Clustering to minimize the maximum intracluster distance

Teofilio F. Gonzalez, UC Santa Barbara

TCS 38. 1985
The Problem

• Input: A weighted undirected graph \( G=(V,E,W) \), integer \( k \)
• Goal: partition \( V \) into \( B_1, \ldots, B_k \) s.t.
  \[ \max_k \{ w(ij) \mid i,j \in B_k \} \text{ is minimum} \]

NP-hard even if the weights are distances betw points in a 2D plane (Gonzalez 85)

Will show: polynomial 2-approximation if distances satisfy the triangle inequality.
FPF* Algorithm

• Pick an arbitrary point, call it the head of cluster 1 (head₁)
• Assign all points to cluster 1
• Iteration i:
  - Pick a furthest point from any current head,
  - Designate it headᵢ,
  - Move to cluster i every point that is closer to headᵢ than to its current head.
• O(nk) time

*Furthest point first
Thm: Alg guarantees a 2-approximation

- Let $h_i(x)$ - head of pt. $x$ after iteration $i$
- $x^*$ - furthest point from its head when alg terminates $\Delta := d(x^*, h^k(x^*))$
- For every pt $x$, iter $i$, $d(x, h_i(x)) \geq d(x, h_{i+1}(x))$
- $\rightarrow$ For every $i$, $d(x^*, h_i(x^*)) \geq \Delta$
- $x^*$ was never chosen as head $\rightarrow$ all heads are at dist $\geq \Delta$ from each other
- $\rightarrow$ The $k$ final heads and $x^*$ are all at dist $\geq \Delta$ from each other $\rightarrow$ OPT $\geq \Delta$
- Triangle inequality implies that the alg has soln value $\leq 2\Delta$
Is better approx possible?

• Thm (Gonzalez, same paper):
  2- \( \varepsilon \) approximation is NP-hard for all \( \varepsilon > 0 \) (actually, even for points in 3D)
K-Boost

Geraci, Leoncini, Montangero, Pellegrini, Renda

JCB 2009
Overview

• Use Gonzalez FPF idea
• Determine the no. of clusters by a stability-based method
• Use boosting to improve robustness of k estimate
• Fast clustering: $O(k|M|)$ for input matrix $M$
From correlation to distance

• Want to use distance measure.
• Pearson correlation is not appropriate

\[ P(p_i, p_j) = \frac{\sum_{t=1}^{m} (p_i, t - \mu_i)(p_j, t - \mu_j)}{\sqrt{\left(\sum_{t=1}^{m} (p_i, t - \mu_i)^2\right) \left(\sum_{t=1}^{m} (p_j, t - \mu_j)^2\right)}} \]

• Soln: Take \( d(p_i, p_j) = \sqrt{1 - P(p_i, p_j)} \)
• Reason: \( q_i := (p_i - \mu) / \sigma_i \) then \( ||q_i|| = 1 \),
  \[ P(p_i, p_j) = q_i * q_j \]
  \[ \Rightarrow ||q_i - q_j||^2 = q_i * q_i + q_j * q_j - 2q_i * q_j = 2(1 - q_i * q_j) \]
Basic FPF Algorithm

• Input: points \( N=\{p_i\} \)
• In iteration \( i \), set of centers \( C_i \)
• Start with a random point as \( C_1 \)
• Repeat till \( k \):
  - Select a point furthest from any current center, set is as a new center
• Complexity: need to compute only distances to new center \( \Rightarrow O(kN) \) or \( O(kmn) \) for \( m\times n \) GE matrix
• Practical speedup: order pts in decreasing dist from each center, stop searching for points \( p \) to move from \( c_j \) to \( c_i \) when \( d(p,c_j)<.5d(c_i,c_j) \)
Using prediction strength to determine $k$

Tibshirani, Walther, Hastie 05

- Given two clustering solutions $V=\{V_1,..V_k\}$ and $W=\{W_1...W_k\}$ obtained independently on same gene set:
- Def: $P_i(k) =$ fraction of mates in cluster $V_i$ that are also mates in $W$

\[
P_i(k) = \frac{\left| \{(x,y) | x,y \in V_i, x<y \text{ and } \exists z \text{ s.t. } x,y \in W_z\} \right| \ast 2}{|V_i| \ast (|V_i|-1)}
\]

Prediction strength : $PS(k) := \min_i P_i(k)$
Using prediction strength to determine $k$

- Sample $S \subset N; T := N - S$
- For $t = 1, 2, \ldots$
  - Cluster $S, T$ into $t$ clusters, get partitions $U = \{U_1, \ldots, U_k\}, V = \{V_1, \ldots, V_k\}$ resp.
  - Use the centers of $U$ to partition $T$. Call the resulting partition $W = \{W_1, \ldots, W_k\}$
- Measure $PS(t)$
- $k = \text{minimal } t > 1 \text{ s.t. } PS(t+1) < PS(t)$
FIG. 1. Plotting $z$-score as a function of $k$ for the Cho et al. data set. Each value of $z$-score is the average of three experiments. One can notice a decreasing trend after a first peak.
The full Alg: (1) finding k

