Spectral clustering for microarray data

Alvin Liang
Grant Wang
Cameron Wheeler
12 May 2005
Outline

• Introduction to spectral algorithm
• Application to Golub et al. leukemia data set
• Application to Devauchelle et al. arthritis data set
Introduction to the spectral algorithm

• Two steps:
  – Create hierarchical tree in “top-down” manner based on eigenvalues, eigenvectors
  – Merges leaves in “bottom-up” fashion to create clusters based on objective function
    • Informally, objective function maximizes:
      (“similarity” inside clusters) - (“similarity” outside clusters)
    • You choose what “similarity” means

• Algorithm has been applied in other contexts (web search)
Golub et al. data

• 38 “training” patients, 34 “test” patients
• Cluster into acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML)
• They consider “class discovery”: What if you didn’t know about the ALL/AML distinction? How would you find it?
• Their answer: clustering
Their results

• Only 4 errors out of 34 using self-organizing maps, in two clusters (they used GENECLUSTER)
• How does spectral clustering do?
• How do we present the data to the spectral clustering algorithm?
Normalization

• Spectral clustering wants similarity between patients (between 0,1)
• How do we do it?
  - Normalize data so mean = 0, variance = 1
    - (same as Golub et al)
  - Put all positive genes in one coordinate
  - Put all negative genes in another
• Now dot product between two patients is similarity
Our results

• Creates one cluster with (14 ALL, 3 AML)
• Creates another cluster with (8 ALL, 9 AML)
• Why doesn’t it work?
• Let’s look at three different clusterings:
  – “Correct” clustering (all ALL in one, all AML in other
  – “Our clustering”
  – Random clustering
What is a good clustering?

• If a clustering of patients is good, then there should be genes that are expressed high in that cluster, and low in the other.
  – (I.e. genes that “differentiate” the two clusters)

• We find:
  – “Correct clustering”: yes, there are such genes
  – Our clustering: yes, there are such genes
  – Random clustering: no, no such genes
Results

• True clustering:

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>0.92</td>
<td>-0.92</td>
<td>GLUTATHIONE S-TRANSFERASE, MICROSONAL</td>
</tr>
<tr>
<td>0.05</td>
<td>0.92</td>
<td>-0.87</td>
<td>APLP2 Amyloid beta (A4) precursor-like protein 2</td>
</tr>
<tr>
<td>0.00</td>
<td>0.85</td>
<td>-0.85</td>
<td>CD33 CD33 antigen (differentiation antigen)</td>
</tr>
</tbody>
</table>

• Our clustering:

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.11</td>
<td>1.00</td>
<td>-0.88</td>
<td>GB DEF = Secreted epithelial tumor mucin antigen</td>
</tr>
<tr>
<td>0.17</td>
<td>1.00</td>
<td>-0.82</td>
<td>KIAA0265 gene, partial cds</td>
</tr>
<tr>
<td>0.11</td>
<td>0.94</td>
<td>-0.82</td>
<td>Hyaluronoglucosaminidase 1 (HYAL1) mRNA</td>
</tr>
</tbody>
</table>

• Random clustering:

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2353</td>
<td>0.8235</td>
<td>-0.5882</td>
<td>n/a</td>
</tr>
<tr>
<td>0.1765</td>
<td>0.7059</td>
<td>-0.5294</td>
<td>n/a</td>
</tr>
<tr>
<td>0.1765</td>
<td>0.7059</td>
<td>-0.5294</td>
<td>n/a</td>
</tr>
</tbody>
</table>
OA vs RA

Four images removed for copyright reasons.
Seek to Classify OA and RA

• Use Traditional Clustering Methods
  – Classification
  – Gene Clustering

• Spectral Clustering
  – Classification
  – Gene Clustering
Data from Devauchelle et al.

- Synovial tissue
  - 13 OA patients (age 72 ± 9.3)
  - 8 RA patients (age 55± 9.2)

- 4652 Genes probed
  - 63 genes selected
  - Normalized to mRNA levels
PCA and K-means

### OA and RA principle components

<table>
<thead>
<tr>
<th>Principal Components</th>
<th>Variance Explained (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0%</td>
</tr>
<tr>
<td>2</td>
<td>10%</td>
</tr>
<tr>
<td>3</td>
<td>20%</td>
</tr>
<tr>
<td>4</td>
<td>30%</td>
</tr>
<tr>
<td>5</td>
<td>40%</td>
</tr>
<tr>
<td>6</td>
<td>50%</td>
</tr>
<tr>
<td>7</td>
<td>60%</td>
</tr>
<tr>
<td>8</td>
<td>70%</td>
</tr>
<tr>
<td>9</td>
<td>80%</td>
</tr>
<tr>
<td>10</td>
<td>100%</td>
</tr>
</tbody>
</table>

The graph shows the variance explained by each principal component, with 100% explained by the first component.
Clustering of OA and RA (2 clusters)
Clustering of OA and RA (3 clusters)
Clustering of OA and RA (4 clusters)
Clustering of OA and RA (5 clusters)
Spectral Clustering
Clusters

• Spectral
  – Cluster 1
    • 13 OA patients
  – Cluster 2
    • 8 RA patients

• K-means
  – 2 Clusters (13 OA patients, 8 RA patients)
  – 3 Clusters (9 OA, 8 RA, 4 OA)
  – 4 Clusters (7 OA, 8 RA, 4 OA, 2 OA)
Clustering Genes (OA)

OA principle components

Variance Explained (%)

Principal Components
Clustering of OA (2 clusters)
Clustering of OA (3 clusters)
Spectral Clustering
Clusters

• Spectral Clustering
  – Genes in Cluster 1
    • RNB6, FLJ10342, CDK7, BRD3
  – Genes in Cluster 2
    • CTSD, TIMP2, FUBP1

• K-means (2 Clusters)
  – Genes in Cluster 1
    • TIMP2, CDK7
  – Genes in Cluster 2
    • CTSD, RNB6, FLJ10342