Subscribe, if you haven’t gotten any msgs

Overview

- What is clustering?
- Similarity/distance metrics
- Hierarchical clustering algorithms
  - E.g. [Eisen et al. 1998]
- K-means
  - E.g. [MacQueen, 1965] [Tavazoie et al. 1999]
- Self-organizing map (SOM)
  - E.g. [Tamayo et al. 1999]

What is clustering?

- Group similar objects together
- Objects in the same cluster (group) are more similar to each other than objects in different clusters
- Data exploratory tool
Clustering Expression Data

• Why cluster gene expression data?
  – Tissue classification
  – Find biologically related genes
  – First step in inferring regulatory networks
  – Look for common promoter elements
  – Hypothesis generation
  – One of the tools of choice for expression analysis

Clustering Expression Data

• What has been done?
  – Partitional
    • CAST (Ben-Dor et al. 1999)
    • k-means, variously initialized (Hartigan 1975)
  – Hierarchical
    • single-, average-, complete-, centroid-link [Eisen et al. 98]
    – Self Organizing Maps (SOM) [Tamayo et al. 99]
    – Support Vector Machines (SVM) [Grundy et al. 00]
  – etc., etc., etc.

Clustering Expression Data

• Why so many methods?
  – Clustering is NP-hard, even with simple objectives, data
  – Hard problem: high dimensionality, noise, …
  – . . . many heuristic, local search, & approximation algorithms
  – No clear winner

Clustering Expression Data

• How to define similarity?
  – Similarity metric:
    – A measure of pairwise similarity or dissimilarity
    – Examples:
      • Correlation coefficient
      • Euclidean distance
Similarity metrics

- Euclidean distance
  \[
  \sqrt{\sum_{j=1}^{p} (X_j - \overline{Y}_j)^2}
  \]
- Correlation coefficient
  \[
  \frac{\sum_{j=1}^{p} (X_j - \overline{X}_j)(Y_j - \overline{Y}_j)}{\sqrt{\sum_{j=1}^{p} (X_j - \overline{X}_j)^2 \sum_{j=1}^{p} (Y_j - \overline{Y}_j)^2}} \quad \text{where} \quad \overline{X} = \frac{\sum_{j=1}^{p} X_j}{p}
  \]

Lessons from the example

- Correlation – direction only
- Euclidean distance – magnitude & direction
- Min # attributes (experiments) to compute pairwise similarity
  - >= 2 attributes for Euclidean distance
  - >= 3 attributes for correlation
- Array data is noisy \( \Rightarrow \) need many experiments to robustly estimate pairwise similarity

Example

<table>
<thead>
<tr>
<th>X</th>
<th>1</th>
<th>0</th>
<th>-1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Z</td>
<td>-1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>W</td>
<td>2</td>
<td>0</td>
<td>-2</td>
<td>0</td>
</tr>
</tbody>
</table>

- Correlation (X,Y) = 1          Distance (X,Y) = 4
- Correlation (X,Z) = -1         Distance (X,Z) = 2.83
- Correlation (X,W) = 1         Distance (X,W) = 1.41

Clustering algorithms

- Inputs:
  - Raw data matrix or similarity matrix
  - Number of clusters or some other parameters
- Many different classifications of clustering algorithms:
  - Hierarchical vs partitional
  - Heuristic-based vs model-based
  - Soft vs hard
Hierarchical Clustering
[Hartigan 1975]

- 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

Hierarchical: Single Link
- cluster similarity = similarity of two most similar members
  - Potentially long and skinny clusters
  - Fast

Example: single link

$$d_{(1,2,3,4,5)} = \min\{d_{1,3}, d_{2,4}, d_{3,5}\} = \min\{6,3\} = 3$$
$$d_{(1,2,3,4,5)} = \min\{d_{1,4}, d_{2,5}\} = \min\{10,9\} = 9$$
$$d_{(1,2,3,4,5)} = \min\{d_{1,5}, d_{2,3}\} = \min\{9,8\} = 8$$
**Example: single link**

\[
d_{(1,2,3)(4,5)} = \min\{d_{(1,2,3)4}, d_{(1,2,3)5}\} = 5
\]

Sometimes drawn to a scale

**Hierarchical: Complete Link**

- cluster similarity = similarity of two least similar members
- tight clusters
- slow

**Example: complete link**

\[
d_{(1,2,3)} = \max\{d_{13}, d_{23}\}_{\text{max} = 6,3} = 6
\]
\[
d_{(1,2,4)} = \max\{d_{14}, d_{24}\}_{\text{max} = 10,9} = 10
\]
\[
d_{(1,2,5)} = \max\{d_{15}, d_{25}\}_{\text{max} = 9,8} = 9
\]
Example: complete link

Hierarchical: Average Link
- cluster similarity = average similarity of all pairs

Example: average link

Example: average link
Example: average link

Hierarchical: Centroid Link
- cluster centroid = average of all points
- cluster similarity = distance between centroids

In Expression literature, often called “Average link”
+ faster
- discards shape

Algorithm Analysis
(see class notes)

Software: TreeView [Eisen et al. 1998]
- Fig 1 in Eisen’s PNAS 99 paper
- Time course of serum stimulation of primary human fibroblasts
- cDNA arrays with approx 8600 spots
- centroid-link
- Free download at: http://rana.lbl.gov/EisenSoftware.htm
- Another Good Package: TMEV
  - http://www.tigr.org/software/tm4/
Hierarchical divisive clustering algorithms

- Top down
  - Start with all the objects in one cluster
  - Successively split into smaller clusters
- Tend to be less efficient than agglomerative
- Resolver implemented a deterministic annealing approach from [Alon et al. 1999]

Partitional: K-Means

[MacQueen 1965]

1

2

3

Details of k-means

- Iterate until converge:
  - Assign each data point to the closest centroid
  - Compute new centroid

Objective function:

\[
\text{Minimize } \sum_x (x - \text{Centroid(Cluster}(x)))^2
\]

Properties of k-means

- Fast
- Proved to converge to local optimum
- In practice, converge quickly
- Tend to produce spherical, equal-sized clusters
- Related to the model-based approach (next lecture)
Summary

• Definition of clustering
• Pairwise similarity:
  – Correlation
  – Euclidean distance
• Clustering algorithms:
  – Hierarchical (single-, complete-, average-, centroid-link)
  – K-means
  – SOM
• Different clustering algorithms ➔ different clusters

Misc Notes

• Greedy algorithms. Can get trapped in local minima. Can be sensitive to addition of new points, order of points,…

+ simple, intuitive algorithms, reasonably fast, ok on simple data, no obvious preconception about structure
- no model of structure; biases unclear