Redefining Class Definitions using Constraint-Based Clustering

An Application to Remote Sensing of the Earth’s Surface

Dan Preston
Stanford University

Carla Brodley
Tufts University

Roni Khardon
Tufts University

Mark Friedl
Boston University

Damien Sulla-Menashe
Boston University

KDD 2010
Where do classes come from?

• Categories and concepts that people find useful
  ‣ May not be supported by the features

• Clustering to find the homogeneous groups in the data
  ‣ May not be of use to humans
  ‣ Many different (equally “good”) clusterings
  ‣ How many clusters are there?
Is there more to this?
Do the features support these distinctions?
Our Approach

• Probabilistic constraint clustering using:
  ‣ original labels as constraints
  ‣ expert-belief as a guide

• A metric to evaluate a clustering result (in a constrained setting)
  ‣ ...and determine $K$

• Application to Remote Sensing
Constraint-Based Clustering: A Brief Review

• Constraints given as instance pairs
• Hard constraints for k-means [Wagstaff, et al 2001]
• Hard constraints with EM [Shental, et al, 2003]
• Probabilistic must-link constraints [Law, et al 2004]:
  expert defines group membership of each instance for n groups
• PPC: Probabilistic must and cannot link constraints, fast approximations [Lu and Leen, 2007]
What are our constraints?
Supplying Expert Belief

• The “C”-matrix allows the domain expert to supply preferences that label pairs exist or do not exist together:

Non-diagonal

\[
\begin{align*}
C(A,B) &= 1.0: \text{ merge classes } A \text{ and } B \\
C(A,B) &= -1.0: \text{ keep classes } A \text{ and } B \text{ separate} \\
C(A,B) &= 0.0: \text{ expert has no opinion}
\end{align*}
\]

Diagonal

\[
\begin{align*}
C(A,A) &= 1.0: \text{ do not split class } A \\
C(A,A) &< 1.0: \text{ ok to split } A \\
C(A,A) &= 0.0: \text{ ignore the labels} \\
C(A,A) &= -1.0: \text{ nonsense}
\end{align*}
\]
### Expert-Belief Matrix: Example

