Cluster Analysis
- Hierarchical Clustering

Sridhar P. Arjunan
School of Electrical and Computer Engineering,
RMIT University
GPO Box 2476V, Melbourne, VIC 3001,
Australia

s3099587@student.rmit.edu.au

22 Sep. 2006

RMIT

Outline

- Introduction - Clustering
- Hierarchical Clustering
- Dendrograms – Visualisation
- Clustering of Features of EMG Signal
- Example

s3099587@student.rmit.edu.au

RMIT

Introduction

- The term cluster analysis (first used by Tryon, 1939) encompasses a number of different algorithms and methods for grouping objects of similar kind into respective categories.
- A general question facing researchers in many areas of inquiry is how to organize observed data into meaningful structures,
- In other words cluster analysis is an exploratory data analysis tool which aims at sorting different objects into groups in a way that the degree of association between two objects is maximal if they belong to the same group and minimal otherwise.
- Given the above, cluster analysis can be used to discover structures in data without providing an explanation/interpretation. In other words, cluster analysis simply discovers structures in data without explaining why they exist.

s3099587@student.rmit.edu.au

RMIT

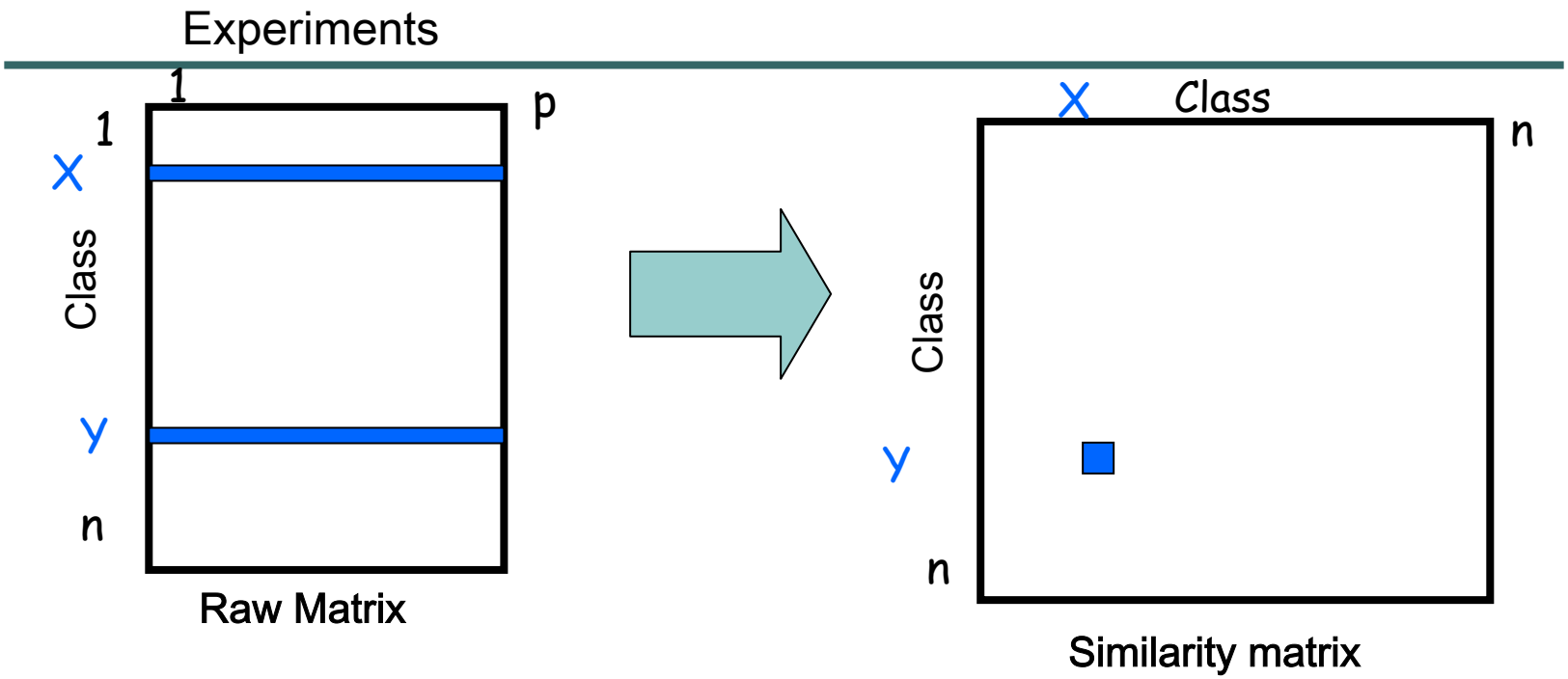
INTRO……

There are three stages to cluster analysis; namely,

- partitioning/similarity (what defines the groups),
- interpretation of clusters (how to use groups), and
- profiling the characteristics of similar/partitioned groups (what explains the groups).

s3099587@student.rmit.edu.au

How to define similarity?



- Similarity metric:
  - A measure of *pair wise* similarity or dissimilarity
  - Examples:
    - Correlation coefficient
    - Euclidean distance

s3099587@student.rmit.edu.au

RMIT

Clustering techniques

There are many clustering techniques. Here are several of them:

Hierarchical clustering: In the hierarchical procedures, we construct a hierarchy or tree-like structure to see the relationship among entities (observations or individuals)

Optimal partitioning: Most popular of them is k-means partitioning. It tries to find g groups in the data by optimising some kind of criteria.

Distribution mixtures. These techniques assume that data came from some distribution that is a mixture of several distributions.

Non-parametric methods use local density estimation. These techniques use direct search at each point of the space and try to avoid distribution assumptions.

s3099587@student.rmit.edu.au

RMIT

Hierarchical Clustering

This technique can also be subdivided into two groups.

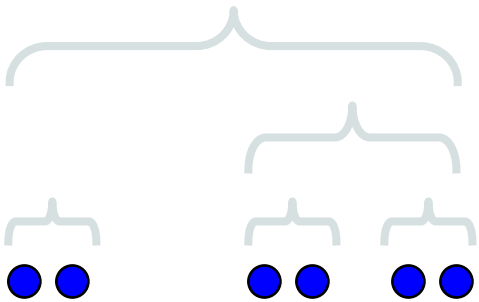
a) Divisive that starts from one group and then divides iteratively to two groups, then to three groups etc.

b) Agglomerative clustering starts from n groups and joins two groups and finds n-1 group. It continues until there is only one group. Both these clustering result in a tree like structure called *dendrograms*.

s3099587@student.rmit.edu.au

RMIT

Hierarchical Clustering [Hartigan 1975]



dendrogram

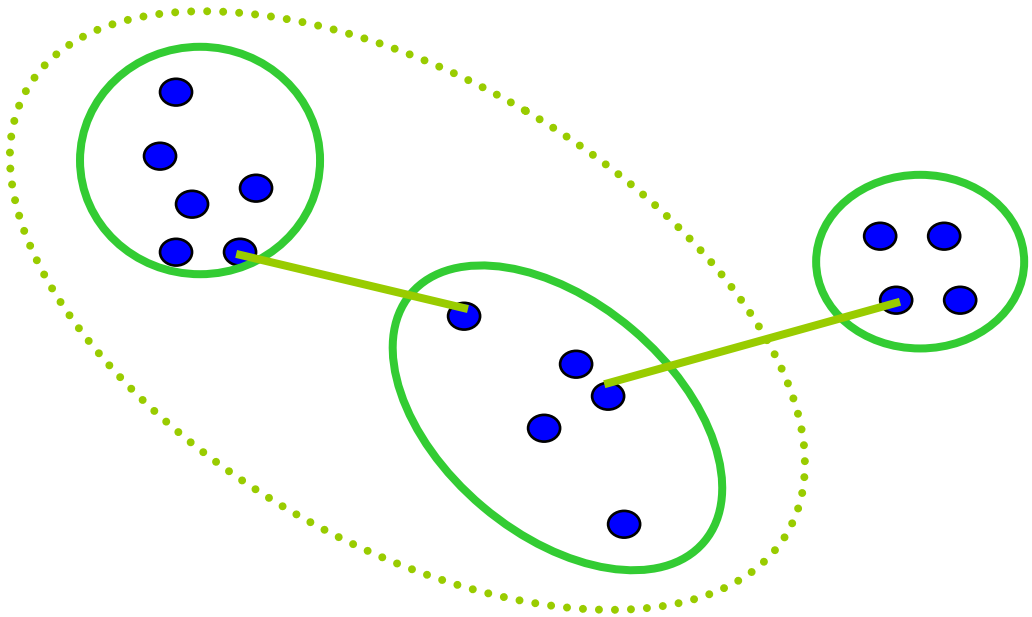
- Agglomerative (bottom-up)
- Algorithm:
  - Initialize: each item a cluster
  - Iterate:
    - select two most *similar* clusters
    - merge them
  - Halt: when required number of clusters is reached

