Clustering
Lecture 7: Clustering Ensemble

Jing Gao
SUNY Buffalo
Outline

• **Basics**
  – Motivation, definition, evaluation

• **Methods**
  – Partitional
  – Hierarchical
  – Density-based
  – Mixture model
  – Spectral methods

• **Advanced topics**
  – Clustering ensemble
  – Clustering in MapReduce
  – Semi-supervised clustering, subspace clustering, co-clustering, etc.
Clustering Ensemble

• Problem
  – Given an unlabeled data set $D=\{x_1,x_2,\ldots,x_n\}$
  – An ensemble approach computes:
    • A set of clustering solutions $\{C_1,C_2,\ldots,C_k\}$, each of which maps data to a cluster: $f_j(x)=m$
    • A unified clustering solutions $f^*$ which combines base clustering solutions by their consensus

• Challenges
  – The correspondence between the clusters in different clustering solutions is unknown
Motivations

• Goal

– Combine “weak” clusterings to a better one

[PTJ05]
## An Example

The goal: get the consensus clustering they may not represent the same cluster!

<table>
<thead>
<tr>
<th></th>
<th>$C_1$</th>
<th>$C_2$</th>
<th>$C_3$</th>
<th>$C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$v_1$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$v_2$</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>$v_3$</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$v_4$</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>$v_5$</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>$v_6$</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

[GMT07]
Methods (1)

• How to get base models?
  – Bootstrap samples
  – Different subsets of features
  – Different clustering algorithms
  – Random number of clusters
  – Random initialization for K-means
  – Incorporating random noises into cluster labels
  – Varying the order of data in on-line methods
Methods (2)

• How to combine the models?

- Correspondence
  - Explicit
  - Implicit

- Consensus Function
- Representation
  - Optimized Method
  - Object-based
  - Cluster-based
  - Object-Cluster-based

- Generative Approaches
Hard Correspondence (1)

- **Re-labeling+voting**
  
  - Find the correspondence between the labels in the partitions and fuse the clusters with the same labels by voting [DuFr03,DWH01]

<table>
<thead>
<tr>
<th></th>
<th>C₁</th>
<th>C₂</th>
<th>C₃</th>
</tr>
</thead>
<tbody>
<tr>
<td>v₁</td>
<td>1</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>v₂</td>
<td>1</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>v₃</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>v₄</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>v₅</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>v₆</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>C₁</th>
<th>C₂</th>
<th>C₃</th>
<th>C*</th>
</tr>
</thead>
<tbody>
<tr>
<td>v₁</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>v₂</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>v₃</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>v₄</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>v₅</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>v₆</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>
Hard Correspondence (2)

• Details
  – Minimize match costs
  – Match to a reference clustering or match in a pairwise manner

• Problems
  – In most cases, clusters do not have one-to-one correspondence
Soft Correspondence* (1)

• Notations
  – Membership matrix $M_1, M_2, ..., M_k$
  – Membership matrix of consensus clustering $M$
  – Correspondence matrix $S_1, S_2, ..., S_k$
  – $M_i S_i = M$

<table>
<thead>
<tr>
<th></th>
<th>$C_1$</th>
<th>$C_2$</th>
<th>$C_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$v_1$</td>
<td>1</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>$v_2$</td>
<td>1</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>$v_3$</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>$v_4$</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>$v_5$</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>$v_6$</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

$$ M_2 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad S_2 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \quad M = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} $$

* [LZY05]
Soft Correspondence (2)

• Consensus function
  – Minimize disagreement \( \min \sum_{j=1}^{k} \| M - M_j S_j \|^2 \)
  – Constraint 1: column-sparseness
  – Constraint 2: each row sums up to 1
  – Variables: \( M, S_1, S_2, \ldots, S_k \)

• Optimization
  – EM-based approach
  – Iterate until convergence
    • Update \( S \) using gradient descent
    • Update \( M \) as \( M = \frac{1}{k} \sum_{j=1}^{k} M_j S_j \)
• How to combine the models?