<table>
<thead>
<tr>
<th>Class</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>evergreen needleleaf forests</td>
<td>1.00</td>
<td>-0.60</td>
<td>-0.80</td>
<td>0.80</td>
<td>0.00</td>
<td>-0.80</td>
<td>0.60</td>
<td>0.00</td>
<td>-0.80</td>
<td>0.00</td>
<td>-0.80</td>
<td>-0.40</td>
<td>-0.80</td>
<td>-1.00</td>
<td>-0.60</td>
</tr>
<tr>
<td>evergreen broadleaf forests</td>
<td>-0.60</td>
<td>1.00</td>
<td>-0.60</td>
<td>-0.60</td>
<td>0.00</td>
<td>-0.80</td>
<td>0.60</td>
<td>0.00</td>
<td>-0.80</td>
<td>0.00</td>
<td>-0.60</td>
<td>0.00</td>
<td>-1.00</td>
<td>-1.00</td>
<td>-0.80</td>
</tr>
<tr>
<td>deciduous broadleaf forests</td>
<td>-0.80</td>
<td>-0.60</td>
<td>1.00</td>
<td>0.80</td>
<td>0.00</td>
<td>-0.80</td>
<td>0.60</td>
<td>0.00</td>
<td>-0.80</td>
<td>0.00</td>
<td>-0.60</td>
<td>0.00</td>
<td>-0.80</td>
<td>-1.00</td>
<td>-0.80</td>
</tr>
<tr>
<td>mixed forests</td>
<td>0.80</td>
<td>-0.60</td>
<td>0.80</td>
<td>1.00</td>
<td>0.00</td>
<td>-0.80</td>
<td>0.60</td>
<td>0.00</td>
<td>-0.80</td>
<td>0.00</td>
<td>-0.60</td>
<td>0.00</td>
<td>-0.80</td>
<td>-1.00</td>
<td>-0.80</td>
</tr>
<tr>
<td>closed shrublands</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>-0.40</td>
<td>-0.40</td>
<td>-0.40</td>
<td>-0.40</td>
<td>-0.60</td>
<td>0.00</td>
<td>-0.80</td>
<td>-0.80</td>
<td>-0.80</td>
</tr>
<tr>
<td>open shrublands</td>
<td>-0.80</td>
<td>-0.80</td>
<td>-0.80</td>
<td>-0.80</td>
<td>0.00</td>
<td>1.00</td>
<td>-0.20</td>
<td>-0.40</td>
<td>0.40</td>
<td>-0.60</td>
<td>-0.40</td>
<td>0.00</td>
<td>-0.60</td>
<td>0.20</td>
<td>-1.00</td>
</tr>
<tr>
<td>woody savannas</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>-0.40</td>
<td>-0.20</td>
<td>1.00</td>
<td>0.80</td>
<td>0.00</td>
<td>-0.60</td>
<td>-0.40</td>
<td>0.00</td>
<td>-0.80</td>
<td>-0.80</td>
<td>-1.00</td>
</tr>
<tr>
<td>savannas</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>-0.40</td>
<td>-0.40</td>
<td>0.80</td>
<td>1.00</td>
<td>0.60</td>
<td>-0.60</td>
<td>-0.40</td>
<td>0.40</td>
<td>-0.80</td>
<td>-0.80</td>
<td>-1.00</td>
</tr>
<tr>
<td>grasslands</td>
<td>-0.80</td>
<td>-0.80</td>
<td>-0.80</td>
<td>-0.80</td>
<td>-0.40</td>
<td>0.40</td>
<td>0.00</td>
<td>0.60</td>
<td>1.00</td>
<td>-0.60</td>
<td>0.40</td>
<td>0.40</td>
<td>-0.80</td>
<td>-0.20</td>
<td>-1.00</td>
</tr>
<tr>
<td>permanent wetlands</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>-0.60</td>
<td>-0.60</td>
<td>-0.60</td>
<td>-0.60</td>
<td>-0.60</td>
<td>1.00</td>
<td>0.00</td>
<td>0.00</td>
<td>-1.00</td>
<td>-1.00</td>
<td>0.60</td>
</tr>
<tr>
<td>croplands</td>
<td>-0.80</td>
<td>-0.60</td>
<td>-0.60</td>
<td>-0.60</td>
<td>-0.40</td>
<td>-0.40</td>
<td>-0.40</td>
<td>0.40</td>
<td>0.00</td>
<td>1.00</td>
<td>0.80</td>
<td>-0.80</td>
<td>-0.80</td>
<td>-1.00</td>
<td>0.60</td>
</tr>
<tr>
<td>urban and built-up lands</td>
<td>-0.40</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.40</td>
<td>0.00</td>
<td>0.80</td>
<td>1.00</td>
<td>-0.80</td>
<td>-0.80</td>
<td>-1.00</td>
</tr>
<tr>
<td>natural vegetation mosaics</td>
<td>-0.80</td>
<td>-1.00</td>
<td>-0.80</td>
<td>-0.80</td>
<td>-0.80</td>
<td>-0.60</td>
<td>-0.80</td>
<td>-0.80</td>
<td>-0.80</td>
<td>-1.00</td>
<td>-0.80</td>
<td>-0.80</td>
<td>1.00</td>
<td>-0.40</td>
<td>-1.00</td>
</tr>
<tr>
<td>barren</td>
<td>-1.00</td>
<td>-1.00</td>
<td>-1.00</td>
<td>-1.00</td>
<td>-0.80</td>
<td>0.20</td>
<td>-0.80</td>
<td>-0.80</td>
<td>-0.20</td>
<td>-1.00</td>
<td>-0.80</td>
<td>-0.80</td>
<td>-0.40</td>
<td>1.00</td>
<td>-0.80</td>
</tr>
<tr>
<td>water bodies</td>
<td>-0.60</td>
<td>-0.80</td>
<td>-0.80</td>
<td>-0.80</td>
<td>-1.00</td>
<td>-1.00</td>
<td>-1.00</td>
<td>-1.00</td>
<td>0.60</td>
<td>-1.00</td>
<td>-1.00</td>
<td>-1.00</td>
<td>-0.80</td>
<td>1.00</td>
<td>-1.00</td>
</tr>
</tbody>
</table>

