

# Data Mining Algorithms

## Cluster Analysis

Graham Williams

Principal Data Miner, ATO  
Adjunct Associate Professor, ANU



## Overview

### Cluster Analysis

Introduction  
Requirements

### Measuring Similarity

Distances  
Data Types

### Algorithms

Cluster Methods  
KMeans



## What is Cluster Analysis?

- 1 How do we understand the behaviour of an individual?
  - Paint everyone with the same brush;
  - Treat everyone as an individual.
- 2 How do we understand the world—through understanding every individual in the world?
- 3 We categorise, for *good or bad*, entities into groups:
  - Socio-economic groups: “the poor”, “the rich”;
  - Political: a lefty, a new right;
  - Racial: religious, geographical,
- 4 We find that to get through in life we generally talk about groups, not individuals, but computers don’t need to—they have the power to build an understanding of the individual, for *better or worse*.



## What is Cluster Analysis?

- **Cluster**: a collection of data objects
  - Similar to one another within the same cluster
  - Dissimilar to the objects in other clusters
- **Cluster analysis**
  - Grouping a set of data objects into clusters
- Clustering is **unsupervised classification**: no predefined classes—descriptive data mining.
- Typical applications
  - As a stand-alone tool to get insight into data distribution
  - As a preprocessing step for other algorithms



## General Applications of Clustering

- Pattern Recognition
- Spatial Data Analysis
  - create thematic maps in GIS by clustering feature spaces
  - detect spatial clusters and explain them in spatial data mining
- Image Processing
- Economic Science (especially market research)
- WWW
  - Document Classification
  - Question Categorisation
  - Weblog Access Patterns



## Specific Examples

- **Marketing**: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs  
Land use: Identification of areas of similar land use in an earth observation database
- **Insurance**: Identifying groups of motor insurance policy holders with a high average claim cost
- **City-planning**: Identifying groups of houses according to their house type, value, and geographical location
- **Earth-quake studies**: Observed earth quake epicenters should be clustered along continent faults



# What Is Good Clustering?

- **High Quality:**
  - high intra-class similarity
  - low inter-class similarity
- Depends on:
  - similarity measure
  - algorithm for searching
- Ability to discover hidden patterns

# Clustering Caveats

Clustering may not be the best way to discover interesting groups in a data set. Often visualisation methods work well, allowing the human expert to identify useful groups. However, as the data set sizes increase to millions of entities, this becomes impractical and clusters help to partition the data so that we can deal with smaller groups. Different algorithms deliver different clusterings.

# Requirements of Clustering in Data Mining

- Scalability
- Different attribute types
- Clusters with arbitrary shape
- Minimal domain knowledge required
- Can cope with noise and outliers
- Insensitive to order of input records
- High dimensionality

# Overview

- Cluster Analysis*
  - Introduction
  - Requirements
- Measuring Similarity*
  - Distances
  - Data Types
- Algorithms*
  - Cluster Methods
  - KMeans

# Similarity and Dissimilarity Between Objects

- Distance measures the **similarity** or **dissimilarity** between two data objects  $a = (a_1, a_2, \dots, a_p)$  and  $b = (b_1, b_2, \dots, b_p)$ .
- Properties
  - $d(a, b) \geq 0$
  - $d(a, a) = 0$
  - $d(a, b) = d(b, a)$
  - $d(a, b) \leq d(a, c) + d(c, b)$

# Minkowski distance

$$d(a, b) = \sqrt[q]{(|a_1 - b_1|^q + |a_2 - b_2|^q + \dots + |a_p - b_p|^q)}$$

- If  $q = 1$ ,  $d$  is the **Manhattan distance**.

$$d(a, b) = |a_1 - b_1| + |a_2 - b_2| + \dots + |a_p - b_p|$$

- If  $q = 2$ ,  $d$  is **Euclidean distance**:

$$d(a, b) = \sqrt{(|a_1 - b_1|^2 + |a_2 - b_2|^2 + \dots + |a_p - b_p|^2)}$$

## Type of data in clustering analysis

- Interval-scaled variables
- Binary variables
- Nominal, ordinal, and ratio variables
- Variables of mixed types



## Overview

Cluster Analysis  
Introduction  
Requirements

Measuring Similarity  
Distances  
Data Types

Algorithms  
Cluster Methods  
KMeans



## Major Clustering Approaches

- Partitioning algorithms (kmeans, pam, clara, fanny): Construct various partitions and then evaluate them by some criterion. A fixed number of clusters,  $k$ , is generated. Start with an initial (perhaps random) cluster.
- Hierarchy algorithms: Create a hierarchical decomposition of the set of data (or objects) using some criterion
- Density-based: based on connectivity and density functions
- Grid-based: based on a multiple-level granularity structure
- Model-based: A model is hypothesized for each of the clusters and the idea is to find the best fit of that model



