UNIFYING DEPENDENT CLUSTERING AND DISPARATE CLUSTERING FOR NON-HOMOGENEOUS DATA

M. Shahriar Hossain, Dept. of CS, Virginia Tech
Satish Tadepalli, Dept. of CS, Virginia Tech
Layne T. Watson, Dept. of CS, Virginia Tech
Ian Davidson, Dept. of CS, UC Davis
Richard F. Helm, Dept. of Biochemistry, Virginia Tech
Naren Ramakrishnan, Dept. of CS, Virginia Tech
## Problem Setting

<table>
<thead>
<tr>
<th>Companies</th>
<th>Avg. salary of Employees</th>
<th>Stock values</th>
<th>Profit margins</th>
</tr>
</thead>
<tbody>
<tr>
<td>x₁</td>
<td>1.0 K</td>
<td>25.11</td>
<td>11%</td>
</tr>
<tr>
<td>x₂</td>
<td>1.1 K</td>
<td>21.32</td>
<td>20%</td>
</tr>
<tr>
<td>x₃</td>
<td>1.2 K</td>
<td>28.81</td>
<td>12%</td>
</tr>
<tr>
<td>x₄</td>
<td>1.2 K</td>
<td>31.85</td>
<td>22%</td>
</tr>
<tr>
<td>x₅</td>
<td>1.1 K</td>
<td>85.32</td>
<td>5%</td>
</tr>
<tr>
<td>x₆</td>
<td>1.2 K</td>
<td>10.71</td>
<td>32%</td>
</tr>
<tr>
<td>x₇</td>
<td>0.9 K</td>
<td>11.61</td>
<td>18%</td>
</tr>
<tr>
<td>x₈</td>
<td>1.1 K</td>
<td>35.81</td>
<td>12%</td>
</tr>
<tr>
<td>x₉</td>
<td>1.2 K</td>
<td>20.81</td>
<td>4%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Countries</th>
<th>GDP</th>
<th>GNP</th>
<th>Inflation</th>
</tr>
</thead>
<tbody>
<tr>
<td>y₁</td>
<td>$11832 B</td>
<td>$12970 B</td>
<td>-0.4%</td>
</tr>
<tr>
<td>y₂</td>
<td>$8219 B</td>
<td>$8153 B</td>
<td>2.0%</td>
</tr>
<tr>
<td>y₃</td>
<td>$6732 B</td>
<td>$7812 B</td>
<td>-0.3%</td>
</tr>
<tr>
<td>y₄</td>
<td>$1761 B</td>
<td>$2852 B</td>
<td>1.8%</td>
</tr>
<tr>
<td>y₅</td>
<td>$5022 B</td>
<td>$4391 B</td>
<td>0.0%</td>
</tr>
<tr>
<td>y₆</td>
<td>$7224 B</td>
<td>$8312 B</td>
<td>1.1%</td>
</tr>
</tbody>
</table>
Fortunes of individual companies are intertwined with the fortunes of the countries.

Performances of companies may not necessarily be tied to the economies of the countries.
Objective Function

- Optimize $F$
  - Disparate clustering:
    - minimize: $F$
  - Dependent clustering:
    - maximize: $F$
    - minimize: $F$

- Quasi Newton Trust Region Algorithm

$v_{x_i}^s \quad v_{y_j}^s$

\[
\begin{array}{c}
\begin{array}{ccc}
C_1 & C_2 & C_3 \\
\hline
C_1' & & \end{array} \\
\begin{array}{ccc}
& C_2' & C_3' \\
C_1' & & \end{array} \\
\begin{array}{ccc}
& & C_3' \\
& & \end{array}
\end{array}
\]

\[
\begin{array}{ccc}
2 & 1 & 1 \\
2 & 1 & 2 \\
1 & 1 & 3
\end{array}
\]
Formulations

\[ v_i^{(x_s)} = \frac{\exp\left(-\frac{D}{B} \| x_s - m_{i,x} \|^2 \right)}{\sum_{i'=1}^{k_x} \exp\left(-\frac{D}{B} \| x_s - m_{i',x} \|^2 \right)} \]

\[ w_{i,j} = \sum_{s=1}^{n_x} \sum_{t=1}^{n_y} B(s, t) v_i^{(x_s)} v_j^{(y_t)} \]

\[ p(\alpha_i = j) = p(C_{(y)} = j \mid C_{(x)} = i) = \frac{w_{i,j}}{w_i} \]

\[ p(\beta_j = i) = p(C_{(x)} = i \mid C_{(y)} = j) = \frac{w_{i,j}}{w_j} \]

\[ \mathcal{F} = \frac{1}{k_x} \sum_{i=1}^{k_x} D_{KL} \left( \alpha_i \mid \mid U \left( \frac{1}{k_y} \right) \right) + \frac{1}{k_y} \sum_{j=1}^{k_y} D_{KL} \left( \beta_j \mid \mid U \left( \frac{1}{k_x} \right) \right) \]
Single Dataset Scenarios

