Clustering Methods:  
From k-means to Gaussian Mixture Model and Louvain Algorithm

Maxwell Lee

High-dimension Data Analysis Group  
Laboratory of Cancer Biology and Genetics  
Center for Cancer Research  
National Cancer Institute

November 16, 2020
Outline of Clustering Methods

- **Traditional method**
  - Hierarchical clustering
  - K-means clustering
- **Distribution based method**
  - GMM
  - LDA
- **Graph based method**
  - Spectral clustering
  - Louvain clustering

GMM: Gaussian Mixture Model
LDA: Latent Dirichlet Allocation
NMF: Non-negative matrix factorization

Contributed by Emily Tai
Hierarchical Agglomerative Clustering Algorithm

- Start with n leaf nodes
- Sequentially merge a pair of nodes with the smallest distance or minimal variance
- End with a single cluster
Spectral Clustering

Spectrum: set of its eigenvalues

Laplacian Eigenmaps

K-means Clustering

Euclidean distance is not appropriate in HD or in non-Euclidean space

- Curse of dimensionality
- Laplacian Eigenmaps is a non-linear dimension reduction method
K-means Clustering

color by group

color by k-means cluster
Two Circles and K-means Clustering

color by group

color by k-means cluster
Two Circles and Spectral Clustering

color by group

color by Spectral Clustering
Three Circles and K-means Clustering

color by group

color by k-means cluster
Three Circles and Spectral Clustering

color by group

color by Spectral Clustering

data set

color by Spectral Clustering

scc
PCA Label by Subtype vs. Spectral Clustering

Label by subtype

Spectral Clustering in HD

977 samples  accuracy 64.5%

HD: high dimension, 5000 genes
Spectral Clustering vs. k-means Cluster (HD)

<table>
<thead>
<tr>
<th></th>
<th>Basal</th>
<th>Her2</th>
<th>LumA</th>
<th>LumB</th>
<th>Normal</th>
</tr>
</thead>
<tbody>
<tr>
<td>clust1</td>
<td>165</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>clust2</td>
<td>8</td>
<td>64</td>
<td>8</td>
<td>13</td>
<td>7</td>
</tr>
<tr>
<td>clust3</td>
<td>0</td>
<td>5</td>
<td>245</td>
<td>47</td>
<td>2</td>
</tr>
<tr>
<td>clust4</td>
<td>0</td>
<td>4</td>
<td>78</td>
<td>132</td>
<td>0</td>
</tr>
<tr>
<td>clust5</td>
<td>0</td>
<td>0</td>
<td>169</td>
<td>1</td>
<td>24</td>
</tr>
</tbody>
</table>

Accuracy = (165 + 64 + 245 + 132 + 24) / 977 = 64.5%

<table>
<thead>
<tr>
<th></th>
<th>Basal</th>
<th>Her2</th>
<th>LumA</th>
<th>LumB</th>
<th>Normal</th>
</tr>
</thead>
<tbody>
<tr>
<td>clust1</td>
<td>169</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>clust2</td>
<td>4</td>
<td>69</td>
<td>17</td>
<td>40</td>
<td>5</td>
</tr>
<tr>
<td>clust3</td>
<td>0</td>
<td>0</td>
<td>268</td>
<td>11</td>
<td>21</td>
</tr>
<tr>
<td>clust4</td>
<td>0</td>
<td>0</td>
<td>125</td>
<td>119</td>
<td>0</td>
</tr>
<tr>
<td>clust5</td>
<td>0</td>
<td>4</td>
<td>90</td>
<td>23</td>
<td>6</td>
</tr>
</tbody>
</table>

Accuracy = (169 + 69 + 268 + 119 + 6) / 977 = 64.6%

5000 genes
Spectral Clustering
Match
Mismatch
K-means
PCA with **pam50**: Label by Subtype vs. Spectral Clustering

**Label by subtype**

**Label by Spectral Clustering**

- **accuracy 70%**

HD: high dimension, 39 genes
## Spectral Clustering vs. k-means Clustering (pam50)

