Comparing Unsupervised Methods For High-Dimensional Data

Parag, Apurva, Sonia, Anugna

University of Massachusetts Dartmouth

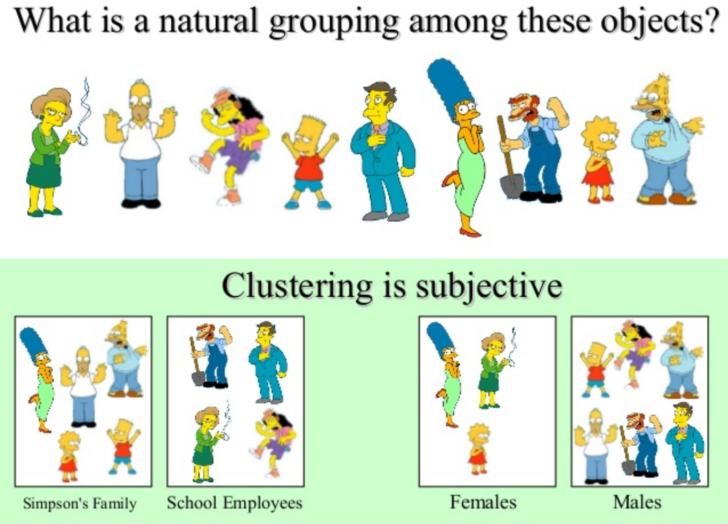
May 3, 2017

Clustering

1. Clustering is one of the most important unsupervised learning methods, which deals with finding a structure in a collection of unlabeled data.

2. Hence, clusters are collection of objects which are similar between them and are dissimilar to the objects belonging to other clusters.

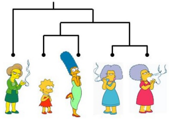
Clustering


What is a natural grouping among these objects?
Clustering is subjective
Simpson's Family
School Employees
Females
Males

Types of Clusters

Types of Cluster

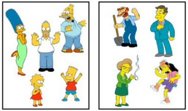
Hierarchical



It creates a hierarchical decomposition of the set of objects using some criterion

It constructs various partitions & evaluate by some criterion.

Partitional



Methods

- Partition
  -K-Means Clustering
- Hierachical
  -Agglomerative(Bottom-up)
- Spectral Clustering

