New Geometric Methods of Mixture Models for Interactive Visualization

Jia Li\textsuperscript{1}, Bruce Lindsay\textsuperscript{1}, Xiaolong (Luke) Zhang\textsuperscript{2}

\textsuperscript{1}Department of Statistics
\textsuperscript{2}College of Information Sciences and Technology
The Penn State University
Goals

• Develop theories and algorithms for revealing prominent geometric features of mixture density.
• Develop approaches to clustering, dimension reduction, and variable selection based on the geometry of mixture density.
• Develop interactive visualization systems empowered by a suite of statistical learning tools.
• Apply the statistical methods and visualization paradigm to meteorology data for weather prediction and engineering design data
Our Work

• Theories and algorithms
  • Modal EM algorithm for solving modes of mixture density.
  • Clustering methods based on mode association.
  • Variable selection based on the geometry of mixture density.
  • Two-way mixture model for high dimensional data.
• Visualization system design
  • A work-centered visual analytics model
  • Explored applications to meteorology data and engineering design data.
    • Preliminary evaluation: engineering design case
• Parallelization of data clustering algorithms
Model EM (MEM)

- Let a mixture density be $f(x) = \sum_{k=1}^{K} \pi_k f_k(x)$.
  - $x \in \mathcal{R}^d$
  - $\pi_k$ is the prior probability of mixture component $k$.
  - $f_k(x)$ is the density of component $k$.
- Given any initial value $x^{(0)}$, MEM solves a local maximum of the mixture by alternating two steps.
Mode Association Clustering (MAC)

- The MAC Algorithm

1. Form kernel density $f(x \mid S, \sigma^2) = \sum_{i=1}^{n} \frac{1}{n} \phi(x \mid x_i, D(\sigma^2))$, where $S = \{x_1, x_2, ..., x_n\}$.

2. Use $f(x\mid S, \sigma^2)$ as the density function. Use each $x_i$, $i = 1, 2, ..., n$, as the initial value in the MEM algorithm to find a mode of $f(x\mid S, \sigma^2)$. Let the mode identified by starting from $x_i$ be $M_\sigma(x_i)$.

3. Extract distinctive values from the set $\{M_\sigma(x_i), i = 1, 2, ..., n\}$ to form a set $G$. Label the elements in $G$ from 1 to $|G|$.

4. If $M_\sigma(x_i)$ equals the $k$th element in $G$, $x_i$ is put in the $k$th cluster.
Hierarchical Mode Association Clustering (HMAC)

- Gradually increase kernel bandwidth:
  \[ \sigma_1 < \sigma_2 < \sigma_3 \cdots \]
- Kernel density at level \( i \): \( f(x \mid S, \sigma_i^2) \)
  - \( \sigma_i \) no other density, fewer modes
- Starting points at level \( i \) are the modes acquired at the previous level \( i - 1 \).
- The hierarchy by design:
  \[ x_i \rightarrow M_{\sigma_1}(x_i) \rightarrow M_{\sigma_2}(M_{\sigma_1}(x_i)) \rightarrow \cdots \]
Geometry of Mixture Models

At level 3, merge the modes from level 2

At level 4, merge the modes from level 3
Cloud Map Segmentation
A Work-Centered Model for Visual Analytics
Visual Analytics System: LIVE

- Intrinsic structures
- New structures produced by algorithms
- User Interaction
  - Interaction with individual view graphs
  - Multiple view coordination
    - E.g., brushing tools, color mapping, etc.
  - Dynamic refining inputs and parameters of algorithms
Evaluation: Conceptual Ship Design

Design input variables:
- Length ($L$), Beam ($B$), Depth ($D$), Draft ($T$),
- Block Coeff ($C_B$), and Speed ($V_k$).

Design output variables:
- Transportation Cost ($TC$), Light Ship Weight ($LSM$)
  and Annual Cargo ($AC$).

Goal
- Minimize $TC$, minimize $LSM$, and maximize $AC$.

Constraints:
- $L/B \geq 6$
- $L/D \leq 15$
- $L/T \leq 19$
- $F_n \leq 0.32$
- $25,000 \leq DWT \leq 50,000$
- $Const_1 = T - 0.45DWT^{0.31} \leq 0$
- $Const_2 = T - (0.7D + 0.7) \leq 0$
- $Const_3 = 0.07B - GM_T \leq 0$

Multi-Objective Optimization (MOO)
Preliminary Result

• Our system can facilitate an iterative design optimization process.
  • Use our algorithm to identify similar design alternatives
  • Use our algorithm to discover the values of design inputs based on desired outputs
  • Control the process of data clustering and classification
    • Step-by-step vs. batch
Preliminary Result

• Our system can facilitate an iterative design optimization process.
  • Use our algorithm to identify similar design alternatives
  • Use our algorithm to discover the values of design inputs based on desired outputs
  • Control the process of clustering
    • Step-by-step vs. batch

• Challenges
  • Knowledge about clustering algorithms by domain experts
    • Validation
  • Speed of clustering algorithms
    • Real-time interaction
Parallelization of HMAC

- Hadoop
- MPI

Image Data: 1,400 * 64
More Results

Ship design data: 2,000 * 17

Image Data: 1,400 * 64
Project Accomplishments

• Algorithms
  • Downloadable from our project website

• Visualization design
  • A work-centered model for visual analytics
  • A system prototype to support engineering design
    • Plan to build a system for meteorology data analysis
Selected Publications

Impact

• Training Ph.D. students
  • Three Ph.D. dissertations
    • Statistics, CSE, Information Sciences and Technology
  • Two other Ph.D. students involved

• Led to new projects
  • Health informatics (NSF –SHB, NIH)
  • Spatial-temporal data analysis (Industrial collaboration)

• Outreach
  • Invited session in Joint Statistical Meetings (JSM), 2010 (J. Li)
  • Invited panelist on the Panel of Visualization in the Annual Workshop of Human-Computer Interaction Consortium, 2010 (X. Zhang)
  • Invited talks
    • Institute of Software at Chinese Academy of Sciences, 2011 (X. Zhang)
    • Xerox Research Center Europe, 2012 (X. Zhang)
    • NSF EarthCube Workshop, 2012 (X. Zhang)