**ALTERNATIVE CLUSTERING**

![Diagram showing two datasets D1 and D2 with vectors and relations labeled X1 to Xn, where D2 is a subset of D1.]
Single Dataset Scenarios

- **CONstrained CLUSTERING**
  - Instance-level constraints

- Clustering of $D_1$ is given.
- The desired constrained clustering is obtained in $D_2$.

$\mathcal{F} = \alpha \mathcal{F}_{dep} + (1-\alpha) \mathcal{F}_{disparate}$
Single Dataset Scenarios

• **CONstrained CLUSTERING**
  – Cluster-level constraints

[Diagrams illustrating the process with data points and clustering results.]

- Clustering of $D_1$ is given.
- The desired constrained clustering is obtained in $D_2$. 

$D_1$ $D_2 [= D_1]$
Experimental Results

• **ALTERNATIVE CLUSTERING**

• Portrait Dataset

- 3 people each in 3 poses and 36 illuminations (i.e., 324 images.)
- 300 features

Prateek Jain et al. 2008
Experimental Results

**ALTERNATIVE CLUSTERING**

Portrait dataset, Iterations=42
Accuracy person=93%, Accuracy pose=79%

(Accuracy axis is at left and the axis for objective function is at right)

- **Method** | **Person** | **Pose**
  - k-means | 0.65 | 0.55
  - Conv-EM | 0.69 | 0.72
  - Dec-kmeans | 0.84 | 0.78
  - Our framework | **0.93** | **0.79**
Experimental Results

- **CONSTRUANCED CLUSTERING**

Iris dataset: 200 random constraints

- # of clusters: 2, 3, 4, 5, 6, 7, 8, 9, 10
- # of constraints violated:
  - k-means
  - MPCK-MEANS
  - PCK-MEANS
  - Our framework

- Normalized mutual information:
  - k-means
  - MPCK-MEANS
  - PCK-MEANS
  - Our framework

![Graphs showing experimental results for different clustering methods and constraints](image)
Experimental Results

- **COMPARING GENE EXPRESSION PROGRAMS**

<table>
<thead>
<tr>
<th>Gene Count</th>
<th>Human</th>
<th>Yeast</th>
<th>Worm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>9,125</td>
<td>3,664</td>
<td>5,987</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pairs Relationships</th>
<th>Human-Worm</th>
<th>Worm-Yeast</th>
<th>Human-Yeast</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>12,000</td>
<td>8,002</td>
<td>9,012</td>
</tr>
</tbody>
</table>

**Experimental Results**

- **Human=Worm (original)**
- **Human=Worm (final)**
- **Worm=Yeast (original)**
- **Worm=Yeast (final)**

**Expression Clusters**

- **Human Clusters**
- **Worm Clusters**
- **Yeast Clusters**

**Human<>Yeast**

<table>
<thead>
<tr>
<th></th>
<th>561</th>
<th>604</th>
<th>312</th>
<th>137</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>284</td>
<td>1217</td>
<td>175</td>
<td>813</td>
</tr>
<tr>
<td></td>
<td>926</td>
<td>41</td>
<td>578</td>
<td>757</td>
</tr>
<tr>
<td></td>
<td>1034</td>
<td>47</td>
<td>800</td>
<td>726</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>298</th>
<th>537</th>
<th>388</th>
<th>499</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>318</td>
<td>376</td>
<td>698</td>
<td>588</td>
</tr>
<tr>
<td></td>
<td>807</td>
<td>452</td>
<td>622</td>
<td>781</td>
</tr>
<tr>
<td></td>
<td>606</td>
<td>729</td>
<td>454</td>
<td>859</td>
</tr>
</tbody>
</table>
Future Work & Conclusion

• Future directions
  – Capture more expressive relationships
    • Dependent and disparate clustering on same set of relationships
    • Different goal for different types of relationships (one-to-one, ML, MNL, etc.)
  – Clustering dependencies

• Conclusion
  – General, expressive framework for clustering non-homogenous datasets
  – The framework subsumes previously defined formulations
    • MDI (Kullback et al. ‘78), Disparate Clustering (Jain et al. ‘08), Clustering over Relation Graphs (Banerjee et al. ‘07), Multivariate Information Bottleneck (Friedman ‘01), etc.
Thank you