Datasets

Iris Identification

Instances:150, Attributes: 4

Yeast Identification

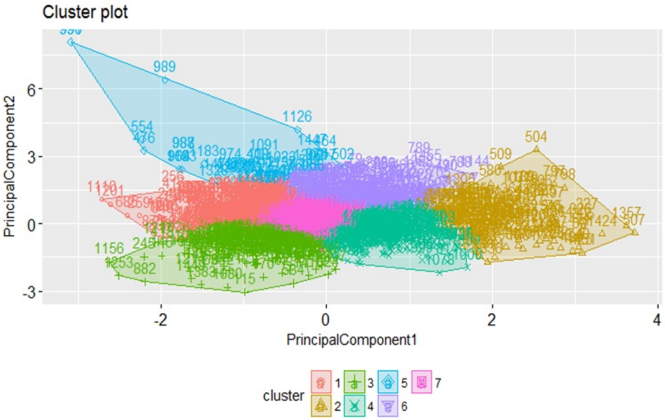
Instances:1484. Attributes: 9

Glass Identification

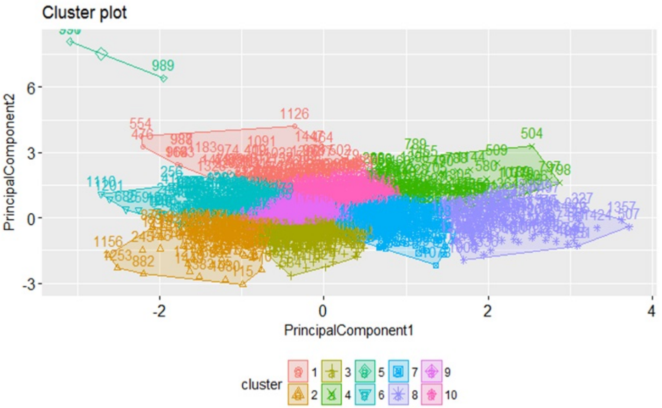
Instances:214. Attributes: 10

Iris Identifictaion

Glass Identificataion



Yeast Identificataion

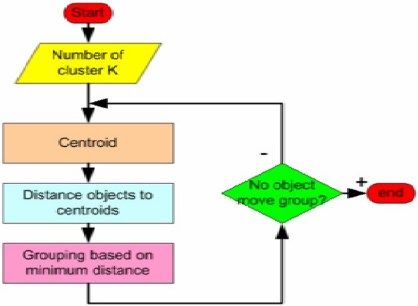


K-means

***Algorithm:** works with numeric data only

1. Select a number (K) of cluster centers (at random)
2. Assign every item to its nearest cluster center (e.g. using Euclidean distance)
3. Move each cluster center to the mean of its assigned items
4. Repeat steps 2,3 until convergence (change in cluster assignments less than a threshold)

   The centroid is (typically) the mean of the points in the cluster.

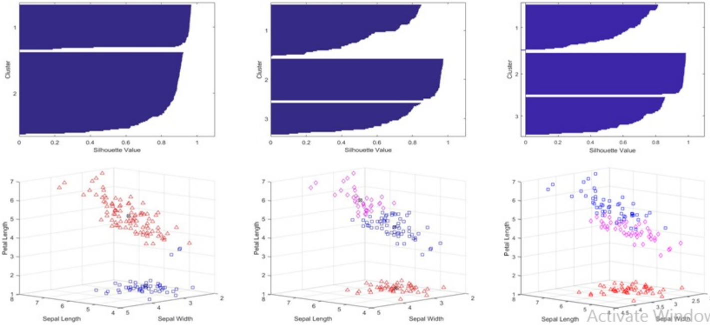


Euclidean Distance

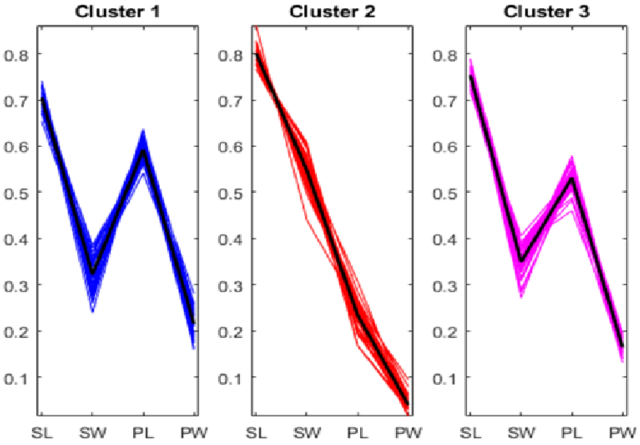
- The Euclidean distance or Euclidean metric is the distance between two points in Euclidean space

$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1-p_1)^2 + (q_2-p_2)^2 + \cdots + (q_n-p_n)^2}$$
$$= \sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}.$$