- Partition N into $A, B, R$ \(|A|=|B|=\eta=\sqrt{|N|}\)
- Run FPF on $A, B$ till $k=\eta$
- $A^+=\text{set of centers extracted in } A \text{ up to iter } t$
- For $i=1,\ldots, \eta$
  - Cluster $R$ using FPF into i clusters $C$
  - Cluster $R$ using $A^i$ into i clusters $C'$
  - Compute $PS(i)$ by comparing $C, C'$
  - Stop when $PS(t-2)<PS(t-1)>PS(t)$. Set $t(A)=t$
- Repeat for $B^+$
- $k \leftarrow \text{rounded ave( } t(A)-1, t(B)-1\text{) }$
The full Alg: (2) clustering

- Use the centers in $C$ (from $R^{FPP}$) as initial centroids
- Enter all the remaining pts in random order, changing the centroids as we go along.

- Complexity: $O(mnk)$ (ex.)
Experiments: Yeast datasets

<table>
<thead>
<tr>
<th>Data set</th>
<th>Cho et al.</th>
<th>Eisen et al.</th>
<th>Spellman et al.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probes</td>
<td>6601</td>
<td>2467</td>
<td>6178</td>
</tr>
<tr>
<td>Conditions</td>
<td>17</td>
<td>79</td>
<td>82</td>
</tr>
<tr>
<td>Problem size</td>
<td>112,217</td>
<td>194,893</td>
<td>506,596</td>
</tr>
</tbody>
</table>

- **Table 1. Yeast: Summary of Data Set Properties**

- **Evaluation criteria:**
  - **Internal:** Homogeneity, Separation
  - **External:** zscore for match to known yeast functions (Gibbons Roth 2002) (more on this in a separate class)
Algs that are not given k:

**Table 2. Experimental Results Comparing Algorithms that Suggest a Plausible Value of k**

<table>
<thead>
<tr>
<th>Method</th>
<th>$k$</th>
<th>$S_g$</th>
<th>$T$</th>
<th>$z_{score}$</th>
<th>$k$</th>
<th>$S_g$</th>
<th>$T$</th>
<th>$z_{score}$</th>
<th>$k$</th>
<th>$S_g$</th>
<th>$T$</th>
<th>$z_{score}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLICK</td>
<td>30</td>
<td>136</td>
<td>540</td>
<td>62.47</td>
<td>8</td>
<td>0</td>
<td>165</td>
<td>42.26</td>
<td>27</td>
<td>17</td>
<td>3000</td>
<td>73.93</td>
</tr>
<tr>
<td>FPF-SB</td>
<td>14</td>
<td>0</td>
<td>16</td>
<td>61.93</td>
<td>12</td>
<td>0</td>
<td>19</td>
<td>57.90</td>
<td>16</td>
<td>0</td>
<td>59</td>
<td>58.20</td>
</tr>
<tr>
<td>K-Boost</td>
<td>10</td>
<td>0</td>
<td>21</td>
<td><strong>79.00</strong></td>
<td>8</td>
<td>0</td>
<td>23</td>
<td><strong>69.03</strong></td>
<td>19</td>
<td>0</td>
<td>134</td>
<td><strong>78.60</strong></td>
</tr>
</tbody>
</table>

For each algorithm and data set, we report the number $k$ of clusters, the number $S_g$ of singleton data points, the running time in seconds, and the $z_{score}$ computed by ClusterJudge. The results shown for $z_{score}$ are the average of three independent runs. *K-Boost* achieves a significantly better $z_{score}$ on all the three yeast data sets using far less computation time.

FPF-SB: An earlier version using a single random sample
Feeding $k$ to fixed-$k$ algs:

**Table 3. Cho et al. Dataset**

<table>
<thead>
<tr>
<th>Method</th>
<th>$k$</th>
<th>$T$</th>
<th>$z_{score}$</th>
<th>Hom</th>
<th>Sep</th>
<th>$k$</th>
<th>$T$</th>
<th>$z_{score}$</th>
<th>Hom</th>
<th>Sep</th>
<th>$k$</th>
<th>$T$</th>
<th>$z_{score}$</th>
<th>Hom</th>
<th>Sep</th>
</tr>
</thead>
<tbody>
<tr>
<td>FPF</td>
<td>14</td>
<td>5</td>
<td>56.5</td>
<td>0.572</td>
<td>0.011</td>
<td>30</td>
<td>10</td>
<td>52.43</td>
<td><strong>0.645</strong></td>
<td>0.016</td>
<td>10</td>
<td>4</td>
<td><strong>60.63</strong></td>
<td>0.548</td>
<td>−0.025</td>
</tr>
<tr>
<td>HAC</td>
<td>14</td>
<td>103</td>
<td>53.1</td>
<td>0.517</td>
<td>−0.143</td>
<td>30</td>
<td>103</td>
<td>56.57</td>
<td><strong>0.617</strong></td>
<td>−0.059</td>
<td>10</td>
<td>102</td>
<td><strong>61.00</strong></td>
<td>0.511</td>
<td>−0.151</td>
</tr>
<tr>
<td>$k$-means</td>
<td>14</td>
<td>18</td>
<td>74.6</td>
<td>0.655</td>
<td>−0.035</td>
<td>30</td>
<td>38</td>
<td>67.80</td>
<td><strong>0.703</strong></td>
<td>−0.001</td>
<td>10</td>
<td>10</td>
<td><strong>95.33</strong></td>
<td>0.631</td>
<td>−0.056</td>
</tr>
</tbody>
</table>

