Multiple Clustering Views via Constrained Projections

Xuan-Hong Dang, Aarhus University, Denmark
(joint work with Ira Assent, Aarhus University, and James Bailey, University of Melbourne)

28 avril 2012
1 Introduction
2 Alternative Clustering Problem
3 Subspace Learning with Constraint
4 Initial Experiments
5 Conclusions
Introduction

- Clustering: categorizes similar objects into same groups
- High dimensional data → Multiple clusterings may exist.

- Other data: text/document data, gene data, ...
- Challenge: How to find all meaningful solutions?
Several algorithms have been developed.

- Seeking alternative clusterings simultaneously.
  
  \[
  \text{Eg: Maximize } L(\Theta^{(1)}; \mathcal{X}) + L(\Theta^{(2)}; \mathcal{X}) - I(C^{(1)}; C^{(2)}|\Theta)
  \]

- Seeking alternative clusterings in sequence.
  
  \[
  \text{Eg: Maximize } L(\Theta^{(2)}; \mathcal{X}) - I(C^{(1)}; C^{(2)})
  \]

\[\Rightarrow \text{model view point: latter approach has limited number of parameters optimized}\]

\[\Rightarrow \text{our approach in this work}\]

Given \(\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}\) in \(\mathbb{R}^d\) and \(C^{(1)}\) as reference, seek \(C^{(2)}\) as an alternative: \(\bigcup_i C_i^{(2)} = \mathcal{X}\) and \(C_i^{(2)} \cap C_j^{(2)} = \emptyset\) for \(\forall i \neq j; \ i, j \leq k\)
**Objective**

Subspace learning:

- **un-correlate from** $C^{(1)} \Rightarrow$ ensure difference.
- **retaining local data proximity** \( \Rightarrow \) ensure quality.

With graph based approach:

- \( F \) : maps \( \{x_i\}_{i=1}^n \) into \( \{y_i\}_{i=1}^n \) (i.e., \( Y = F^T X \))
  \( \Rightarrow \) \( f \) in \( F \) combines \( X \) into 1-dim: \( f^T X = \{y_1, \ldots, y_n\} = y^T \).
- Define objective:
  \[
  \arg\min_f \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (f^T x_i - f^T x_j)^2 K_{ij} \quad \text{s.t.} \quad S^T X^T f = 0
  \]
  \( \Rightarrow S \) is a feature subspace capturing \( C^{(1)} \)
  \( \Rightarrow \) Penalize : \( K_{ij} \) large but \( y_i, y_j \) are mapped far apart
Learn $S$ with LDA

- Learn $S$ as a subspace best capturing $C^{(1)}$.
  $\Rightarrow C^{(1)}$’s clusters represented in $S$ are most separable.

- Fisher LDA is a good choice:
  \[
  \max_w \frac{w^T S_B w}{w^T S_W w}
  \]
  \[
  S_B = \sum_k n_k (m^{(k)} - \mu)(m^{(k)} - \mu)^T
  \]
  \[
  S_W = \sum_k \sum_i (x_i^{(k)} - m^{(k)})(x_i^{(k)} - m^{(k)})^T
  \]
  Optimal $w$’s are eigenvectors of $S_W^{-1} S_B$
  $S$ is chosen with leading $w$’s.
Solving Constrained Function(1)

- Define $D$ with $D_{ii} = \sum_j K_{ij}$ and $L = D - K$
- Deploying summation:

$$\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (f^T x_i - f^T x_j)^2 K_{ij} = f^T XLX^T f$$

- Adding $f^T XDX^T f = 1$ to remove $f$'s freedom:

$$\mathcal{L}(\alpha, \beta, f) = f^T XLX^T f - \alpha(f^T XDX^T f - 1) - \beta S^T X^T f$$

- For simplicity:

$$\left\{ \begin{array}{l}
\tilde{L} = XLX^T \\
\tilde{D} = XDX^T \\
\tilde{S} = XS
\end{array} \right.$$
Solving Constrained Function(2)

- $\tilde{D}$ is symmetric, pos.semi-definite. Change $f = \tilde{D}^{-1/2}z$ :
  \[ f^T \tilde{L} f = z^T \tilde{D}^{-1/2} \tilde{L} \tilde{D}^{-1/2} z = z^T Qz \]

and two constraints :
  \[ \begin{align*}
    f^T \tilde{D} f &= z^T z = 1 \\
    \tilde{S}^T f &= \tilde{S}^T \tilde{D}^{-1/2} z = 0
  \end{align*} \]

- Lagrange function can be re-written :
  \[ \mathcal{L}(\alpha, \beta, z) = \frac{1}{2} z^T Qz - \frac{1}{2} \alpha (z^T z - 1) - \beta U^T z \]
  where $U^T = \tilde{S}^T \tilde{D}^{-1/2}$.

- Taking derivative and with little algebra :
  \[ \alpha z = Qz - U(U^T U)^{-1} U^T Qz \]
  \[ = (I - U(U^T U)^{-1} U^T) Qz \]
  \[ = PQz \]

$\Rightarrow$ eigenvalue problem
Solving Constrained Function (3)

- Solving: \( \alpha \mathbf{z} = P \mathbf{Qz} \)
- Notice \( P \mathbf{Q} \) might not be symmetric; yet, \( \alpha(P \mathbf{Q}) = \alpha(P \mathbf{QP}) \) due to \( P^T = P \) and \( P^2 = P \).
  \[ \Rightarrow \text{not solving } P \mathbf{Qz} = \alpha \mathbf{z} \text{ but } P \mathbf{QPv} = \alpha \mathbf{v}, \text{ with } \mathbf{v} = P^{-1} \mathbf{z} \]
- Eigenvalues of \( P \mathbf{QP} \) are non-negative: 
  \[ \Rightarrow \alpha_0 = 0 \text{ is smallest} \]
  \[ \Rightarrow \mathbf{v}_0 = P^{-1} \mathbf{D}^{1/2} \mathbf{1} \text{ is trivial} \]
- Optimal direction \( \mathbf{f} : \)
  \[ \mathbf{f} = \mathbf{D}^{-1/2} P \mathbf{v} \]

with corresponding smallest non-zero eigenvalue \( \alpha \).
\[ \Rightarrow \mathbf{F} \text{ is formed based on } q \text{ leading eigenvectors of } P \mathbf{QP} \text{ corresponding to smallest non-zero } \alpha 's. \]
Initial experimental results

(a) Synthetic data with 4 Gaussians

(b) Cloud data from UCI repository

(c) Housing data from UCI repository
Conclusions

- Novel approach from subspace learning
  - not only being uncorrelated from provided clustering
  - but also retaining local geometrical data proximity

- Global optimum solution can be achieved

- Capability of seeking multiple clusterings (adding more subspaces into $S$).

- The approach is extendable for non-linear cases.

- Future work:
  - More experiments required on diverse datasets
  - Soft constraint with tradeoff factor (subspace independence vs. local data structure retaining)
  - Alternative clustering interpretation.