s3099587@student.rmit.edu.au

RMIT

Hierarchical: Single Link

- cluster similarity = similarity of two most similar members



- Potentially long and skinny clusters

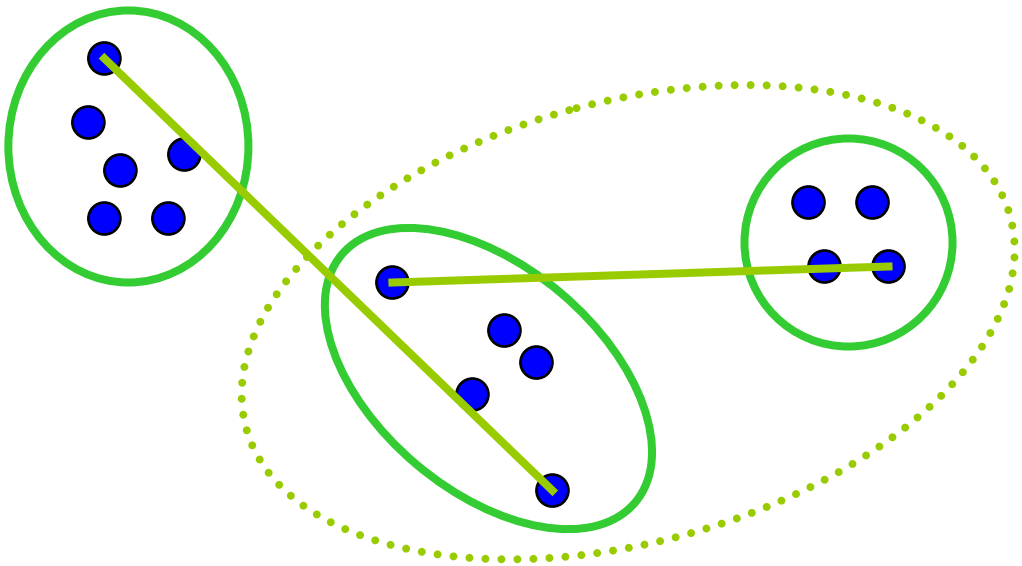
+ Fast

s3099587@student.rmit.edu.au

RMIT

Hierarchical: Complete Link

- cluster similarity = similarity of two least similar members



+ tight clusters

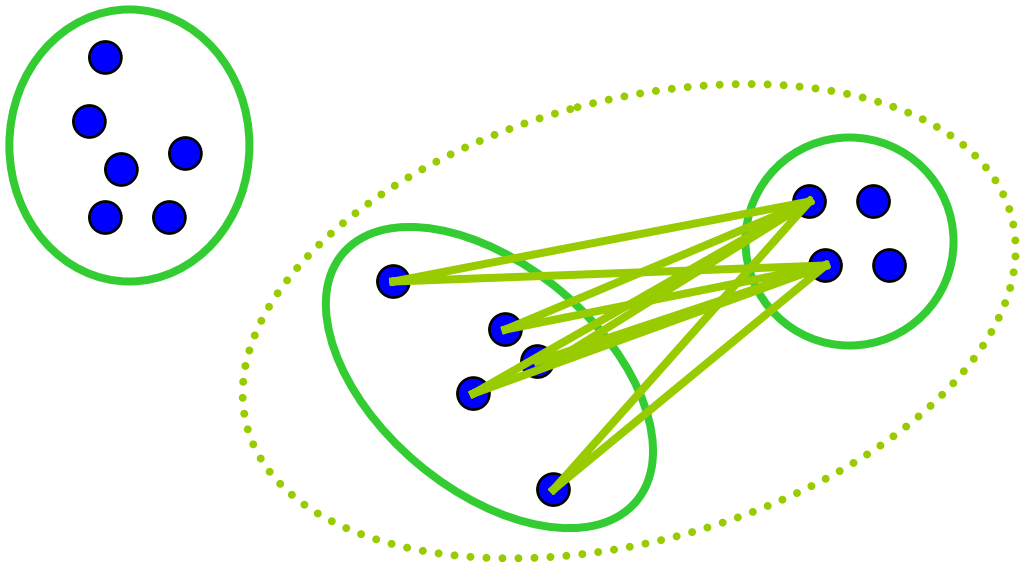
- slow

s3099587@student.rmit.edu.au

RMIT

Hierarchical: Average Link

