iCluster (vanilla)

Stephen Coleman

09/01/2020

Assumptions

- Data is in the form $X = (x_1, \ldots, x_n)$, where $x_i = (x_{i1}, \ldots, x_{ip})^T$ (p genes and n samples);
- X is mean-centred.

Model

Shen, Olshen, and Ladanyi (2009) created iCluster to partition the columns of a dataset. They imagined the application in the context of clustering samples). In brief, iCLuster combines Factor Analysis and K-means clustering.

K-means clustering

Given a partition $C = (c_1, \ldots, c_n)$ where $c_i \in \{1, \ldots, K\}$, and K associated mean vectors $\mathbf{m} = (\mathbf{m}_1, \ldots, \mathbf{m}_K)$, the samples are paritioned such that the sum of within-cluster squared distances is minimised:

$$\min \sum_{k=1}^{K} \mathbb{I}(c_i = k) ||x_i - \mathbf{m}_k||^2$$

The method then updates the mean vectors before updating the partition. When the means stabilise the model is considered to have converged to some local minimum.

This method is highly sensistive to the choice of initial starting points for $\mathbf{m}$. Zha et al. (2002) show that applying PCA first can improve the performance of K-means clustering.

PCA K-means clustering

Let $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z_K})^T$, with $\mathbf{z}_k$ being the normalised indicator vector of cluster k, i.e. if n_k is the number of samples assigned to cluster k, then:

$$\mathbf{z}_k^T = (0, \ldots, 0, \frac{1}{\sqrt{n_k}}, \ldots, \frac{1}{\sqrt{n_k}}, 0, \ldots, 0)$$

References

Shen, Ronglai, Adam B Olshen, and Marc Ladanyi. 2009. “Integrative Clustering of Multiple Genomic Data Types Using a Joint Latent Variable Model with Application to Breast and Lung Cancer Subtype Analysis.” *Bioinformatics* 25 (22). Oxford University Press: 2906–12.

Zha, Hongyuan, Xiaofeng He, Chris Ding, Ming Gu, and Horst D Simon. 2002. “Spectral Relaxation for K-Means Clustering.” In *Advances in Neural Information Processing Systems*, 1057–64.