Clustering Overview

1

Informal goal

- Given set of objects and measure of similarity between them, group similar objects together
- What mean by “similar”?
- What is good grouping?
- Computation time / quality tradeoff

2

Applications:

Many

- biology
- astronomy
- computer aided design of circuits
- information organization
- marketing
- …

3

Issues

- What attributes represent items for clustering purposes?
- What is measure of similarity between items?
  - General objects and matrix of pairwise similarities
  - Objects with specific properties that allow other specifications of measure
    - Most common: Objects are d-dimensional vectors
      - » Euclidean distance
      - » cosine similarity
- What is measure of similarity between clusters?

4

Issues continued

- Cluster goals?
  - Number of clusters?
  - flat or hierarchical clustering?
  - cohesiveness of clusters?
- How evaluate cluster results?
  - relates to measure of closeness between clusters
- Efficiency of clustering algorithms
  - large data sets => external storage
- Maintain clusters in dynamic setting?
- Clustering methods? - MANY!

5

General types of clustering

- "Soft" versus "hard" clustering
  - Hard: partition the objects
    - each object in exactly one partition
  - Soft: assign degree to which object in cluster
    - view as probability or score
- flat versus hierarchical clustering
  - hierarchical = clusters within clusters

6

General types of clustering methods

- agglomerative versus divisive algorithms
  - agglomerative = bottom-up
    - build up clusters from single objects
  - divisive = top-down
    - break up cluster containing all objects into smaller clusters
- both agglomerative and divisive give hierarchies
  - hierarchy can be trivial:

    1 (. .) . . . 2 ((. .) .) . . .

    3 (((. .) .) .) . 4 ((((. .) .) .) .)

7

General types of clustering methods cont.

- constructive versus iterative improvement
  - constructive: decide in what cluster each object belongs and don't change
    - often faster
  - iterative improvement: start with a clustering and move objects around to see if can improve clustering
    - often slower but better

8

Quality of clustering

- In applications quality of clustering depends on how well solves problem at hand
- Algorithm uses measure of quality that can be optimized, but that may or may not do a good job of capturing application needs.
- Underlying graph-theoretic problems usually NP-complete
  - e.g. graph partitioning
- Usually algorithm not finding optimal clustering

9

Distance between clusters

Possible definitions:

I. distance between closest pair of objects with one in each cluster
   - called single link

II. distance between furthest pair objects, one from each cluster
   - called complete linkage

10

Distance between clusters, cont.

Possible definitions:

III. average of pairwise distance between all pairs of objects, one from each
  - more computation

- Generally no representative point for a cluster;
- If Euclidean distance
  - centroid
  - bounding box

11

Vector model: K-means algorithm

- Uses Euclidean distance
- Uses centroid

- Well known, well used
- Flat clustering
- Number of clusters picked ahead of time

12

K-means overview

- Choose k points among set to cluster
  - Call them k centroids
- For each point not selected, assign it to its closest centroid
  - All assignment give initial clustering
- Until “happy” do:
  - Recompute centroids of clusters
    - New centroids may not be points of original set
  - Reassign all points to closest centroid
    - Updates clusters

13

K-means performance

- Sum of squared distances of each point to its centroid is measure trying to optimize (RSS)
- Can prove RSS decreases with each iteration, so converge
- Can achieve local optimum
  - No change in centroids
- Running time depends on how demanding stopping criteria
- Works well in practice
  - speed
  - quality

14