Cop k-means

Danilo Vasconcellos Vargas
If we had background knowledge...

- Sometimes there is information about the problem
  - Constraints
  - Also called Semi-supervised clustering
Constraints (Cop-kmeans)

- Must-link
- Cannot-link

- Constraints are never broken!
• Initialize k cluster centers
• Assign Phase
  - objects are assigned to closest cluster center
• Update Cluster Centers
  - update the cluster centers to the mean of constituent objects
Cop k-means

- Initialize k cluster centers

- Assign Phase
  - objects are assigned to closest cluster center *without violating constraints*

- Update Cluster Centers
  - update the cluster centers to the mean of constituent objects
Handling Constraints

For all objects try to assign it to closest $k$

• 1. No constraint broken:
  - Assign object $o$ to cluster $k$.

• 2. Broken $\rightarrow$ is there a next closest cluster?
  - Yes $\rightarrow$ Back to 1.
  - No $\rightarrow$ 3.

• 3. fail
Experimental Results

- 48% can be determined solely by constraints.
- 100 trials
- Held-out test is not directly affected by constraints

(Wagstaff et al, 2001)
More Results

(Wagstaff et al, 2001)
Comments on Results

- Are constraints worthwhile?
  - Depends on the dataset
  - Constraints can be generalized to the full dataset?

- Sensitivity to assignment order
  - Studied and solved by (Hong and Kwong, 2009) using an ensemble algorithm.
Comments on Results

- The set of constraints can vary and so their impact on the accuracy (Wagstaff, Basu and Davidson 2006)

<table>
<thead>
<tr>
<th>Data set</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Min</td>
</tr>
<tr>
<td>Glass</td>
<td>67.6</td>
</tr>
<tr>
<td>Iris</td>
<td>82.2</td>
</tr>
<tr>
<td>Ionosphere</td>
<td>58.2</td>
</tr>
<tr>
<td>Wine</td>
<td>68.0</td>
</tr>
</tbody>
</table>
Semi Supervised Clustering

- Similarity-adapting methods
  - Example: modifying the Euclidian Distance

- Search-based methods
  - Example: Cop-kmeans
Extensions

- Soft Constraints

Wagstaff, Kiri L., Basu, Sugato, Davidson, Ian “When is constrained clustering beneficial, and why?” National Conference on Artificial Intelligence, Boston, Massachusetts 2006.
