Evolutionary Clustering Methods

Hankui Peng

Supervisors: Nicos Pavlidis, Idris Eckley
STOR-i CDT, Lancaster University

EPSRC Engineering and Physical Sciences Research Council

STOR-i excellence with impact

Office for National Statistics

1 Introduction

What is evolutionary clustering?
Clustering is the process of grouping a set of unlabelled data objects (usually represented as a vector of measurements in a multidimensional space) into a number of clusters. The general objective of clustering is to obtain a partitioning of the data objects such that data within the same cluster are more similar to each other compared to data in different clusters. In some applications, we not only want to obtain static clustering results for one time step, but we are also interested in clustering data objects for an extended period of time. We want to make use of the clustering information from previous time steps to help produce consistent clustering results for the current time step. Ideally, we aim to produce interpretable and efficient clustering results for a set of data objects that evolve over time.

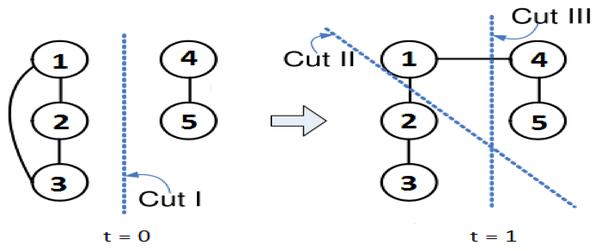

Figure 1: A simple evolutionary setting.

Motivation

- The Office for National Statistics (ONS) is interested in developing novel price indices using web-scraped sales data from three leading websites (Waitrose, TESCO, Sainsbury's) in the UK.
- The challenge come from the size and dimensionality of the data, the frequency of the data being scraped, and the mixed nature of the data (continuous and categorical).
- Consistent and efficient clustering results are desired for roughly the same set of data objects over an extended period of time.

2 Spectral Clustering

Spectral clustering builds a proximity matrix W based on the input features of a set of data objects, and solves a relaxed version of the graph cut problem [3]. We consider the data objects as a set of nodes, and the proximities among them as weighted edges. The relaxed problem is solved through making use of a graph Laplacian matrix L that can be constructed based on W. For the graph cut problem, two popular measures are:

- **Average Association (AA)**

$$AA = \sum_{c=1}^{k} \frac{\text{assoc}(V_c, V_c)}{|V_c|}, \tag{1}$$

- **Normalised Cut (NC)**

$$NC = \sum_{c=1}^{k} \frac{\text{assoc}(V_c, V \backslash V_c)}{\text{assoc}(V_c, V)}, \tag{2}$$

where we use $\text{assoc}(V_{c_1}, V_{c_2})$ to denote the edge weights that link set V_{c_1} and set V_{c_2}. Given a set of N data objects, the classic spectral clustering procedure is described as follows:

1. Build an $N \times N$ proximity matrix W that describes the closeness between different data objects. A common choice for similarity measure is the Gaussian similarity function

$$W(i,j) = \exp\left\{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma^2}\right\}. \tag{3}$$

2. Construct the unnormalised Graph Laplacian matrix as $L = D - W$, where D is a degree matrix that contains information about the edge weights attached to each node.

3. Conduct eigen-decomposition on L, and apply K-means clustering on the k eigenvectors associated with the k smallest eigenvalues.

3 Evolutionary Spectral Clustering (ESC)

Evolutionary spectral clustering conducts static spectral clustering at every time step for an extended period of time. Given that sensible clustering results are desired for roughly the same set of data objects throughout time, we want to incorporate the proximity information from previous time steps into the current time step. We define a cost function that we want to minimise at each time step as follows:

$$\text{Cost} = \alpha \cdot \text{SC} + (1-\alpha) \cdot \text{HC}, \tag{4}$$

where SC denotes the snapshot cost that measures the quality of the current clustering for the current data objects, and HC denotes the history cost that measures the quality of the current clustering applied to the data objects in the previous time step. We introduce two frameworks that differ in how they calculate the history cost: preserving cluster quality (PCQ) and preserving cluster membership (PCM).

