Clustering

Luca Citi
lciti@essex.ac.uk

School of Computer Science and Electronic Engineering
University of Essex (UK)

CE802

Outline

- Clustering
  - Clustering
  - Agglomerative hierarchical clustering
  - K-means method
  - Overview clustering methods

CLUSTERING

THE TASK

Given a set of ***unclassified*** training examples:

- Find a good way of partitioning the training examples into classes.
- Construct a representation that enables the class of any new example to be determined.

Although the two subtasks are logically distinct, they are usually performed together.

Terminology

Statisticians call this ***clustering***.

Neural net researchers usually call it ***unsupervised learning***.

THE BASIC PROBLEM

Classification learning programs are successful if the predictions they make are correct.

i.e. If they agree with an externally defined classification.

In clustering, there is no externally defined notion of correctness.

There are a huge number of ways in which a training set could be partitioned.

Some of these are better than others.

What do we mean by a good partition?

PARTITIONING CRITERIA

Common sense suggests members of a class should resemble each other more than resemble members of other classes.

Hence a good partition should:

- Maximise similarity within classes
- Minimise similarity between classes.

N.B. This implies the existence of a similarity metric – c.f. instance based learning.

Is this enough to identify good partitions?

No.

Consider the partitioning in which every item is assigned to its own class.

Such a partition would be of no use.

This suggests a further criterion:

- Minimise the number of classes created.

Clearly there will be a trade off between this and the other criteria.

Why Do We Want To Form Classes?

What do we gain by assigning two examples to the same class?

One important reason for grouping individuals into classes is that being told the class of an item conveys a lot of information about it.

Example:

Suppose I tell you that Fido is a dog:

Immediately you are reasonably confident of the following:

Fido had four legs

Fido barks

Fido has sharp teeth

Fido probably chases cats

etc

Thus we could also define a good partition as one that:

- Maximises the ability to predict unknown attribute values from class membership

AGGLOMERATIVE HIERARCHICAL CLUSTERING

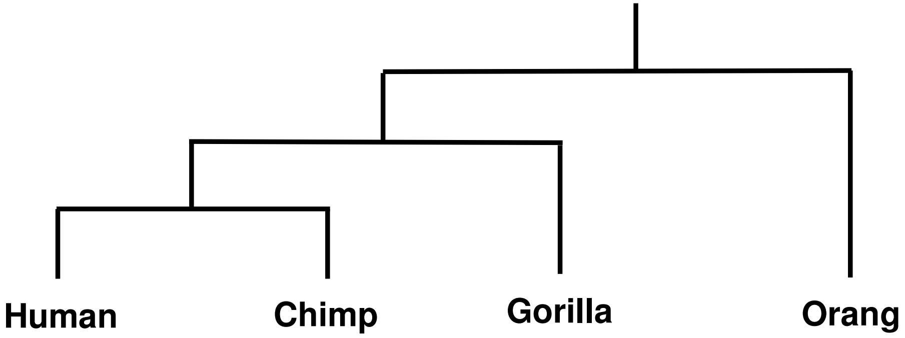
A family of methods.

Basic Idea

```
Assign each example to its own cluster.
WHILE there are at least two clusters
   Find the most similar pair of clusters
   Merge them into a new larger cluster
```

Results are usually presented as a tree called a dendogram.

e.g.



Dendogram for great apes using DNA as similarity metric.

This approach

- Requires a similarity metric that can determine the distance between groups.
- Requires all examples to be available at the start
- Requires the human analyst to decide on the optimal number of classes.

K-Means Method

- An iterative distance based method
- Only suitable for numeric data sets.
- User must specify how many clusters should be formed.

```
k = number of clusters to be formed;
Choose k items randomly as cluster centres;
Set initial cluster centroids to the k
items;
REPEAT
    Assign each item to cluster whose
    centroid is closest to it;
    Update cluster centroids to mean value
    for all items currently in that cluster;
UNTIL no item changes clusters
```

Number of iterations needed will depend on how well formed the clusters are.

Compact well separated clusters will converge rapidly.

Limitations

- May converge on a local maximum
- Only suitable for convex clusters

Numerous elaborations of the basic k-means method have been developed.

Evaluating Clusterings

A program such as k-means will find k clusters in any set of data, even if there is no such structure present.

Thus we need a way to measure how well formed the clusters that have been discovered actually are.

This requires us to develop metrics based on the criteria discussed earlier:

Cohesion – Similarity of items within a cluster

Separation – Difference between items in different clusters

We would also like our metrics to provide overall measures of the quality of the entire clustering scheme.

One approach is to consider:

Distances between items in a cluster

Distances between clusters

Distances between items within a cluster

Suppose there are K clusters

Let C_i and C_j be clusters and let c_i and c_j be their centroids.

Define the cohesion of a single cluster, C_i , as :

$$\text{cohesion}(C_i) = \sum_{x \in C_i} \text{dist}(c_i, x)^2$$

where $\text{dist}(c_i, x)$ is the Euclidean distance between c_i and x.

The smaller the cohesion, the more the members of the cluster resemble one another.