Experimental results comparing algorithms that take $k$ as input with the values computed by FPF-SB, CLICK, and K-Boost. For each algorithm and data set, we report the number $k$ of clusters, the running time $T$ in seconds, the $z_{score}$ computed by ClusterJudge Homogeneity Hom, and Separation Sep. The results shown for $z_{score}$ are the average of three independent runs. HAC has been run with average linkage. $k$-means has been run for 30 iterations.
### Table 4. Eisen et al. Dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>$k$</th>
<th>$T$</th>
<th>$z_{score}$</th>
<th>Hom</th>
<th>Sep</th>
<th>$k$</th>
<th>$T$</th>
<th>$z_{score}$</th>
<th>Hom</th>
<th>Sep</th>
</tr>
</thead>
<tbody>
<tr>
<td>FPF</td>
<td>12</td>
<td>6</td>
<td>53.4</td>
<td>0.483</td>
<td>0.079</td>
<td>8</td>
<td>4</td>
<td>56.87</td>
<td>0.524</td>
<td>-0.076</td>
</tr>
<tr>
<td>HAC</td>
<td>12</td>
<td>7</td>
<td>34.3</td>
<td>0.440</td>
<td>0.042</td>
<td>8</td>
<td>7</td>
<td>37.10</td>
<td>0.439</td>
<td>-0.292</td>
</tr>
<tr>
<td>$k$-means</td>
<td>12</td>
<td>14</td>
<td>62.3</td>
<td>0.528</td>
<td>0.102</td>
<td>8</td>
<td>10</td>
<td>64.86</td>
<td>0.572</td>
<td>-0.021</td>
</tr>
</tbody>
</table>

Experimental results comparing algorithms that take $k$ as input with the values computed by FPF-SB, CLICK, and K-Boost. For each algorithm and data set, we report the number $k$ of clusters, the running time $T$ in seconds, the $z_{score}$ computed by ClusterJudge, Homogeneity Hom, and Separation Sep. The results shown for $z_{score}$ are the average of three independent runs. HAC has been run with average linkage. $k$-means has been run for 30 iterations. Note that FPF, HAC, and $k$-means attain significantly better performance when fed with K-Boost’s estimate of $k$.

### Table 5. Spellman et al. Dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>$k$</th>
<th>$T$</th>
<th>$z_{score}$</th>
<th>Hom</th>
<th>Sep</th>
<th>$k$</th>
<th>$T$</th>
<th>$z_{score}$</th>
<th>Hom</th>
<th>Sep</th>
<th>$k$</th>
<th>$T$</th>
<th>$z_{score}$</th>
<th>Hom</th>
<th>Sep</th>
</tr>
</thead>
<tbody>
<tr>
<td>FPF</td>
<td>16</td>
<td>19</td>
<td>62.2</td>
<td>0.456</td>
<td>0.188</td>
<td>27</td>
<td>32</td>
<td>46.16</td>
<td>0.489</td>
<td>0.066</td>
<td>19</td>
<td>22</td>
<td>62.47</td>
<td>0.481</td>
<td>0.054</td>
</tr>
<tr>
<td>HAC</td>
<td>16</td>
<td>92</td>
<td>55.8</td>
<td>0.420</td>
<td>0.176</td>
<td>27</td>
<td>92</td>
<td>56.63</td>
<td>0.463</td>
<td>-0.018</td>
<td>19</td>
<td>92</td>
<td>57.00</td>
<td>0.448</td>
<td>-0.017</td>
</tr>
<tr>
<td>$k$-means</td>
<td>16</td>
<td>83</td>
<td>80.10</td>
<td>0.507</td>
<td>0.149</td>
<td>27</td>
<td>130</td>
<td>79.66</td>
<td>0.559</td>
<td>0.047</td>
<td>19</td>
<td>75</td>
<td>81.07</td>
<td>0.538</td>
<td>0.035</td>
</tr>
</tbody>
</table>

Experimental results comparing algorithms that take $k$ as input with the values computed by FPF-SB, CLICK, and K-Boost. For each algorithm and data set, we report the number $k$ of clusters, the running time $T$ in seconds, the $z_{score}$ computed by ClusterJudge, Homogeneity Hom, and Separation Sep. The results shown for $z_{score}$ are the average of three independent runs. HAC has been run with
Epilogue (from Geraci et. al)

Clustering is still an open and active research area, and there is no universally recommended method of choice. Accordingly, we used several quality measures and observed that no algorithm is the winner in all situations; so, depending on the purpose of the analysis, one might prefer methods that are stronger in separation or those that are stronger in homogeneity and zscore.