Centroids: Iris

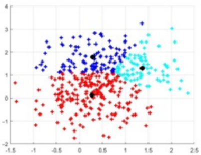


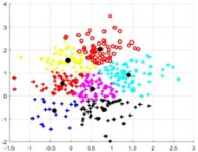
Iris



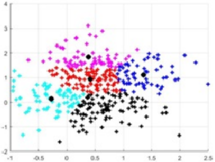
Centroids: Glass

3,2





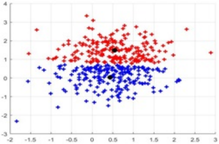
7,9



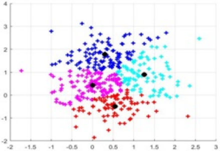
5,5

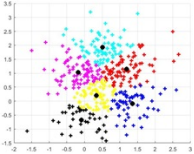
Centroids: Yeast

2,1



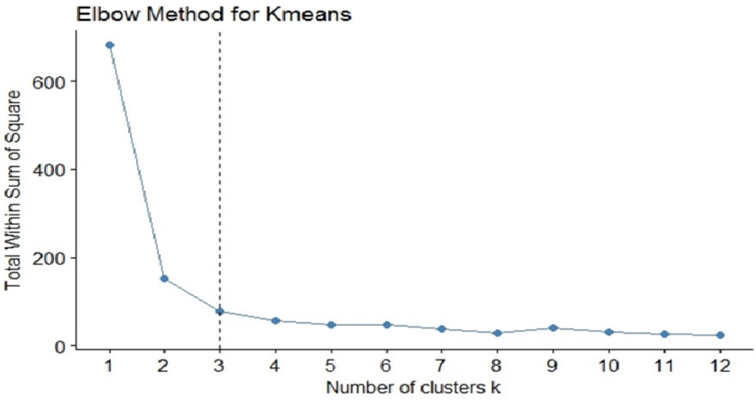
4,6





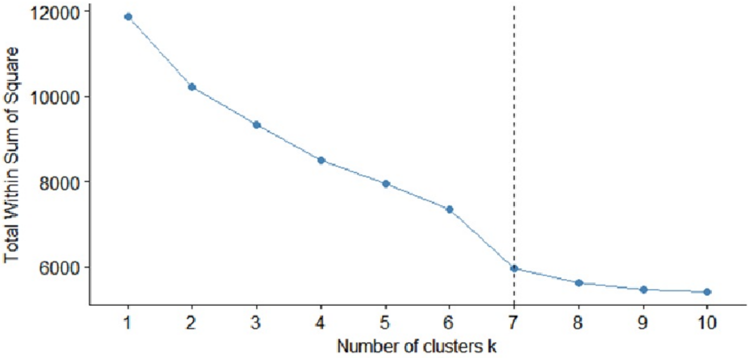
6,8

Elbow Method: Iris



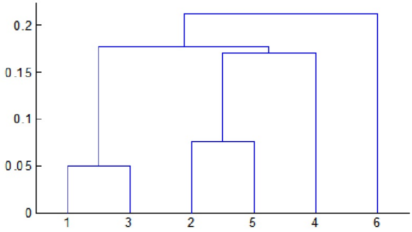
Elbow Method: Glass

Elbow Method: Yeast



Hierarchical Clustering

1. Produces a set of growing clusters organized as a hierarchical tree
2. Can be visualized as a dendrogram A tree -like diagram that records the sequences of merges or splits
3. Types: 1)Agglomerative , 2)Divisive



Agglomerative Algorithm:

1. Compute the distance matrix between the input data points
2. Let each data point be a cluster
3. Repeat
4. Merge the two closest clusters
5. Update the distance matrix
6. Until only a single cluster remains

Key operation is the computation of the distance between two clusters

Closest Pair of Clusters

Many Variants to Defining Closest Pair of Clusters

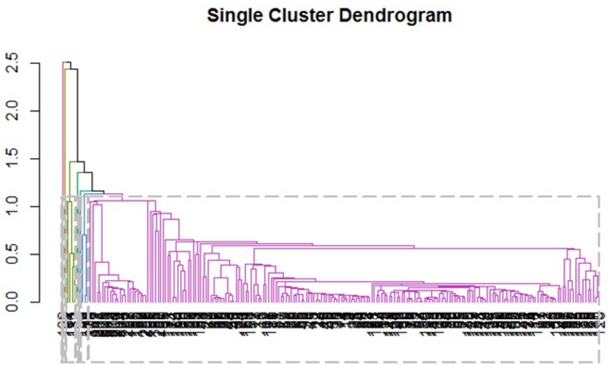
1. SINGLE-LINK: Distance of the Closest points.
2. COMPLETE-LINK: Distance of the Furthest points.
3. AVERAGE-LINK: Average distance between pairs of Elements.

