Core-sets for Fair and Diverse Data Summarization
Sepideh Mahabadi and Stojan Trajanovski

Constrained / Fair Diversity Maximization

Input:
- sets of vectors $P_1, \ldots, P_m$, $P = U_1 P_1$ and $k_1, \ldots, k_m \leq d$, $k = \sum_i k_i$
- Goal: pick $k_i$ points $S_i \subseteq P_i$ s.t. the diversity of the picked points $S = \cup_i S_i$ is maximized

Diversity measures for a subset $S$ of points
- MIN-PARWISE DIST = $\min_{\prod_i \text{dist}(p_i, q)}$
- SUM-PARWISE DIST = $\sum_{\prod_i \text{dist}(p_i, q)}$
- SUM-NN DIST = $\sum_{p \in S \cap \text{dist}(p_i, q)}$

Table 1. Theoretical results.

<table>
<thead>
<tr>
<th>Diversity Notion</th>
<th>FDM</th>
<th>Core-set setting</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIN-PARWISE DIST</td>
<td>$\theta(m)$</td>
<td>$\theta(k)$ per group</td>
<td>[MM20]</td>
</tr>
<tr>
<td>SUM-PARWISE DIST</td>
<td>$\theta(1)$</td>
<td>$\theta(k)$ per group</td>
<td>[CPP18]</td>
</tr>
<tr>
<td>SUM-NN DIST</td>
<td>$\theta(1)$</td>
<td>$\theta(k^2)$ per group</td>
<td>[This work]</td>
</tr>
</tbody>
</table>

Algorithm 1 Core-set Construction Algorithm for SUM-PARWISE
Input: a point set $P$, together with parameters $k_i$ and $k$ (where $k = k_1 + \cdots + k_m$)
Output: a subset $S \subseteq P$

1. $S = \{p_1, \ldots, p_k\} \leftarrow \text{GMM}(P, k)$
2. $T = \emptyset$
3. for $P \in S$ do
4. for $j = 1$ to $k_i$ do
5. $T = T \cup$ any point $p_j \in P \setminus T$ s.t. argmin$_{i \in S \setminus T}$ dist($p_j, q_i$) = $p$
6. end for
7. end for
8. $S = S \cup T$
9. return $S$

Algorithm 2 Core-set Construction Algorithm for SUM-NN
Input: a point set $P$, together with parameters $k_i$ and $k$ (where $k = k_1 + \cdots + k_m$)
Output: a subset $S \subseteq P$

1. $S = \emptyset$
2. for $j = 1$ to $k_i$ do
3. $G_i = \{p_1, \ldots, p_{k+1}\}$ \leftarrow $\text{GMM}(P, k + 1)$
4. $S = S \cup G_i$
5. $P_i = P \setminus G_i$
6. end for
7. return $S$

Applications in Summarization

Modelling 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 frequent 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 \$A$
- Produces each $P_i$ independently
- Produces a small summary $S_i = \mathcal{A}(P_i) \subseteq P_i$

Fair diversity of the data is approximated preserves, i.e.,
$\max_{S_1, \ldots, S_m} \frac{\sum_{i=1}^m \text{dist}(S_i)}{\sum_{i=1}^m \text{dist}(P)} \geq \sum_{i=1}^m \text{dist}(S_i)$

$\max_{S_1, \ldots, S_m} \frac{\sum_{i=1}^m \text{dist}(S_i)}{\sum_{i=1}^m \text{dist}(P)} \leq \frac{1}{\alpha} \text{dist}(S)$

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 (balance)] 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.

References

[MM2022] Addanki et al., Improved Approximation and Scalability for Fair Max-Min

[BGMS16] Bhaskara et al., Linear Relaxations for Finding Diverse Elements in Metric Spaces.

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

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

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