An Introduction to Some Popular Clustering Methods

Teck Por Lim

12 Aug 2015

Outline of the Talk

1. Introduction
2. Partition Based Clustering
3. Hierarchical Clustering

Clustering

- Type of unsupervised learning
- Seeks to group data into subsets
- Typically points within the same cluster are more closely related to each other than other points

Terms

- Partition
  - Given a non-empty set A, a partition of A is a collection of disjoint subsets of A whose union is A
- Clustering
  - Division of data into groups of similar objects [Berkhin(2006)]
  - Definition surprisingly useful, as it encompasses both hard partitions and soft partitions
- Unsupervised
  - No labelled data are available [Xu and Wunsch II(2009)]
- Dissimilarities
  - Typically, dissimilarities are metrics
  - Common example is the Euclidean distance

Applications

- Classification / taxonomy [Everitt et al.(2001)Everitt, Landau, and Leese]
  - Psychology: personality types
  - Astronomy: star types
- Bioinformatics
  - Finding co-expressed genes from microarray data
- Business analytics
  - Grouping customers based on consumption patterns
  - Provide customized marketing strategies to each group

Biclustering of Microarray Data

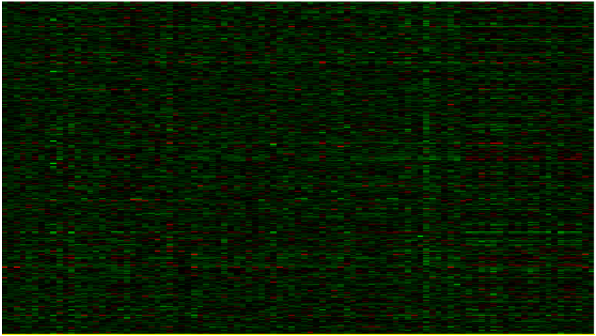

Figure: Biclustering of breast cancer microarray data [Hoshida et al.(2007)Hoshida, Brunet, Tamayo, Golub, and Mesirov] using the algorithm of Cheng and Church [Cheng and Church(2000)]. Rows of data matrix are 1213 genes, whilst the columns are the 97 samples. Colours range from bright green (negative, under-expressed) to bright red (positive, over-expressed).

Classification of Some Clustering Algorithms [Berkhin(2006)]

- Partitioning Methods
  - k-means
  - k-medoids
- Fuzzy Partitioning Methods
- Hierarchichal Methods
  - Agglomerative
  - Divisive
- Density based alogithms

k-means

The first versions of the k-means algorithm are attributed to Lloyd [Lloyd(1957)] and Forgy [Forgy(1965)]. The algorithm converges to a local optimum because both types of steps optimize the within-cluster sum of squares (WCSS) objective.

1. Each data point is assigned to closest centroid, with ties broken arbitrarily
2. The centroid positions are recomputed, based on the new memberships

Problems with k-means [Berkhin(2006)]

- Results dependent upon intialization of centroids
- Computed local optimum may be far from the global optimum
- Not obvious what value of k to use
- Process is sensitive to outliers
- Algorithm lacks scalability
- Only numerical data can be clustered
- Resulting clusters can be unbalanced

Clustering via Expectation-Maximization (EM)

- k-means is a limiting case of fitting data by a mixture of k Gaussians with identical, isotropic covariance matrices
- Soft assignment of data points to mixture components are hardened to label each data point using the most likely component.
- If data does not consist of well separated spherical clouds, k-means can have problems.
- EM clustering allows for "ellipsoidal" clouds of data
- Latent variable is class label
- Expectation step (E step): Calculate expected value of the log likelihood function
- Maximization step (M step): Find parameter that maximizes this quantity

Kernel k-means

- Kernel based methods enable us to deal with clusters which are not linearly separated
- The intuition behind the kernel method is to transform the original problem into a linearly separable problem
- Input data points are mapped nonlinearly into feature space via kernel function
- Kernel $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathfrak{R}$ measures similarity between any pair of inputs $\mathbf{x}, \mathbf{c} \in \mathcal{X}$
- For the often used RBF kernel, $K(\mathbf{x}, \mathbf{c}) = exp(-\frac{\|\mathbf{x}-\mathbf{c}\|^2}{2\sigma^2})$

Finding *k*

- There are many methods for finding the number of clusters in an automated manner
- One can for instance try
  - Bootstrapping approach (fpc::clusterboot)
  - Bayesian approach

Hierarchical Clustering

- Does not depend on random number seed
- Can be computationally complex, but there are efficient variants [Murtagh(1983)]
- Agglomerative
  - Each observation starts in its own cluster
  - Pairs of clusters are merged as one moves up the hierarchy
- Divisive
  - All observations start in one cluster
  - Splits are performed recursively as one moves down the hierarchy

Linkage criteria

Linkage	Formula
complete	$\max\{d(a,b) : \ a \in A, \ b \in B\}$
single	$\min\{d(a,b) : \ a \in A, \ b \in B\}$
average	$\frac{1}{\lvert A \rvert \lvert B \rvert} \sum_{a \in A} \sum_{b \in B} d(a,b)$
centroid	$\lVert c_i - c_j \rVert$, where c_i and c_j are centroids of clusters i and j

Table: Some linkage criteria between two sets of observations A and B.

Bibliography I

P. Berkhin.
A survey of clustering data mining techniques.
In Jacob Kogan, Charles Nicholas, and Marc Teboulle, editors, *Grouping Multidimensional Data*, pages 25–71. Springer Berlin Heidelberg, 2006.

Y Cheng and G M Church.
Biclustering of expression data.
Proc Int Conf Intell Syst Mol Biol, 8:93–103., 2000.

Brian Everitt, Sabine Landau, and Morven Leese.
Cluster analysis.
Arnold, London, 4th edition, 2001.

Bibliography II

E. W. Forgy.
Cluster analysis of multivariate data: efficiency versus interpretability of classifications.
Biometrics, 21(3):768–769, 1965.

Yujin Hoshida, Jean-Philippe Brunet, Pablo Tamayo, Todd R. Golub, and Jill P. Mesirov.
Subclass mapping: Identifying common subtypes in independent disease data sets.
PLoS ONE, 2(11):e1195, 2007.

S. Lloyd.
Least square quantization in pcm, 1957.

F. Murtagh.
A survey of recent advances in hierarchical clustering algorithms.
The Computer Journal, 26(4):354–359, 1983.

Bibliography III

Rui Xu and Donald C. Wunsch II.
Clustering.
IEEE Press, Piscataway, N.J., 2009.

Acknowledgements

Many thanks to Vik Gopal and Alex You for helpful comments