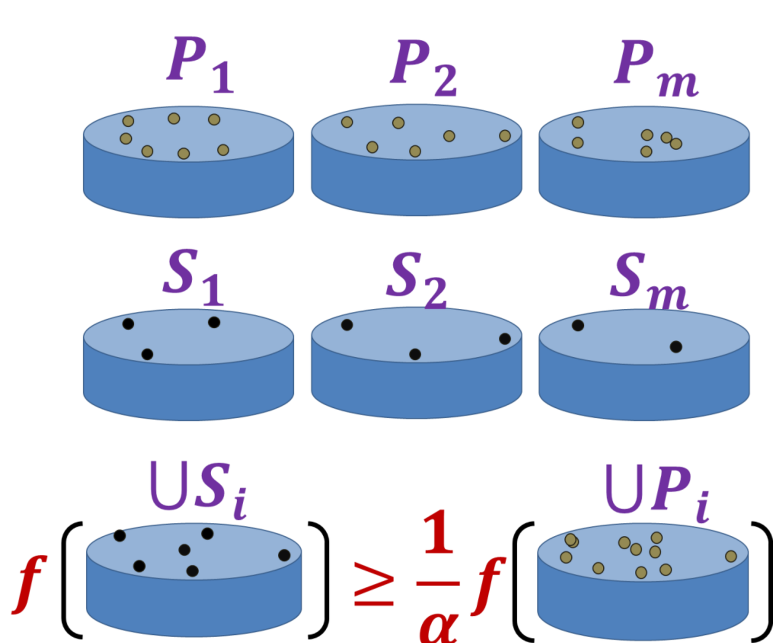
- Processes each P_i independently
- produces a small summary $S_i = \mathcal{A}(P_i) \subseteq P_i$

Main Property
Fair diversity of the data is approximately preserved, i.e.,

$$div_{k_1,k_2,\cdots,k_m}(S) \ge \frac{1}{\alpha} div_{k_1,k_2,\cdots,k_m}(P)$$

where S is the union of all core-sets $S = \cup_i S_i$ and

$$div_{k_1,k_2,\cdots,k_m}(P) = \max_{T_1 \subseteq P_1,\ldots,T_k \subseteq P_k, |T_i| = k_i} div\left(\bigcup_i T_i\right)$$



Theoretical results

Table 1.

Diversity Notion	FDM	Core-set setting: Approx.	Core-set setting: Core-set size	Core-set setting: Reference
MIN-PAIRWISE DIST	$\theta(m)$ [MMM20, AMMM22]	$O(1)$	$O(k)$ per group	[MMM20]
SUM-PAIRWISE DIST	$\theta(1)$ [AMT13]	$(1+\epsilon)$	depends on n or aspect ratio	[CPP18]
		$O(1)$	$O(k_i^2)$ per group	[This work]
SUM-NN DIST	$\theta(1)$ [BGMS16]	$O(m \cdot \log k)$	$O(k^2)$ per group	[This work]

Algorithm 1 Core-set Construction Algorithm for SUM-PAIRWISE

Input a point set P_i, together with parameters k_i and k (where $k = k_1 + \cdots + k_m$)
Output a subset $S_i \subseteq P_i$

```
1: S_i = {p_1, ..., p_{k_i}} ← GMM(P_i, k_i)
2: T ← ∅
3: for p ∈ S_i do
4:     for j = 1 to k_i do
5:         T ← T ∪ any point p_j ∈ P_i \ T s.t. argmin_{q∈S_i} dist(p_j, q) = p.
6:     end for
7: end for
8: S_i ← S_i ∪ T
9: return S_i
```

Algorithm 2 Core-set Construction Algorithm for SUM-NN

Input a point set P_i, together with parameters k_i and k (where $k = k_1 + \cdots + k_m$)
Output a subset $S_i \subseteq P_i$

```
1: S_i ← ∅
2: for j = 1 to k do
3:     G_i = {p_1, ..., p_{k+1}} ← GMM(P_i, k+1)
4:     S_i ← S_i ∪ G_i
5:     P_i ← P_i \ G_i
6: end for
7: return S_i
```

Experiments

Our experiments show the effectiveness of our core-set approach.

- [**need for FDM**] We demonstrate why we need to resort to FDM as DM outcome does not provide the desired fairness (Figure 1);
- [**price of fairness (balancedness)**] Applying FDM, we have a small loss of diversity while we achieve the desired fairness (Table 2);
- [**effectiveness of our core-sets**] We achieve a 100x speed-up while losing the diversity by only a few percent (Table 3) when applying FDM to the union of core-sets vs. FDM on the full data.

