Interactive Visual Analytics for High Dimensional Data

Haesun Park
School of Computational Science and Engineering
Georgia Institute of Technology
Atlanta, GA, U.S.A.

KDD IDEA Workshop, August 2013

This work is supported in part by NSF/DHS and DARPA.
Contributors

• Jaegul Choo (Georgia Tech)
• Jingu Kim (Nokia)
• Da Kuang (Georgia Tech)
• Yunlong He (Georgia Tech)
• Changhyun Lee (Georgia Tech)
• Hanseung Lee (Univ. of Maryland)
• Zhicheng Liu (Stanford University)
• Fuxin Li (Georgia Tech)
• Jaeyeon Kihm (Cornell University)
• Chandan Reddy (Wayne State University)
• Barry Drake (Georgia Tech Research Institute)
• Ed Clarkson (Georgia Tech Research Institute)

• Polo Chau (Georgia Tech)
• Alexander Gray (and many of his students, Georgia Tech)
• Vladimir Koltchinskii (Georgia Tech)
• John Stasko (Georgia Tech)
Data and Visual Analytics

The Science of Analytical Reasoning facilitated by Automated Data Analysis and Interactive Visual Interfaces. (modified based on Thomas and Cook, Illuminating the Path: the research and development agenda for visual analytics, 2005)

Data → Information → Knowledge → Visualization → Data Mining/Data Analysis
“Solving a problem simply means representing it so that the solution is obvious.”

Herbert Simon, 96
The FODAVA Mission:
To develop and advance the mathematical and computational foundations of data and visual analytics through innovative research, educational programs, and the development of workforce to address the challenges of extracting knowledge from massive, complex data.

fodava.gatech.edu
Challenges in Computational Methods for High Dimensional Large-scale Data on Visual Analytics System

• Data challenges
  – Massive, High-dimensional, Nonlinear
  – Vast majority of data is unstructured
  – Noisy, errors and missing values are inevitable in real data set
  – Heterogeneous format/sources/reliability
  – Time varying, dynamic, …

• Visualization challenges
  – Screen Space and Visual Perception
    • High dimensional data: Effective dimension reduction
    • Large data sets: Informative representation of data
  – Speed: necessary for real-time, interactive use
    • Scalable algorithms
    • Adaptive algorithms
Key Foundational Components of VA System for High Dim. Large Data

- **Dimension Reduction**
  - Dimension reduction with prior info/interpretability constraints
  - Manifold learning

- **Informative Presentation of Large Scale Data**
  - Clustering, semi-supervised clustering
  - Multi-resolution data approximation

- **Fast Algorithms**
  - Large-scale optimization/matrix computations
  - Adaptive updating algorithms for dynamic and time-varying data, and interactive vis.

- **Information Fusion**
  - Fusion of different types of data from various sources, vis. comparisons

- **Integration with DAVA systems**
  - FODAVA Testbed, VisIRR, UTOPIAN, iVisClassifier, iVisClustering, PIVE, Jigsaw, …
FODAVA Research Testbed for Visual Analytics of High Dimensional Data

- Library of key computational methods for visual analytics of high dimensional data
  - Foundational data analysis methods with visual interactions implemented
  - Easily accessible to a wide community of researchers and readily available for applications
- Identifies effective methods for specific problems (evaluation)
- Modular: A base for specialized VA systems
  (e.g. iVisClassifier, iVisClustering, VisIRR, UTOPIAN, …)

\[
\begin{align*}
S_w &= \sum_{j \leq k} \sum_{j \in N_i} (a_i - c_j) (a_i - c_j)^T \\
S_b &= \sum_{j \leq k} \sum_{j \in N_i} (c_i - c_j) (c_i - c_j)^T \\
S_i &= \sum_{j \leq k} (a_i - c) (a_i - c)^T
\end{align*}
\]
FODAVA Research Testbed Software: Available at 
http://fodava.gatech.edu/fodava-testbed-software

- Supports various dimension reduction, clustering, and their visual representations and comparisons through alignments for high-dimensional data
- Application domains: document analysis, bioinformatics, seismic data analysis, healthcare, communications, computer vision, ...
- Language used: backend library in Matlab, GUI in JAVA (no need for Matlab installed)
- System support: Windows 32/64 bit, Linux 32/64 bit
Other Related VA Systems