Single-Link Clustering

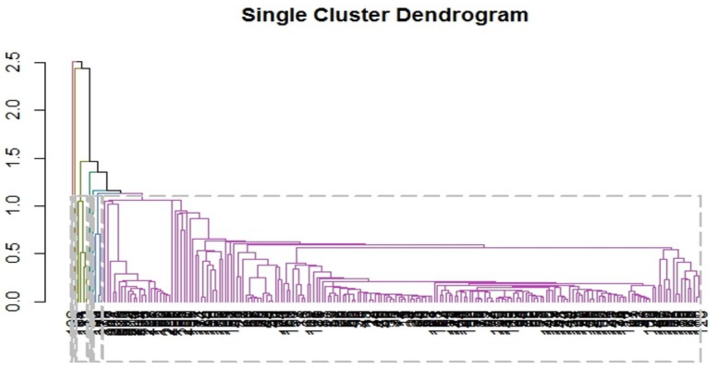
- Single-link distance between clusters C_i and C_j is the minimum distance between any object in C_i and any object in C_j
- The method is also known as nearest neighbor clustering
- Type equation here.

$$D_{sl}(C_i, C_j) = \min_{x,y} \{ d(x,y) \mid x \in C_i, y \in C_j \}$$

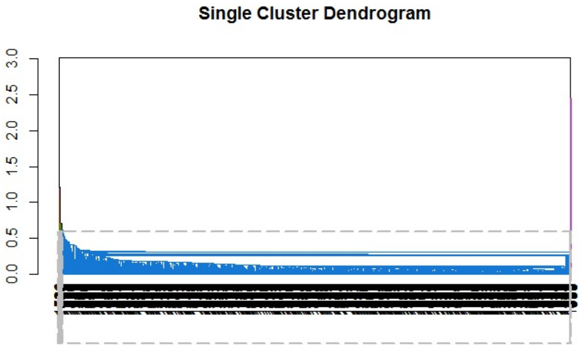
Iris Dataset



Glass Dataset



Yeast Dataset



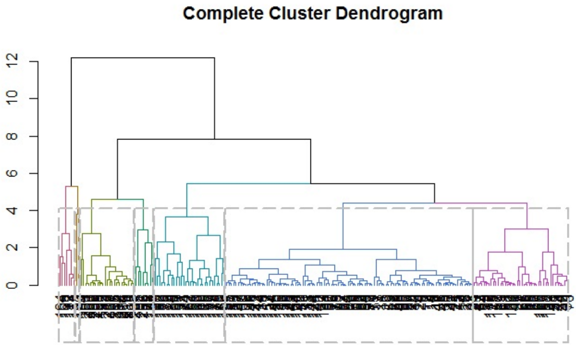
Complete-linkage clustering

Complete-link distance between clusters Ci and Cj is the maximum distance between any object in Ci and any object in Cj

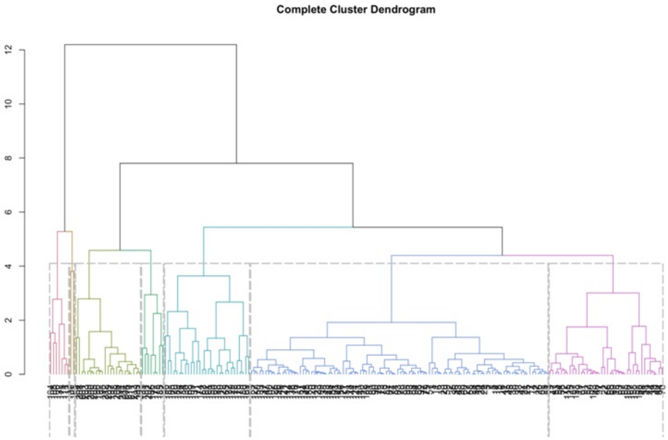
The method is also known as farthest neighbor clustering