Experimental results

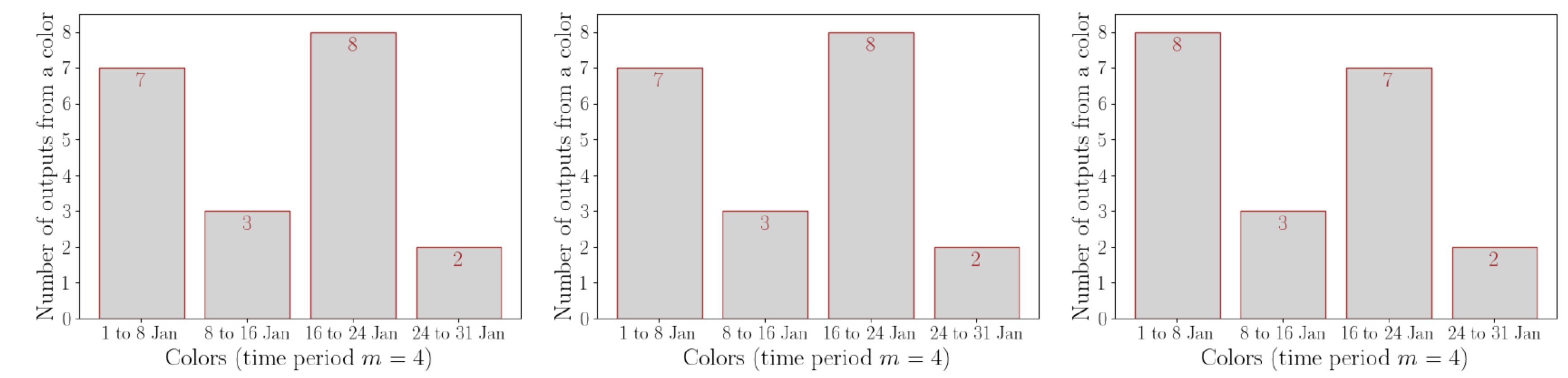


Figure 1. DM algorithm outcomes with equidistant time periods as colors ($m = 4$) with $k = 20$.

Table 2. The loss of diversity (% Div. loss) between DM vs. FDM for the Reddit dataset.

DM vs. FDM		SUM-PAIRWISE	SUM-NN	MIN-PAIRWISE
colors k_i	$\sum k_i$	% Div. loss	% Div. loss	% Div. loss
[2, 2, 2, 2]	8	1.22%	9.66%	51.57%
[3, 3, 3, 3]	12	0.98%	14.27%	49.99%
[4, 4, 4, 4]	16	0.50%	13.72%	48.78%
[5, 5, 5, 5]	20	0.47%	18.96%	48.05%
[6, 6, 6, 6]	24	0.19%	9.48%	47.20%
[2, 4, 6, 8]	20	0.42%	15.40%	48.05%
[3, 6, 9, 12]	30	0.29%	13.29%	46.34%
[4, 8, 12, 16]	40	0.25%	1.98%	45.52%
[5, 10, 15, 20]	50	0.16%	9.62%	44.48%
[6, 12, 18, 24]	60	0.12%	3.98%	43.60%

Table 3. The *loss* of diversity (% Div. loss), and the running time *gains* (x times faster) of the FDM when applied to the union of core-sets compared to FDM applied to the full data.

FDM full data vs. core-sets		SUM-PAIRWISE		SUM-NN		MIN-PAIRWISE	
colors k_i	$\sum k_i$	% Div. loss	Time gain (×)	% Div. loss	Time gain (×)	% Div. loss	Time gain (×)
[2, 2, 2, 2]	8	1.35%	196.24	2.22%	1 769.70	0.00%	208.64
[3, 3, 3, 3]	12	0.67%	333.13	0.29%	888.55	0.00%	152.48
[4, 4, 4, 4]	16	1.21%	539.69	−1.59%	474.26	0.00%	122.29
[5, 5, 5, 5]	20	1.17%	432.68	−0.44%	294.23	0.00%	89.08
[6, 6, 6, 6]	24	0.94%	130.87	−3.03%	183.28	0.00%	63.69
[2, 4, 6, 8]	20	1.50%	845.98	−1.80%	285.68	0.00%	91.44
[3, 6, 9, 12]	30	1.06%	134.76	2.27%	110.36	0.00%	53.05
[4, 8, 12, 16]	40	1.02%	182.06	−0.88%	57.88	0.00%	36.51
[5, 10, 15, 20]	50	1.16%	194.36	0.71%	34.90	0.00%	26.97
[6, 12, 18, 24]	60	1.27%	172.25	−0.49%	23.71	0.00%	20.53

References

[MMM20] Moumoulidou *et al.*, Diverse Data Selection under Fairness Constraints. arXiv, 2020.
[AMMM22] Addanki *et al.*, Improved Approximation and Scalability for Fair Max-Min Diversification. arXiv, 2022.
[AMT13] Abbassi *et al.*, Diversity Maximization Under Matroid Constraints. In KDD'13.
[BGMS16] Bhaskara *et al.*, Linear Relaxations for Finding Diverse Elements in Metric Spaces. In NIPS'16.
[CPP18] Ceccarello *et al.*,.Fast Coreset-based Diversity Maximization under Matroid Constraints. In WSDM'18.