- **IN-SPIRE** [http://in-spire.pnnl.gov]
  - Uses $K$-means and PCA for document clustering and visualization
  - Limited algorithms and interaction capability

- **Jigsaw** [http://www.cc.gatech.edu/gvu/ii/jigsaw]
  - Uses $K$-means and SVD/LSI for document clustering and similarity computation
  - More entity based, offers different types of interactions

- **GGobi** [http://www.ggobi.org]
  - Provides an animated projection view called grand tour
  - Limited applicability to very high-dim data

- **iPCA** [Jeong et al., iPCA: An Interactive System for PCA-based Visual Analytics]
  - PCA-centric interactive visualization tool for generic data
  - Limited applicability to very high-dim data

- **WEKA** [http://www.cs.waikato.ac.nz/ml/weka]
  - Supports various machine learning algorithms with basic interactions
  - Lacks data exploration/interaction capabilities
Testbed Modules

- Computational modules
  - Vector encoding
  - Pre-processing
  - Clustering
  - Dimension reduction

- Interactive visualization modules
  - Parallel coordinates
  - Scatter plot
  - Cluster summary
  - Brushing and Linking
  - Space alignment
  - Raw data view
Dimension Reduction

• Visualizes high-dimensional data by parallel coordinates and/or scatter plot

• Methods
  • Linear methods
    • PCA, FA, ProbPCA, LDA, OCM, NPE, LPP, LLTSA, NCA, MCML
  • Nonlinear methods
    • MDS, Isomap, LLE, LTSA, Sammon, HessLLE, MVU, LandMVU, KernPCA, GDA, DiffMaps, SPE, AutoEnc, LLC, ManiChart, CFA, GPLVM, SNE, T-SNE

• Provides initial parameters and can change them interactively.
• Can recursively apply dimension reduction on user-selected data.
• Fast algorithms are implemented.
Clustering and Classification

- Generates cluster/class labels of data, which are color-coded in visualization.
- Methods
  - Clustering
  - Classification (on-going work)
    - $K$-nearest neighbors classifier, SVM, Logistic regression, Naïve Bayes
- Provides cluster summary
- Provides GUI for semi-supervision, e.g., must/cannot link
- Can hierarchically construct cluster structures
Testbed Overview
Role of Parallel Coordinates
Fast Comp. Modules for Interactive Vis.

- Essential for real-time interaction
- Low-precision computation is often good enough as visual insight is the initial goal
- Iterative/hierarchical refinement
- Adaptive algorithms

p-Isomap computing time vs. $k$ value in $k$-NN graph

PCA timing: double vs single precision computation time and results

48x36 vs 80x60
Key Computational Methods

• LDA/GSVD: for 2D representation of clustered data
• NMF: for dimension reduction and clustering
• Orthogonal Procrustes and MDS: for space alignment and comparisons of visual representations
LDA/GSVD for 2D Representation of High Dimensional Clustered Data

2D representation of 700x1000 data with 7 clusters: LDA vs. SVD vs. PCA
Want to represent data/cluster/outlier info even after a severe dim. reduction
Linear Discriminant Analysis (LDA) for 2D/3D Representation of Clustered Data
(J. Choo, S. Bohn, HP, VAST09)

max \text{trace}(G^T S_b G) \quad \text{min} \text{trace}(G^T S_w G)

LDA/GSVD
\[ \alpha^2 H_b H_b^T x = \beta^2 H_w H_w^T x \]

\[ \text{max trace} \quad (G^T (S_w + \mu I) G)^{-1} (G^T S_b G) \]

- Regularization in LDA for Computational Zooming

Small regularization \quad \text{Large regularization}
2D Visualization of Clustered Text, Image, Audio Data

Medline Data (Text)  Facial Data (Image)  Spoken Letters (Audio)
Nonnegative Matrix Factorization (NMF)


Why Nonnegativity Constraints?
- Better Approx. vs. Better Representation/Interpretation
- Nonnegative Constraints often *physically meaningful*
- *Interpretation of analysis results easier*