$$D_{cl}(C_i, C_j) = \max_{x,y}\{d(x,y) \mid x \in C_i, y \in C_j\}$$

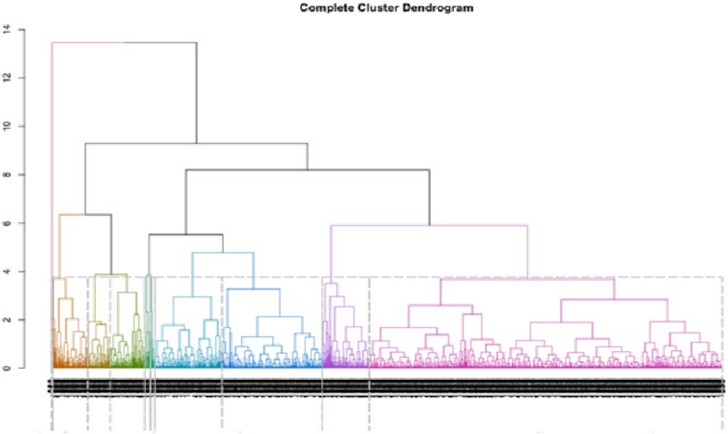
Iris Dataset



Glass Dataset



Yeast Dataset

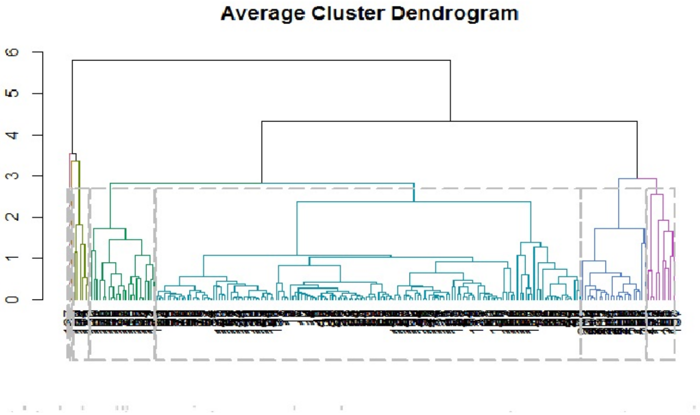


Average-linkage clustering

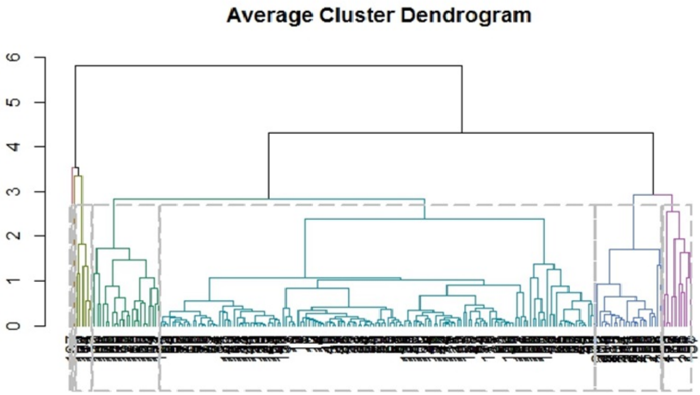
Group average distance between clusters Ci and Cj is the average distance between any object in Ci and any object in Cj

$$D_{avg}(C_i, C_j) = \frac{1}{|C_i| \times |C_j|} \sum_{x \in C_i, y \in C_j} d(x, y)$$

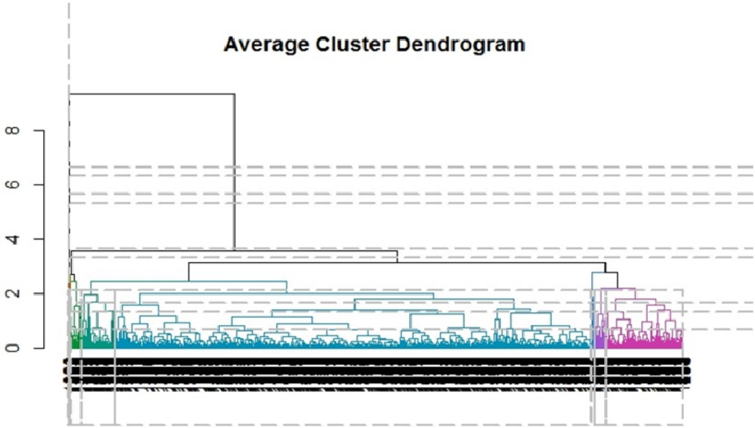
Iris Dataset



Glass Dataset



Yeast Dataset



Spectral Clustering

- The idea in spectral clustering is to construct similarity graphs that repre the local neighborhood relationships between observations.

