Multi-Source Visual Analytics

Jieping Ye
Computer Science and Engineering
The Biodesign Institute
Arizona State University

Co-PIs: Anshuman Razdan, Peter Wonka
Dimensionality Reduction for Data Visualization

Left: Visualizing data points as rectangles. Right: A magnifying lens
Dimensionality Reduction Algorithms

- **Supervised:**
  - Linear discriminant analysis (LDA)
  - Canonical correlation analysis (CCA)
  - Partial least squares (PLS)

- **Unsupervised:**
  - Principal component analysis (PCA)
  - Manifold learning (Isomap, LLE, Laplacian Eigenmap)

Original data \[\rightarrow\] Dimensionality reduction \[\rightarrow\] Reduced data

- LDA, CCA, PLS, PCA
- SVM, NN
Clustering and Dimensionality Reduction (1)

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups.

Intra-cluster distances are minimized

Inter-cluster distances are maximized
Standard PCA fails to detect these two natural clusters, whereas the proposed cluster sensitive dimensionality reduction (CSDR) does a much better job of separating the data.

**How can we combine clustering and dimensionality reduction to improve visual analytics tasks?**
Multi-source Data Transformations

• Processing heterogenous data is a significant challenge in visual analytics.
  – For example, an analyst may want to analyze data from multiple sources like images, text (emails), and telephone conversations.

• We propose to investigate techniques to transform entities that come from different sources.
Multiple Kernel Learning for Data Fusion
Research Aims

• Clustering and dimensionality reduction
  – Single source data transformation

• Clustering and dimensionality reduction
  – Multi-source data transformation

• MSVA a novel Visual Analytics Framework
The Proposed MSVA framework

A user can draw from a number of data transformation and visual analysis tools. A typical sequence of data processing is shown by the arrows. The user can interactively provide feedback and update the transformation.
Preliminary Work: Problem Setup

Given \( \{x_1, x_2, \ldots, x_n\} \in \mathbb{R}^m \)

Let \( X = \begin{bmatrix} x_1, x_2, \ldots, x_n \end{bmatrix} \) be the data matrix

\[ W \in \mathbb{R}^{m \times l} : x_i \in \mathbb{R}^m \implies \hat{x}_i = W^T x_i \in \mathbb{R}^l \]

Clustering \( C_1, C_2, \ldots, C_k \)

• It has been shown that for most high-dimensional data sets, almost all low dimensional projections are nearly normal.
Mahalanobis Distance

\[ \{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n\} \in \mathbb{R}^l \quad \text{nearly normal for large } m \]

Mahalanobis distance

\[ d_M(\hat{x}_i, \hat{x}_j) = \sqrt{\left(\hat{x}_i - \hat{x}_j\right)^T \hat{S}^{-1} \left(\hat{x}_i - \hat{x}_j\right)} \]

where

\[ \hat{S} = \frac{1}{n} \sum_{i=1}^{n} (\hat{x}_i - \hat{\mu})(\hat{x}_i - \hat{\mu})^T = W^T SW \]

\[ S = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^T \]

regularization

\[ S = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^T + \lambda I_m \]
Sum of Squared Error

Under this new distance measure, K-means clustering assigns the data into $k$ disjoint clusters, which minimizes the Sum of Squared Error (SSE):

$$\text{SSE} \left( \left\{ C_j \right\}_{j=1}^k \right) = \sum_{j=1}^k \sum_{\hat{x}_i \in C_j} d_M (\hat{x}_i, \mu_j)^2.$$
Sum of Squared Inter-Cluster Error

As the summation of all pair-wise distances is a constant for a fixed $W$, the minimization of SSE is equivalent to the maximization of Sum of Squared Inter-Cluster Error (SSIE):

$$\text{SSIE}\left(\left\{C_j\right\}_{j=1}^k\right) = \sum_{j=1}^k n_j d_M (\hat{\mu}_j, \hat{\mu})^2.$$
Compact Matrix Formulation

Sum of Squared Intra-Cluster Error (SSIE) can be expressed in a compact matrix form as follows:

$$SSIE\left(\{C_j\}_{j=1}^k\right) = \text{trace}\left(L^T X^T W^T \left(W^T S W\right)^{-1} W X L\right)$$

$L$ is the weighted cluster indicator matrix, whose $i$-th row is

$$L_i = \frac{1}{\sqrt{n_i}} \begin{pmatrix} 0, \ldots, 0, 1, \ldots, 1, 0, \ldots, 0 \end{pmatrix}^T$$

Joint dimensionality reduction and clustering formulation:

$$\max_{W,L} \text{trace}\left(L^T X^T W^T \left(W^T S W\right)^{-1} W X L\right)$$
## Preliminary Study

<table>
<thead>
<tr>
<th></th>
<th>Proposed</th>
<th>PCA</th>
<th>LLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCM</td>
<td>0.583</td>
<td>0.568</td>
<td>0.569</td>
</tr>
<tr>
<td>Soybean</td>
<td>0.725</td>
<td>0.671</td>
<td>0.668</td>
</tr>
<tr>
<td>Segment</td>
<td>0.644</td>
<td>0.552</td>
<td>0.551</td>
</tr>
<tr>
<td>Letter (a-d)</td>
<td>0.662</td>
<td>0.606</td>
<td>0.606</td>
</tr>
<tr>
<td>USPS</td>
<td>0.726</td>
<td>0.708</td>
<td>0.709</td>
</tr>
<tr>
<td>YaleFaceB</td>
<td>0.771</td>
<td>0.733</td>
<td>0.733</td>
</tr>
<tr>
<td>Average</td>
<td>0.685</td>
<td>0.64</td>
<td>0.639</td>
</tr>
</tbody>
</table>
Semi-supervised Setting

Domain knowledge
Use feedback
Proposed research

• Single source data transformation

• Multi-source data transformation

• Sparse data transformation

• Applications
  – Visual document analysis
  – Geo-spatial analysis
  – Health information analysis
Application I: Visual Document Analysis

• The capability to quickly process, tag/annotate, triage and classify volumes of information is key to enabling effective and useful information analysis.
Application II: Geo-spatial Analysis

The hyperspectral image is shown on the left, and a thematic map of land cover classes is on the right.
Application III: Health Information Analysis

Demographic, genetic, cognitive measures
Questions!