### Spectral Clustering

<table>
<thead>
<tr>
<th></th>
<th>Basal</th>
<th>Her2</th>
<th>LumA</th>
<th>LumB</th>
<th>Normal</th>
</tr>
</thead>
<tbody>
<tr>
<td>clust1</td>
<td>167</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>clust2</td>
<td>6</td>
<td>64</td>
<td>6</td>
<td>12</td>
<td>5</td>
</tr>
<tr>
<td>clust3</td>
<td>0</td>
<td>6</td>
<td><strong>361</strong></td>
<td><strong>114</strong></td>
<td>2</td>
</tr>
<tr>
<td>clust4</td>
<td>0</td>
<td>3</td>
<td>67</td>
<td>67</td>
<td>0</td>
</tr>
<tr>
<td>clust5</td>
<td>0</td>
<td>0</td>
<td>66</td>
<td>0</td>
<td><strong>25</strong></td>
</tr>
</tbody>
</table>

**Accuracy** = \((167 + 64 + 361 + 67 + 25) / 977 = 70\%\)

### K-means

<table>
<thead>
<tr>
<th></th>
<th>Basal</th>
<th>Her2</th>
<th>LumA</th>
<th>LumB</th>
<th>Normal</th>
</tr>
</thead>
<tbody>
<tr>
<td>clust1</td>
<td>170</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>clust2</td>
<td>2</td>
<td>72</td>
<td>11</td>
<td>31</td>
<td>4</td>
</tr>
<tr>
<td>clust3</td>
<td>0</td>
<td>0</td>
<td><strong>221</strong></td>
<td>75</td>
<td>0</td>
</tr>
<tr>
<td>clust4</td>
<td>1</td>
<td>1</td>
<td>81</td>
<td><strong>87</strong></td>
<td>0</td>
</tr>
<tr>
<td>clust5</td>
<td>0</td>
<td>0</td>
<td><strong>187</strong></td>
<td>0</td>
<td><strong>26</strong></td>
</tr>
</tbody>
</table>

**Accuracy** = \((170 + 72 + 221 + 87 + 26) / 977 = 59\%\)
Spectral Clustering

Spectrum: set of its eigenvalues

Laplacian Eigenmaps

K-means Clustering

Euclidean distance is not appropriate in HD or in non-Euclidean space

- Curse of dimensionality
- Laplacian Eigenmaps is a non-linear dimension reduction method
In high dimension, for any given point, all other points are on the surface of n-ball (hypersphere).
The Problems of Curse of Dimensionality

High density of local data points is required to estimate parameters in k-means and GMM
Euclidean Distance

\[ d(p, q)^2 = (q_1 - p_1)^2 + (q_2 - p_2)^2 \]

\[ d(p, q) = d(q, p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} \]
Triangle on a Curved Surface vs on a Plane

Non-Euclidean
\[ A^2 + B^2 > C^2 \]
\[ \alpha + \beta + \gamma > 180^0 \]

Euclidean
\[ A^2 + B^2 = C^2 \]
\[ \alpha + \beta + \gamma = 180^0 \]
Matrix Representation of Graph

undirected graph

adjacency matrix

\[
\begin{array}{ccccccc}
 & A & B & C & D & E \\
A & 0 & 1 & 1 & 0 & 0 \\
B & 1 & 0 & 1 & 0 & 0 \\
C & 1 & 1 & 0 & 1 & 0 \\
D & 0 & 0 & 1 & 0 & 1 \\
E & 0 & 0 & 0 & 1 & 0 \\
\end{array}
\]
Finding One-Stop Flight by Matrix Multiplication

directed graph

adjacency matrix (M)

A | B | C | D | H
---|---|---|---|---
A | 0 | 0 | 1 | 0 | 1
B | 1 | 0 | 0 | 0 | 1
C | 0 | 0 | 0 | 1 | 1
D | 0 | 1 | 0 | 0 | 1
H | 1 | 1 | 1 | 1 | 0

one-stop flight (M%*%M)

A | B | C | D | H
---|---|---|---|---
A | 1 | 1 | 2 | 1
B | 1 | 1 | 2 | 1 | 1
C | 1 | 2 | 1 | 1
D | 2 | 1 | 1 | 1 | 1
H | 1 | 1 | 1 | 1 | 4
Algorithm of Spectral Clustering