- General steps of spectral clustering:

  -- finds the m eigenvectors Z_{N*m} corresponding to the m smallest eigenvalues of L (ignoring the trivial constant eigenvecto

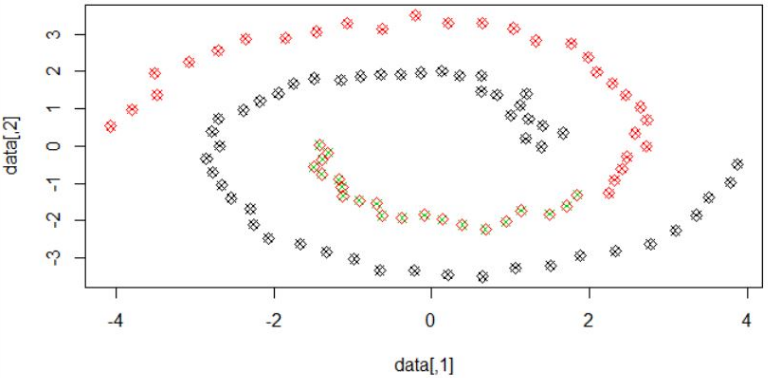
Algorithm

- Obtain data representation in the low-dimensional space that can be easily clustered
- Use k eigenvectors (k chosen by user)
- Directly compute k-way partitioning
- Experimentally has been seen to be "better
- project your data into R^(n)
- define an Affinity matrix , using a Gaussian Kernel or say just an Adjacency matrix (i.e.)
- construct the Graph Laplacian from A (i.e. decide on a normalization)

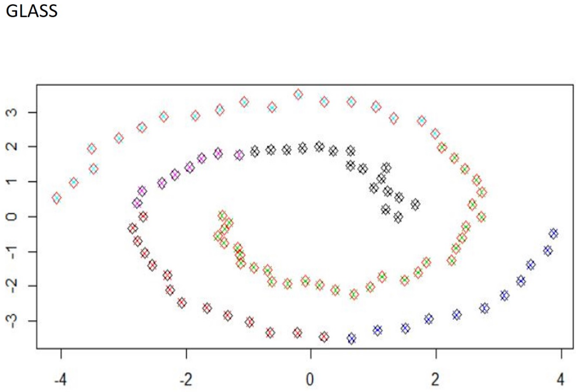
Continue......

- project your data into R^(n)
- define an Affinity matrix , using a Gaussian Kernel or say just an Adjacency matrix (i.e.
- construct the Graph Laplacian from A (i.e. decide on a normalization)
- solve an Eigenvalue problem (or a Generalized Eigenvalue problem)
- select k eigenvectors corresponding to the k lowest (or highest) eigenvalues , to define a k-dimensional subspace
- form clusters in this subspace using, say, k-means

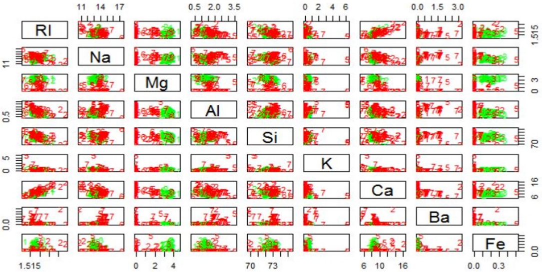
Spectral Iris:



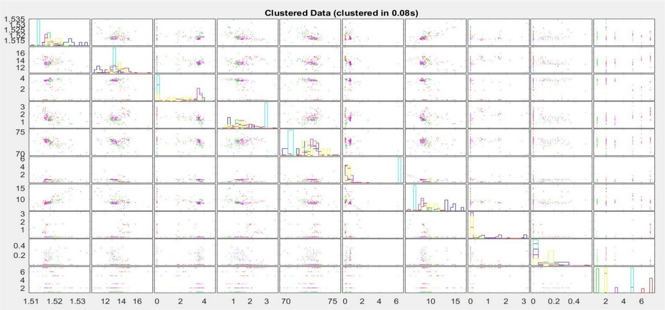
Spectral Glass:



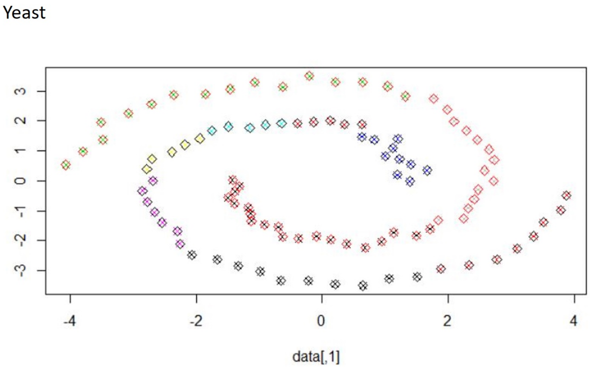
Glass Grid Graph



Glass Cluster grid



Yeast:



Issue:

1. Choice of k, the number of clusters
2. Choice of scaling factor
3. Realistically, search over and pick value that gives the tightest clusters
4. Choice of clustering method

Cluster Evaluation:

The quality of a clustering is very hard to evaluate because We do not know the correct clusters Some methods are used:

User Inspection:

* Study centroids, and spreads
* Rules from a decision tree

Cluster evaluation: evaluation based on internal information

1. Intra-cluster cohesion (compactness):
Study centroids, and spreadsCohesion measures how near the data points in a cluster are to the cluster centroid. Sum of squared error (SSE) is a commonly used measure.

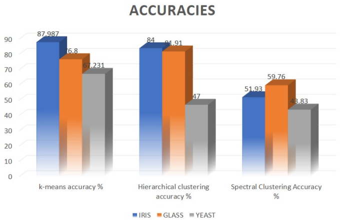
2. Inter-cluster separation (isolation):
Separation means that different cluster centroids should be far away from one another. In most applications, expert judgments are still the key

Result:

* Accuracy
* Error Rate

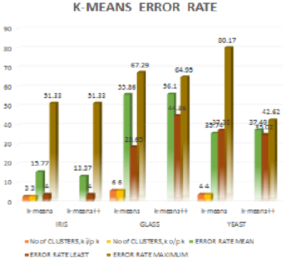
Accuracy

Dataset	K-means Accuracy(%)	Hierarchical clustering Accuracy(%)	Spectral Clustering Accuracy(%)
Iris	87.987	84	51.93
Glass	76.8	81.91	59.76
Yeast	67.231	47	43.83



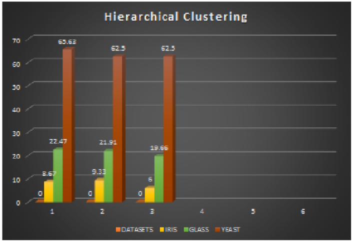
K-Means

DATASET	ALGORITHM	No of CLUSTERS,k		ERROR RATE		
		i/p k	o/p k	MEAN	LEAST	MAXIMUM
IRIS	k-means	3	3	15.77	4	51.33
	k-means++			13.37	4	51.33
GLASS	k-means	6	6	55.86	28.65	67.29
	k-means++			56.1	44.86	64.95
YEAST	k-means	4	4	35.74	37.38	80.17
	k-means++			37.49	35.02	42.62



Hierarchical

DATASETS	SINGLE	COMPLETE	AVEREAGE
IRIS	8.67	9.33	6
GLASS	22.47	21.91	19.66
YEAST	65.63	62.5	62.5



Spectral

DATASETS	ERROR RATES
IRIS	42
GLASS	62
YEAST	72.75

Conclusion:

1. The three Algorithms are compared as:
* The Size of Dataset
* No. of Clusters
* Type of Dataset
* Type of Software

2. Inter-cluster separation (isolation):The performance of k-means Spectral algorithm is better than Hierarchical clustering algorithm.

* K-means and Spectral have less accuracy.

* Partitional Algorithm is recommended for Huge Dataset(better result).

* Hierarchical clustering is recommended for Small Datasets.

References

Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel (2012)
OPTICS: Ordering Points To Identify the Clustering Structure. In Proceedings of the International Conference on Management of Data
SIGMOD , volume 28(2) of SIGMOD Record, pages 4960, Philadelphia, PA.USA, 13 June 1996.ACM Press .

Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, and Prabhakar Raghavan
Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications.
SIGMOD In Proceedings of the International Conference on Management of Data, (), volume 27(2) of SIGMOD Record, pages 94105, Seattle,WA, USA, 14 June 1998. ACM Press.

Michael R. Anderberg. M. 1973.
Cluster Analysis and Applications..
Academic Press New York.

www.wikipedia.com

Thank You