Distances between clusters

Define the separation of two clusters, C_i and C_j, as:

$$\text{separation}(C_i, C_j) = \text{dist}(c_i, c_j)^2$$

Overall Cohesion

Summing the cohesion over all clusters gives the overall cohesion:

$$\mathrm{overallcohesion} = \sum_{i=1}^{K} \sum_{x \in C_i} \mathrm{dist}(c_i, x)^2$$

This is a measure of how much of the variability is within clusters.

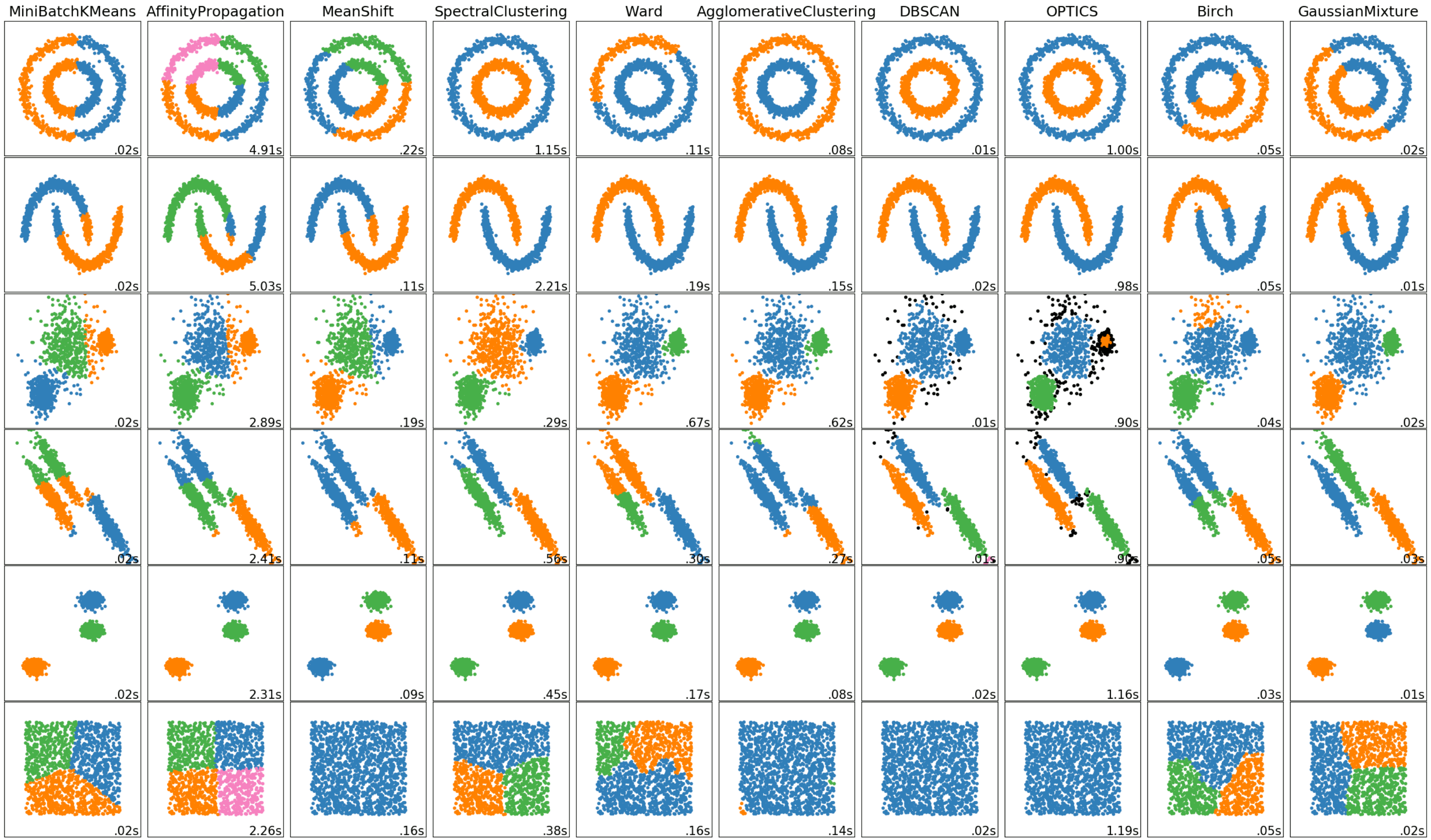
It can be compared with the amount of variability in the entire data set which is defined as:

$$\sum_{i=1}^{K} \sum_{x \in C_i} \mathrm{dist}(c, x)^2$$

where c is the centroid of the entire data set

Overview of clustering methods

There are a number of clustering methods with different properties



From `https://scikit-learn.org/stable/modules/clustering.html`

Outline

- Clustering
  - Clustering
  - Agglomerative hierarchical clustering
  - K-means method
  - Overview clustering methods

References

Required course material reading:

Scott's notes on Clustering

pp. 1–8!

From http://stats.stackexchange.com/questions/133656/how-to-understand-the-drawbacks-of-k-means

Q&A