**“C”-Matrix**
Incorporating Constraints into EM

\[ q_{ik} \sim P_k P(y_i^*|\theta_k) \exp \left( 2 \sum_{j=1, j \neq i} \lambda C(\ell_i, \ell_j) q_{jk} \right) \]

\[ \Sigma_k = \frac{\sum_{i=1}^N q_{ik} (y_i^* - \mu_k^*) (y_i^* - \mu_k^*)^T}{\sum_{i=1}^N q_{ik}} \]

\[ \mu_k^* = \frac{\sum_{i=1}^N y_i q_{ik}}{\sum_{i=1}^N q_{ik}} \]

\[ P_k = \frac{1}{N} \sum_{i=1}^N q_{ik} \]

Same as EM. Makes simplifying approximations
Algorithm Speed Up

• Compact Expert-belief matrix creates redundant computation in $q_{ik}$

• We can exploit this to reduce complexity from $O(N^2)$ to $O(NL)$:

$$S^*(\ell, k) = \sum_{i=1}^{N} \lambda C(\ell, \ell_i) q_{ik}$$

$$q_{ik} \sim P_k P(\tilde{y}_i | \theta_k) \exp(2(S^*(\ell_i, k) - \lambda C(\ell_i, \ell_i) q_{ik}))$$
What is a good clustering?
Evaluating a clustering

- Heuristic criteria have a fit term and a complexity penalty term:
  \[ BIC = N \log\left(\frac{RSS}{N}\right) + k \log N \]
  \[ AIC = N \log\left(\frac{RSS}{N}\right) + 2k \]

- Our heuristic will use BIC and add a term for constraint adherence.
Assessing Constraint Adherence

- Create an L x L matrix $V$, where $V(A, B) \in [-1, 1]$ measures how frequently instances from class A and class B cluster together.
- $V(A,B) = 1$: all instances from classes A and B found together in the same cluster.
- $V(A,B) = -1$: no instances from A and B are in the same cluster.

$V(A,B)$ is a normalized count (to $[-1,1]$), where:

+1 for each pair of A,B that appear in the same cluster,
-1 for each that appear in different clusters.
Calculating Adherence

\[ G(C, \{z_i\}) = \sum_{A \in L} \sum_{B \in L} (C(A, B) - V(A, B))^2 \]

- Sum of the squared difference between the expert-defined constraints and the appearance measure \( V(A,B) \)
- Measures adherence to constraints
Constrained BIC

Fit term + Complexity Penalty term:

\[ BIC = N \log\left(\frac{RSS}{N}\right) + k \log N \]

Fit term + Complexity Penalty term + Constraint Adherence term:

\[ cBIC = (1 - \lambda)N \log\left(\frac{RSS}{N}\right) + \lambda N \log \frac{G(C, \{z_i\})}{4L^2} + k \log N \]
Evaluation Review

• What you need to know:

\[ G(C, \{z_i\}) = \text{how well we adhere to constraints} \]
\[ \text{BIC} = \text{balances complexity and cluster fit} \]
\[ \text{cBIC} = \text{a blending of these} \]
EM versus PPC
PPC with Strong Constraints

<table>
<thead>
<tr>
<th>Class</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.95</td>
<td>0.02</td>
</tr>
<tr>
<td>2</td>
<td>0.02</td>
<td>0.99</td>
</tr>
</tbody>
</table>
PPC with Weak Constraints

<table>
<thead>
<tr>
<th>Class</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.20</td>
<td>0.01</td>
</tr>
<tr>
<td>2</td>
<td>0.01</td>
<td>0.15</td>
</tr>
</tbody>
</table>
Choosing $k$ with cBIC

Strong Constraints

$\text{k = 2}$

Weak Constraints

$\text{k = 4}$
Remote Sensing Data Set
Merging Classes

IGBP

CPPC

EM

New England & Montreal
Splitting Classes

IGBP
CPPC
EM

Iowa (agriculture)
Retaining Structure

Mexico

IGBP

CPPC

EM
Summary

- Framework for redefining our classes, with expert preferences and reusing our labels
- Compact representation provides CPPC complexity reduction from $O(N^2)$ to $O(NL)$
- cBIC provides a tool to evaluate a constrained clustering, and helps to determine $k$
- Promising results in Remote Sensing
Thank you

Email: dpreston@cs.stanford.edu