- cluster similarity = average similarity of all pairs



+ tight clusters

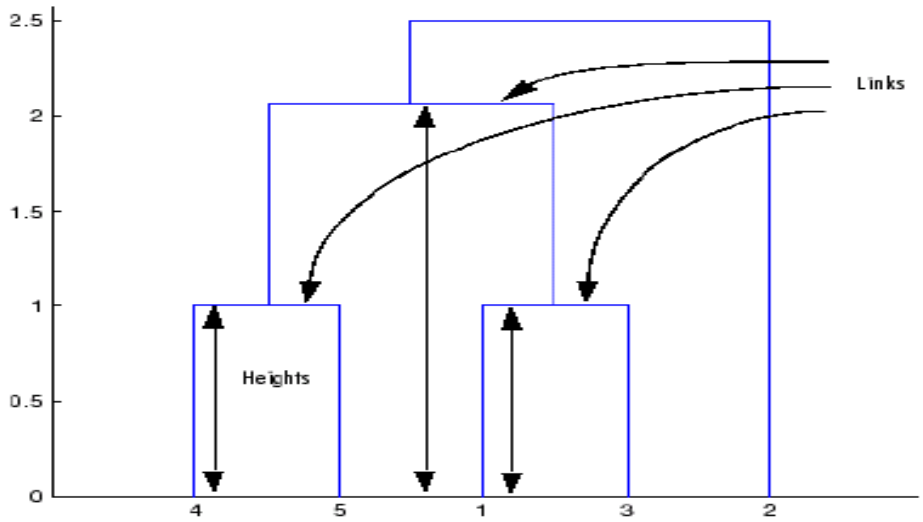
- slow

s3099587@student.rmit.edu.au

Dendrograms

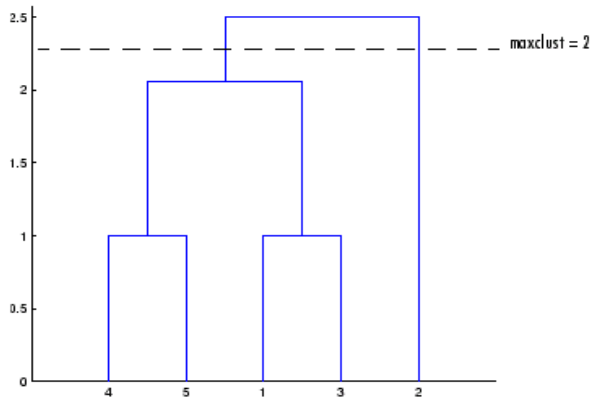
The result of hierarchical clustering is usually represented by a tree like structure. Top of tree represents one single group (root) of all data points and at the bottom individual items are given. Ordinate scale shows the dissimilarity measure at which groups are joined.