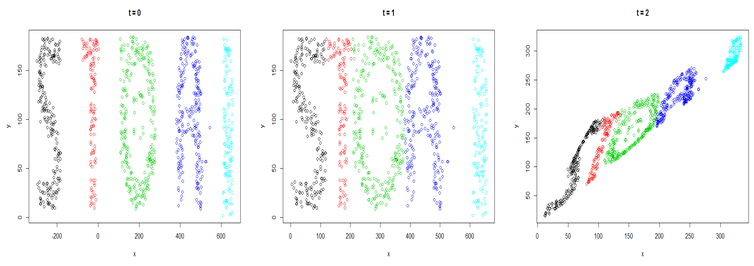

Figure 2: A top example that shows the evolvment of the same set of data objects over 3 time steps.

3.1 Preserving Cluster Quality (PCQ)

The PCQ framework focuses on how well the partition for current time step t clusters the data at the previous time step $t-1$. Static spectral clustering is applied at every time step t based on the following adjusted proximity matrix:

$$\alpha \cdot W_t + (1-\alpha) \cdot W_{t-1}. \tag{5}$$

- The full proximity matrix from the previous time step needs to be stored.
- Could potentially be extended to contain proximities from more previous time steps.

3.2 Preserving Cluster Membership (PCM)

Let X_{t-1} denote an $N \times k$ matrix formed by the k eigenvectors from the previous time step $t-1$. The PCM framework directly takes the first k eigenvectors from the previous time step, and conduct spectral clustering based on the following adjusted proximity matrix:

$$\alpha \cdot W_t + (1-\alpha) \cdot X_{t-1} X_{t-1}^T. \tag{6}$$

- Only the k eigenvectors of the proximity matrix from the previous time step is needed.
- A rotation-invariant distance is defined between the set of eigenvectors at time step t and $t-1$:

$$\text{dist}(X_t, X_{t-1}) = \frac{1}{2}\left\|X_t X_t^T - X_{t-1} X_{t-1}^T\right\|^2. \tag{7}$$

4 Experiments & Results

- Set up two well separated Gaussian datasets as shown in Fig. 2 (left).
- Move the bottom left set of data objects towards the upper right set.
- Apply static and evolutionary spectral clustering (PCQ) on this evolving set of data over 50 time steps.
- Compare the performance of these different methods in terms of the snapshot cost and misclassification error by using the ground truth .

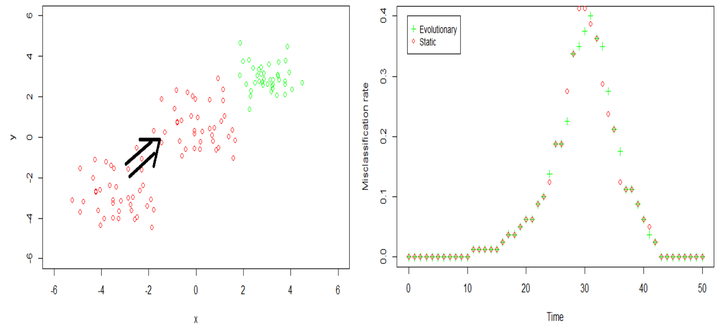

Figure 3: Evolutionary spectral clustering on the colliding Gaussian dataset.

5 Conclusions & Future Work

- Both PCQ and PCM framework only include the clustering information from one previous time step, a clustering scheme which includes information from all previous time steps is desired.
- We will further consider reasonable ways to construct proximity matrix when the data objects contain mixed data types (both continuous and categorical).
- More computationally efficient approaches as a workaround for the eigendecomposition step will be further investigated.

References

[1] Yun Chi, Xiaodan Song, Dengyong Zhou, Koji Hino, and Belle L Tseng. Evolutionary spectral clustering by incorporating temporal smoothness. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2007.

[2] Rocco Langone, Marc Van Barel, and Johan AK Suykens. Efficient evolutionary spectral clustering. *Pattern Recognition Letters*, 84:78–84, 2016.

[3] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.