- Correspondence
  - Explicit
  - Implicit
    - Consensus Function
      - Optimization Method
      - Object-based
      - Cluster-based
      - Object-Cluster-based
    - Representation
      - Generative Approaches
Object-based Methods (1)

- Clustering objects
  - Define a similarity or distance measure:
    - Similarity between two objects can be defined as the percentage of clusterings that assign the two objects into same clusters
    - Distance between two objects can be defined as the percentage of clusterings that assign the two objects into different clusters
  - Conduct clustering on the new similarity (distance) matrix
  - Result clustering represents the consensus
  - Can view this approach as clustering in the new feature space where clustering results are the categorical features
Object-based Methods (2)

<table>
<thead>
<tr>
<th>$v_1$</th>
<th>$v_2$</th>
<th>$v_3$</th>
<th>$v_4$</th>
<th>$v_5$</th>
<th>$v_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Diagram:

- Nodes labeled $v_1$ to $v_6$ connected by edges.
- Nodes are grouped in clusters:
  - Cluster 1: $v_1$, $v_2$, $v_3$ connected.
  - Cluster 2: $v_4$, $v_5$, $v_6$ connected.
Co-association matrix $T$ from $\lambda^{(1)}$, $\lambda^{(2)}$, $\lambda^{(3)}$, and $\lambda^{(4)}$. 

[StGh03]
Consensus Function

- Minimizing disagreement
  - Information-theoretic [StGh03]
    \[
    \max \frac{1}{k} \sum_{j=1}^{k} NMI(T, T_j) \quad \text{NMI}(T, T_j) = \frac{I(T, T_j)}{\sqrt{H(T)H(T_j)}}
    \]
  - Median partition [LDJ07]
    \[
    \overline{T} = \frac{1}{k} \sum_{j=1}^{k} T_j \quad \min \| \overline{T} - T \|^2
    \]
  - Correlation clustering [GMT07]
    \[
    \max \sum_{C(u) = C(v)} T_{uv} + \sum_{C(u) \neq C(v)} (1 - T_{uv})
    \]
Optimization Method

• Approximation
  – Agglomerative clustering (bottom-up) [FrJa02, GMT07]
    • Single link, average link, complete link
  – Divisive clustering (top-down) [GMT07]
    • Furthest
  – LocalSearch [GMT07]
    • Place an object into a different cluster if objective function improved
    • Iterate the above until no improvements can be made
  – BestClustering [GMT07]
    • Select the clustering that maximize (minimize) the objective function
  – Graph partitioning [StGh03]
  – Nonnegative matrix factorization [LDJ07, LiDi08]
[GMT07]
Overall Distance on Votes data set

BestClustering, Agglomerative, Furthest, Balls, LocalSearch, Rock, Limbo

[GMT07]
• How to combine the models?

Correspondence

Explicit

Consensus Function

Optimization Method

Implicit

Representation

Object-based

Cluster-based

Object-Cluster-based

Generative Approaches
Cluster-based Methods

• Clustering clusters
  – Regard each cluster from a base model as a record
  – Similarity is defined as the percentage of shared common objects
    • eg. Jaccard measure
  – Conduct clustering on these clusters
  – Assign an object to its most associated consensus cluster
Meta-Clustering Algorithm (MCLA)*

<table>
<thead>
<tr>
<th></th>
<th>$C_1$</th>
<th>$C_2$</th>
<th>$C_3$</th>
<th>$C$</th>
<th>$M_1$</th>
<th>$M_2$</th>
<th>$M_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$v_1$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$v_2$</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>$v_3$</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$v_4$</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>$v_5$</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>$v_6$</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>

* [StGh03]
• How to combine the models?

- Correspondence
  - Explicit
  - Implicit
    - Consensus Function
      - Optimization Method
      - Object-based
      - Cluster-based
      - Object-Cluster-based
    - Representation
HyperGraph-Partitioning Algorithm (HGPA)*

• Hypergraph representation and clustering
  – Each node denotes an object
  – A hyperedge is a generalization of an edge in that it can connect any number of nodes
  – For objects that are put into the same cluster by a clustering algorithm, draw a hyperedge connecting them
  – Partition the hypergraph by minimizing the number of cut hyperedges
  – Each component forms a consensus cluster

* [StGh03]
Hypergraph-Partitioning Algorithm (HGPA)

<table>
<thead>
<tr>
<th></th>
<th>$C_1$</th>
<th>$C_2$</th>
<th>$C_3$</th>
<th>$C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$v_1$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$v_2$</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>$v_3$</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$v_4$</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>$v_5$</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>$v_6$</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Hypergraph representation – a circle denotes a hyperedge
Bipartite Graph Partitioning*

• Hybrid Bipartite Graph Formulation
  – Summarize base model output in a bipartite graph
  – Lossless summarization—base model output can be reconstructed from the bipartite graph
  – Use spectral clustering algorithm to partition the bipartite graph
  – Time complexity $O(nkr)$—due to the special structure of the bipartite graph
  – Each component represents a consensus cluster