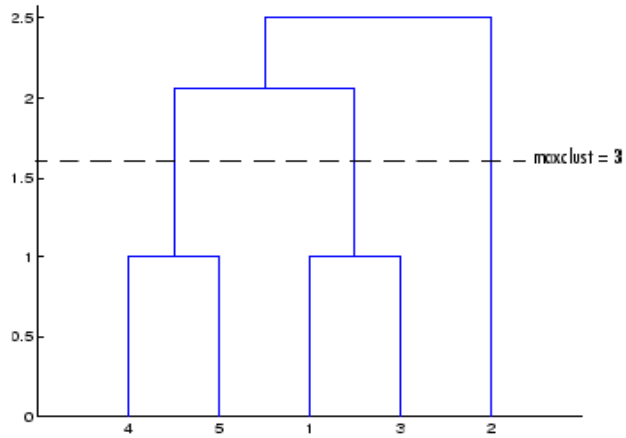
Dendrograms are built usually from bottom to top but it is natural to read them from top to bottom.

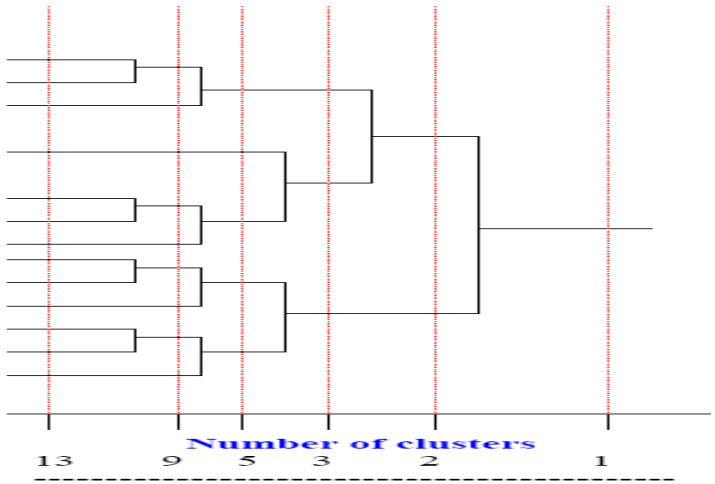


s3099587@student.rmit.edu.au

Dendrogram – Clustering Data



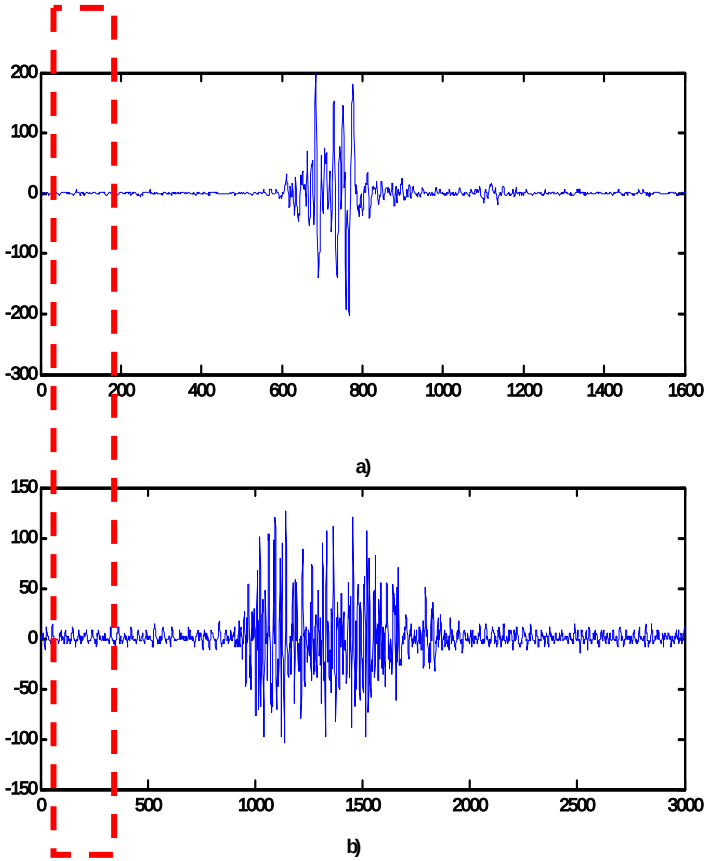


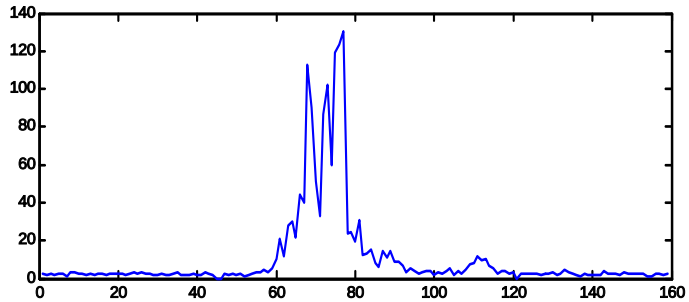


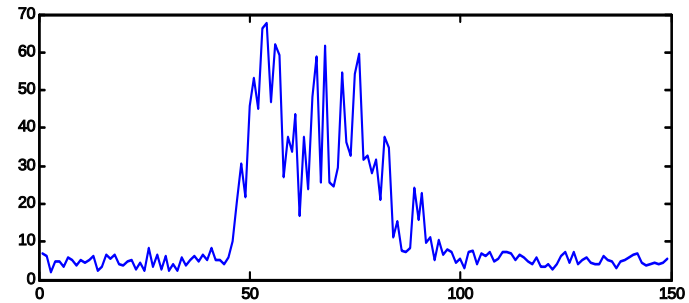