## Basic Partitioning Algorithm

- Partition database  $D$  of  $n$  objects into  $k$  clusters
  - Given  $k$ , find  $k$  clusters that optimises partitioning criterion
  - Global optimal: exhaustively enumerate all partitions
- Heuristic methods: k-means and k-medoids algorithms
  - k-means: Each cluster represented by center of the cluster
  - k-medoids or PAM (partition around medoids): Each cluster represented by one of the objects in the cluster

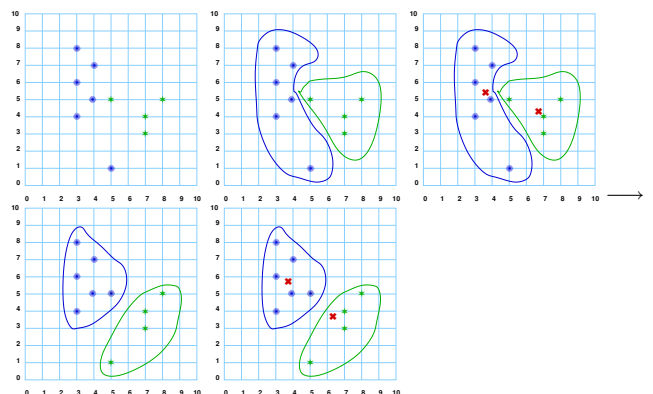


## The K-Means Clustering Method

- Given  $k$ , the k-means algorithm is implemented in 4 steps:
  - 1 Partition objects into  $k$  nonempty subsets
  - 2 Compute seed points as the centroids of the clusters of the current partition. The centroid is the center (mean point) of the cluster.
  - 3 Assign each object to the cluster with the nearest seed point.
  - 4 Go back to Step 2, stop when no objects change clusters.



## The K-Means Clustering Method



## Comments on K-Means

- Strengths
  - Relatively efficient:  $O(tkn)$ , where  $n$  is the number objects,  $k$  is the number of clusters, and  $t$  is the number iterations. Normally,  $k, t \ll n$ .



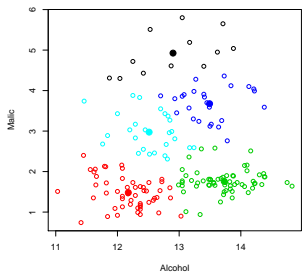
## Comments on K-Means

- Weakness
  - Often terminates at a local optimum. The global optimum may be found using techniques such as: deterministic annealing and genetic algorithms
  - Applicable only when the mean is defined—what about categorical data?
  - Need to specify  $k$ , the number of clusters, in advance.
  - Unable to handle noisy data and outliers.
  - Not suitable for non-convex clusters.

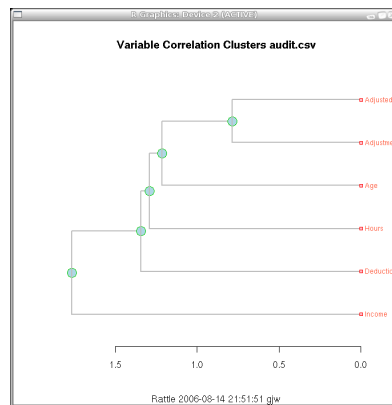


## KMeans in R

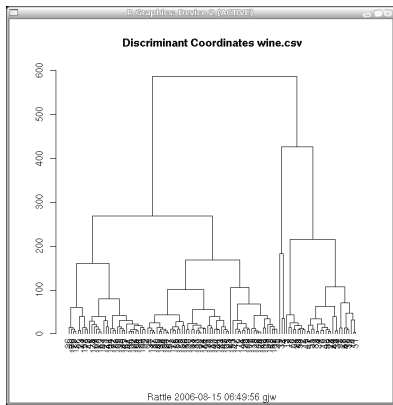
```
clusters <- 5
load("wine.Rdata")
wine.cl = kmeans(wine[,2:3], clusters)
plot(wine[,2:3], col=wine.cl$cluster)
points(wine.cl$centers, pch=19, cex=1.5, col=1:clusters)
dev.copy(device=pdf, file="wine-clusters.pdf")
dev.off()
```



## Rattle: Hierarchical Variable Cluster



## Rattle: Hierarchical Data Cluster



## Summary

- Cluster analysis is unsupervised learning.
- Useful for partitioning a very large population, perhaps for data mining each sub-population separately.
- Often more effective under expert guidance.