- **Data.Matrix**

KNN: k nearest neighbor

- **KNN.Graph**

Weight matrix: $w_{ij} = \exp\left(\frac{(x_i - x_j)^2}{\sigma}\right)$

- **Similarity.Matrix**

L: Laplacian Matrix (D – W)

- **Laplacian.Matrix**

- **Graph.Partition**

- **Eigen.Decomposition**

LD: low dimension
K-Nearest Neighbor Graph


$k$NN graph ($k = 3$)

$\epsilon$-ball graph
Weight matrix: \( w_{ij} = \exp\left(-\frac{(x_i - x_j)^2}{2\sigma^2}\right) \)

Similarity Matrix

Data.Matrix \( \xrightarrow{\quad} \) KNN.Graph \( \xrightarrow{\quad} \) Similarity.Matrix \( \xrightarrow{\quad} \) Laplacian.Matrix \( \xrightarrow{\quad} \) Graph.Partition \( \xrightarrow{\quad} \) Eigen.Decomposition \( \xrightarrow{\quad} \) Clustering.LD

\[
W = \begin{pmatrix}
0 & .8 & .8 & 0 & 0 \\
.8 & 0 & .8 & 0 & 0 \\
.8 & .8 & 0 & .1 & 0 \\
0 & 0 & .1 & 0 & .9 \\
0 & 0 & 0 & .9 & 0 \\
\end{pmatrix}
\]

\( \text{diag}(W) = (0,0,0,0,0) \)
Laplacian Matrix

\[ D \text{ is 5 by 5 square matrix} \]
\[ \text{diag}(D) = (1.6,1.6,1.7,1,0.9) \]

\[ L: \text{Laplacian Matrix} = (D - W) \]

\[ W = \begin{pmatrix} 0 & .8 & .8 & 0 & 0 \\ .8 & 0 & .8 & 0 & 0 \\ .8 & .8 & 0 & .1 & 0 \\ 0 & 0 & .1 & 0 & .9 \\ 0 & 0 & 0 & .9 & 0 \end{pmatrix} \]
Graph Partitioning with Ratio Cut

\[ \text{Cut}(A, B) = \sum_{i \in A, j \in B} w_{ij} \]

\[ \text{RatioCut}(A,B) = \text{cut}(A,B)(1/a + 1/b) \]

- a: number of nodes in A
- b: number of nodes in B

Jianbo Shi and Jitendra Malik, 2000
Graph Partitioning with Ratio Cut

\[ f_i = \begin{cases} 
  (b/a)^{1/2}, & i \in A \\
  -(a/b)^{1/2}, & i \in B 
\end{cases} \]

\[ \sum_{i,j} w_{ij} (f_i - f_j)^2 \]

\[ = 2(b/a + a/b + 2) \text{cut}(A,B) \]
\[ = 2n(1/a + 1/b) \text{cut}(A,B) \]
\[ = 2n \text{RatioCut}(A,B) \]

n: total number of nodes (a+b)
Eigen Decomposition of Laplacian Matrix

\[ f = (f_1, f_2, \ldots, f_n) \]

\[ f^T L f = f^T (D - W) f = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij} (f_i - f_j)^2 \]

\( L \) is positive semidefinite

\[ 0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n \]

Cost function

\[ \text{Cost function} = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij} (f_i - f_j)^2 \]

subject to \( f^T f = 1 \)

Eigen decomposition of Laplacian matrix \( L \)
K-means Clustering in Low Dimension

- Data Matrix
- KNN Graph
- Similarity Matrix
- Laplacian Matrix
- Graph Partition
- Eigen Decomposition
- Clustering LD

\[ 0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n \]

Eigen decomposition of Laplacian matrix \( L \)

- \( k \)-means clustering/GMM in low-dimension
- with eigen vectors \( f_2, \ldots, f_k \)
Outline of Clustering Methods

- Traditional method
  - Hierarchical clustering
  - K-means clustering

- Distribution based method
  - GMM
  - LDA
  - NMF

- Graph based method
  - Spectral clustering
  - Louvain clustering

Contributed by Emily Tai

GMM: Gaussian Mixture Model
LDA: Latent Dirichlet Allocation
NMF: Non-negative matrix factorization