*[FeBr04]
Bipartite Graph Partitioning

<table>
<thead>
<tr>
<th></th>
<th>$C_1$</th>
<th>$C_2$</th>
<th>$C_3$</th>
<th>$C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$v_1$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$v_2$</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>$v_3$</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$v_4$</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>$v_5$</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>$v_6$</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>
• How to combine the models?
A Mixture Model of Consensus*

• Probability-based
  – Assume output comes from a mixture of models
  – Use EM algorithm to learn the model

• Generative model
  – The clustering solutions for each object are represented as nominal features--$v_i$
  – $v_i$ is described by a mixture of $k$ components, each component follows a multinomial distribution
  – Each component is characterized by distribution parameters $\theta_j$

*[PTJ05]*
EM Method

• Maximize log likelihood

\[ \sum_{i=1}^{n} \log \left( \sum_{j=1}^{k} \alpha_j P(v_i | \theta_j) \right) \]

• Hidden variables
  – \( z_i \) denotes which consensus cluster the object belongs to

• EM procedure
  – E-step: compute expectation of \( z_i \)
  – M-step: update model parameters to maximize likelihood
Table 1: Clustering ensemble and consensus solution

<table>
<thead>
<tr>
<th>y_1</th>
<th>( \pi_1 )</th>
<th>( \pi_2 )</th>
<th>( \pi_3 )</th>
<th>( \pi_4 )</th>
<th>( E[Z_{11}] )</th>
<th>( E[Z_{12}] )</th>
<th>Consensus</th>
</tr>
</thead>
<tbody>
<tr>
<td>y_1</td>
<td>2</td>
<td>B</td>
<td>X</td>
<td>( \beta )</td>
<td>0.999</td>
<td>0.001</td>
<td>1</td>
</tr>
<tr>
<td>y_2</td>
<td>2</td>
<td>A</td>
<td>X</td>
<td>( \alpha )</td>
<td>0.997</td>
<td>0.003</td>
<td>1</td>
</tr>
<tr>
<td>y_3</td>
<td>2</td>
<td>A</td>
<td>Y</td>
<td>( \beta )</td>
<td>0.943</td>
<td>0.057</td>
<td>1</td>
</tr>
<tr>
<td>y_4</td>
<td>2</td>
<td>B</td>
<td>X</td>
<td>( \beta )</td>
<td>0.999</td>
<td>0.001</td>
<td>1</td>
</tr>
<tr>
<td>y_5</td>
<td>1</td>
<td>A</td>
<td>X</td>
<td>( \beta )</td>
<td>0.999</td>
<td>0.001</td>
<td>1</td>
</tr>
<tr>
<td>y_6</td>
<td>2</td>
<td>A</td>
<td>Y</td>
<td>( \beta )</td>
<td>0.943</td>
<td>0.057</td>
<td>1</td>
</tr>
<tr>
<td>y_7</td>
<td>2</td>
<td>B</td>
<td>Y</td>
<td>( \alpha )</td>
<td>0.124</td>
<td>0.876</td>
<td>2</td>
</tr>
<tr>
<td>y_8</td>
<td>1</td>
<td>B</td>
<td>Y</td>
<td>( \alpha )</td>
<td>0.019</td>
<td>0.981</td>
<td>2</td>
</tr>
<tr>
<td>y_9</td>
<td>1</td>
<td>B</td>
<td>Y</td>
<td>( \beta )</td>
<td>0.260</td>
<td>0.740</td>
<td>2</td>
</tr>
<tr>
<td>y_{10}</td>
<td>1</td>
<td>A</td>
<td>Y</td>
<td>( \alpha )</td>
<td>0.115</td>
<td>0.885</td>
<td>2</td>
</tr>
<tr>
<td>y_{11}</td>
<td>2</td>
<td>B</td>
<td>Y</td>
<td>( \alpha )</td>
<td>0.124</td>
<td>0.876</td>
<td>2</td>
</tr>
<tr>
<td>y_{12}</td>
<td>1</td>
<td>B</td>
<td>Y</td>
<td>( \alpha )</td>
<td>0.019</td>
<td>0.981</td>
<td>2</td>
</tr>
</tbody>
</table>
Base Model Accuracy

Ensemble Accuracy

Number of Base Models $k$

[TLJ+04]
Take-away Message

• Clustering ensemble
• Different approaches to combine multiple clustering solutions to one solution