We developed one of the Fastest Algorithms for NMF & conv. analysis

Matlab codes available (J. Kim and H. Park, IDCM08, SISC11)

http://www.cc.gatech.edu/~hpark/nmfsoftware.php

NMF is better and faster than K-means in clustering
- *K*-means: $W$: $k$ cluster centroids, $h_i$: cluster membership indicator
- NMF: $W$: basis vectors for rank-$k$ approx., $h_i$: $k$-dim rep. of $a_i$
- SymNMF (Kuang, Ding, Park, SDM12), Sparse NMF for clustering (Kim and Park, Bioinfo., 07)
NMF Algorithm Comparison (Residual vs. Time, J. Kim and HP 2011)

20 Newsgroups: 26, 214x11, 314, k=160

PIE 64 image: 4, 096x11,554, k = 80
NMF and K-means

- Clustering and Lower Rank Approximation are related.
  - NMF for Clustering: (Ding et al. SDM 05; Kim & Park, TR 08)
  - Document (Xu et al. SIGIR 03), Image (Cai et al. ICDM 08), Microarray (Kim & Park, Bio 07), etc.
  - \( \min \sum_{1 \leq i \leq n} \| a_i - w_{\sigma i} \|^2 = \min \| A - WH \|^2_F \)
    \( \sigma_i = j \) when \( i \)-th point is assigned to \( j \)-th cluster (\( j \in \{1, ..., k\} \))

**K-means**: \( W \): \( k \) cluster centroids, \( h_j \): cluster membership indicator

**NMF**: \( W \): basis vectors for rank-\( k \) approx., \( h_j \): \( k \)-dim rep. of \( a_i \)

**Sparse NMF** (for sparse \( H \)) (H. Kim and Park, Bioinformatics, 07)

\[
\min_{W,H} \left\{ \|A-WH\|_F^2 + \eta \|W\|_F^2 + \beta \sum_{1 \leq j \leq n} \|H(:,j)\|_1^2 \right\}, \ \forall \ i,j, \ W_{ij}, H_{ij} \geq 0
\]

Obj. fun. of K-means and NMF are related when \( h_i \in \{e_1...e_k\} \) and \( A \geq 0 \), but their performances may be very different.
NMF for Clustering

- NMF more accurate and faster on document and image data
  (Xu et al. 03; Pauca et al. 04; Li et al. 07; Kim & Park, 08; Ding et al. 10 ...)

  - Clustering accuracy averaged over 20 runs:

<table>
<thead>
<tr>
<th></th>
<th>K</th>
<th>K-means</th>
<th>SphKmeans</th>
<th>NMF/ANLS</th>
<th>GNMF</th>
<th>Spectral</th>
<th>SymNMF</th>
</tr>
</thead>
<tbody>
<tr>
<td>TDT2</td>
<td>4</td>
<td>.7994</td>
<td>.7978</td>
<td>.9440</td>
<td>.9150</td>
<td>.9093</td>
<td>.9668</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>.6147</td>
<td>.6208</td>
<td>.8292</td>
<td>.8200</td>
<td>.7357</td>
<td>.8819</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>.5286</td>
<td>.5305</td>
<td>.6709</td>
<td>.6812</td>
<td>.5959</td>
<td>.7635</td>
</tr>
<tr>
<td>Reuters</td>
<td>4</td>
<td>.5755</td>
<td>.5738</td>
<td>.7737</td>
<td>.7798</td>
<td>.7171</td>
<td>.8077</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>.5170</td>
<td>.5049</td>
<td>.6747</td>
<td>.6758</td>
<td>.6452</td>
<td>.7343</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>.3712</td>
<td>.3687</td>
<td>.4608</td>
<td>.5338</td>
<td>.5001</td>
<td>6688</td>
</tr>
</tbody>
</table>

- Timing results averaged over 20 runs, on 20 Newsgroups data set:

<table>
<thead>
<tr>
<th></th>
<th>TDT2</th>
<th>Reuters</th>
<th>COIL20</th>
<th>NIPS</th>
<th>ORL</th>
<th>PIE</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-means</td>
<td>.6734</td>
<td>.4289</td>
<td>.6184</td>
<td>.4650</td>
<td>.6499</td>
<td>.7384</td>
<td>.5957</td>
</tr>
<tr>
<td>NMF/ANLS</td>
<td>.8534</td>
<td>.5770</td>
<td>.6312</td>
<td>.4877</td>
<td>.7020</td>
<td>.7912</td>
<td>.6404</td>
</tr>
<tr>
<td>SymNMF</td>
<td>.8979</td>
<td>.5305</td>
<td>.7258</td>
<td>.5129</td>
<td>.7798</td>
<td>.7517</td>
<td>.6998</td>
</tr>
</tbody>
</table>