s3099587@student.rmit.edu.au

RMIT

An Example







s3099587@student.rmit.edu.au

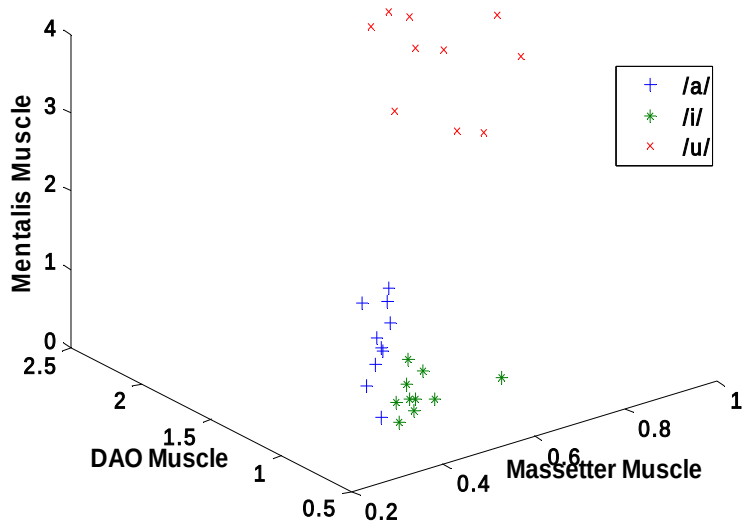
RMIT

Normalised Features

Vowel /a/		
A1	A2	A3

Vowel /i/		
A1	A2	A3

Vowel /u/		
A1	A2	A3



s3099587@student.rmit.edu.au

RMIT

Dendrogram – Using Matlab

s3099587@student.rmit.edu.au

RMIT

Continued……

s3099587@student.rmit.edu.au