CLUTO
A Clustering Toolkit

Mahmoud O. EL-Haj

Essex University

March 3, 2011

Clustering Introduction

Clustering algorithms divide data into meaningful or useful groups, called clusters.
These discovered clusters can be used to explain the characteristics of the underlying data distribution.

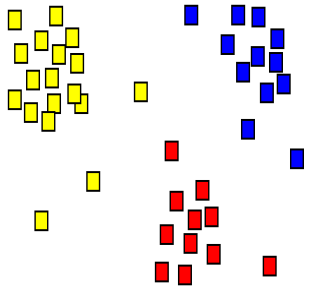

Figure: Clusters Example

Clustering Applications

The applications of clustering include:

- Characterisation of different customer groups based upon purchasing patterns
- Categorisation of documents on the World Wide Web.
- Grouping of genes and proteins that have similar functionality.
- Grouping of spatial locations prone to earth quakes from seismological data, etc

Types of Clustering

- Hierarchical algorithms: find successive clusters using previously established clusters.
  - Agglomerative ("bottom-up"), begin with each element as a separate cluster and merge them into larger clusters.
  - Divisive ("top-down"), begin with the whole set and proceed to divide it into smaller clusters.
- Partitional algorithms typically determine all clusters at once.
- Density-based clustering algorithms are devised to discover arbitrary-shaped clusters (regarded as a region).
- Subspace clustering methods look for clusters that can only be seen in a particular projection

Distance Measure

- An important step in most clustering is to select a distance measure
- determines how the similarity of two elements is calculated.
- This will influence the shape of the clusters
- some elements may be close to one another according to one distance and farther away according to another.

Common Distance Functions

- Euclidean distance.
- Manhattan distance (aka taxicab norm or 1-norm).
- Maximum norm (aka infinity norm).
- Mahalanobis distance.
- The angle between two vectors.
- Hamming distance measures

CLUTO

CLUTO Clustering Toolkit:

- Department of Computer Science, University of Minnesota, Minneapolishttp://www-users.cs.umn.edu/ karypis/
- platform
  - Linux 2.4.18
  - Sun OS 5.7
  - Win32
- programs
  - CLUTO's user callable library
  - Vcluster
  - Scluster

What is CLUTO?

CLUTO is a software package for clustering low and high dimensional datasets and for analysing the characteristics of the various clusters.

CLUTO Provides Three Types of Clustering Algorithms:

- Partitional clustering
- Agglomerative clustering
- Graph-partitioning clustering

CLUTO clustering criterion function

- Provide seven different criterion functions
- Both the partitional and agglomerative clustering algorithms provide some of the more traditional local criteria.

(e.g., single-link, complete-link, and UPGMA)

Analyze discovered clusters

- relations between the objects assigned to each cluster
- relations between the different clusters
- identify the features that best describe and/or discriminate each cluster.
- relationships between the clusters, objects, and features.

Operate on very large datasets

- the number of objects
- the number of dimensions.

CLUTO Programs

- Programs
  - vcluster : operate in the objects feature space
  - scluster : operate in the objects similarity space.
  - both cluster a collection of objects into a predetermined number of clusters k.
  - Vcluster program treats each object as a vector in a high-dimensional space
  - Scluster program operates on the similarity space between the objects
  - they both use a set of five different approaches. Four partitional and one agglomerative.

K-Means Clustering

k-means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean.

- Applying CLUTO K-Means to our Multidocument Summariser
  - Clustering Sentences from a set of related documents.
  - Select from each cluster the sentence with highest similarity.
  - Combine sentences to generate the summary