- Problem sizes:

<table>
<thead>
<tr>
<th></th>
<th>TDT2</th>
<th>Reuters</th>
<th>COIL20</th>
<th>NIPS</th>
<th>ORL</th>
<th>PIE</th>
<th>20 Newsgroups</th>
</tr>
</thead>
<tbody>
<tr>
<td>m</td>
<td>26618</td>
<td>12998</td>
<td>4096</td>
<td>17583</td>
<td>5796</td>
<td>4096</td>
<td>36982</td>
</tr>
<tr>
<td>n</td>
<td>8741</td>
<td>8095</td>
<td>1440</td>
<td>420</td>
<td>400</td>
<td>232</td>
<td>18669</td>
</tr>
<tr>
<td>k</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>9</td>
<td>40</td>
<td>68</td>
<td>20</td>
</tr>
</tbody>
</table>
Information Fusion based on Space Alignment (J. Choo, S. Bohn, G. Nakamura, A. White, HP)

• Want: Unified vector representations of heterogeneous data sets
• Utilize: Reference correspondence information between data pairs, cluster correspondence, etc.

Two conflicting criteria: maximize alignment and minimize deformation

Our methods:
Orthogonal Procrustes analysis, Graph Embedding
Space Alignment by Orthogonal Procrustes

$$\min \| (A-\mu_A 1^T) - kQ(B-\mu_B 1^T) \|_F$$, where $Q^TQ=I$

Alignment of Different Dimension Reduction Results of Clustered Data

Reference Aligned Un-Aligned
Cluster Label Matching of Different Clustering Results

- Hungarian algorithm
- Maximizes the sum of diagonal elements in confusion matrix
- Can better recognize cluster correspondences
Cluster Alignment of Different Clustering Results

Reference

Un-Aligned

- InfoVis and VAST paper data set
- Help refine cluster results and obtain consensus clustering
Testbed Applications
UTOPIAN: User-driven TOPic modeling based on InterActive Nmf [Choo, Lee, Reddy, HP, 2013, TVCG, To appear]

- NMF is superior to pLSI or LDA in topic modeling: faster, results more consistent, less parameters to tune, intermediate results easier to understand and interact with
- Semi-supervised / Weakly-supervised NMF [Choo et al., DMKD, accepted sub. to rev.] allow more flexible interactions
- Rich interactions for user-driven clustering & topic modeling

SS-NMF:

\[
\min_{W \geq 0, H \geq 0} ||A - WH||_F^2 \\
+ ||(W - W_r)M_w||_F^2 \\
+ \sum ||(H - H_rD_H)M_H||_F^2
\]
NMF vs. LDA (Latent Dirichlet Allocation)

$A \sim WH$

W: Keyword-wise distribution of topics
H: Topic-wise distribution of documents

Topic membership changes (InfoVis/VAST paper data set):
* among 10 runs
* throughout iterations

For more results on NMF for topic modeling, visit the poster on Rank 2 NMF for Hierarchical Clustering, Kuang & Park, Monday.
Usage Scenario
Hyundai Genesis Car Review Data
Initial Result

t-SNE
Van der Maaten and Hinton 08

Modified t-SNE
pairwise distances shrunk within each cluster
Improves personalization and understandability via integrated visualizations of document retrieval and recommendation.

**VisIRR**
Interactive Visual Information Retrieval & Recommendation System for Large-scale Document Data
Visualization Example of Queried Set

Keyword query, ‘dimension reduction’

Clear topics

Query

Computational zoom-in
Recommendation Example

Preference-assigned item as ‘highly like’: ‘Enhancing the visualization process with principal component analysis to support the exploration of trends’

• Recommendations based on content, co-authorship, or citation
• Heat kernel based propagation algorithm
• Out-of-sample embedded vis in previously computed space
VisIRR Recommendation Demo
Concluding Remarks

• Interactive Analysis provides more meaningful solutions in many real-life applications
• Visual Analytics is an important discipline for understanding/analyzing large scale data
• True integration of both automated algorithms and interactive visualization is the key
• Practice of multi-disciplinary research and problem driven foundational research required

Thank you!