1. Motivation

- Many clustering applications involve time-varying data
- Tracking changes in community structure of a social network

2. Objectives

- To develop an evolutionary clustering algorithm specifically to deal with time-varying data
- To separate true evolutions from short-term variations due to noise?

3. Overview of spectral clustering

- Spectral clustering is a popular modern clustering technique inspired by spectral graph theory
- Create similarity graph of data points represented by affinity matrix \(W\)
- Use eigenvectors of scaled version of \(W\) to assigning points to clusters
- Can be interpreted as approximate solution for NP-hard problem of minimizing surface area (green edges) to volume (union of green and blue edges) ratio of each cluster

4. Evolutionary clustering framework

- Assume data are realizations from a mixture of random processes
- Let \(Y_t\) denote the affinity matrix at time \(t\)
- Define true affinity matrix \(\Phi^*\) at time \(t\) to be the expected affinity matrix \(E[\Phi(t)]\)
- Define smoothed affinity matrix \(\Phi^\alpha_0 = (1 - \alpha)\Phi_{t-1} + \alpha \Phi(t)\)
- Perform spectral clustering on \(\Phi^\alpha_0\) instead of \(\Phi(t)\) (Chakrabarti et al., 2006)
- \(\alpha\) controls trade-off between bias and variance of estimator
- A small \(\alpha\) is biased but has low variance
- A large \(\alpha\) is unbiased but has high variance

5. Selection of forgetting factor

- \(\alpha^*\) is not implementable (requires knowledge of \(\Phi^*\) and variance of \(\Phi(t)\))
- Replace unknowns by sample equivalents (Ledoit and Wolf, 2003)
- Problem: samples are not iid!
- Solution: iteratively estimate \(\alpha^*\) and component memberships (a) Fix component memberships to be cluster memberships from previous time step (b) Compute sample means and variances of each component to obtain estimate \(\alpha^*\) of \(\alpha\)
- (c) Fix \(\alpha^*\) and perform spectral clustering on \(\Phi^*\)
- (d) Repeat (b) and (c) until \(\alpha^*\) converges
- Re-estimate \(\alpha^*\) at each time step to achieve adaptive forgetting factor

6. Synthetic data experiment

- Mixture of two Gaussian components
- Slowly move one toward the other until they overlap
- Alter mixture proportion to simulate change in cluster

7. Real data experiment

- MIT reality mining data set (Eagle and Pentland, 2006)
- Measured proximity between people in study over 8-month period using cell phones with Bluetooth
- Sudden decreases in \(\alpha^*\) correlate with end of fall term and beginning of spring term
- Fixed \(\alpha^*\) result in trading off smoothing against change detection
- Adaptive \(\alpha^*\) provides effective smoothing during school terms and is also able to detect both change periods

8. Conclusions

- Proposed a method for adaptively selecting the forgetting factor in evolutionary spectral clustering
- Outperforms fixed forgetting factor and incremental clustering
- Future work: analysis of convergence and accuracy of iterative estimation

References