Gene Clustering & Classification

Young-Rae Cho
Associate Professor
Department of Computer Science
Baylor University

Overview

- Introduction to Gene Clustering
  - Partition-based Clustering Methods
  - Hierarchical Clustering Methods
  - Validation of Clustering
- Gene Classification & Sample Classification
What is Clustering?

- Cluster
  - A group of data objects
  - Similar (or related) to one another within the same group
  - Dissimilar (or unrelated) to the objects in different groups

- Clustering (or Cluster Analysis)
  - Finding similarities between data objects
  - Grouping similar data objects into the same clusters
  - Unsupervised learning: no pre-defined classes

- Applications
  - A stand-alone method for data analysis
  - A preprocessing step for other data analysis

Measuring Quality of Clustering

- High Quality Clusters have
  - High intra-class similarity: cohesiveness within clusters
  - Low inter-class similarity: separability between clusters

- Quality of Clustering Depends on
  - Clustering methods
    - Handling both cohesiveness and separability
    - Ability to discover hidden patterns
    - Defining “similar enough” – problem of determining a threshold
  - Data sets
    - Amount of data
    - Complexity of data type
    - High dimensionality
Gene Clustering

- **Definition**
  - Grouping the genes which have similar features

- **Purpose**
  - Finding a group of genes that perform the same functions

- **Features**
  - Sequence similarity
  - Motif inclusion
  - Structure similarity
  - Expression profile coherence
  - Interaction evidence

Microarray Experiment (1)

- **Microarrays**
  - Estimate gene expression levels under varying conditions / time points by measuring the amount of mRNA for that particular gene

- **Microarray Experiment Process**
  - Produce cDNA from mRNA (DNA is more stable)
  - Attach phosphor to cDNA to see when a particular gene is expressed
  - Different color phosphors are available to compare many samples at once
  - Hybridize cDNA over the microarray
  - Scan the microarray with a phosphor-illuminating laser
  - Illumination reveals transcribed genes
  - Scan microarray multiple times for the different color phosphor’s
Microarray Experiment (2)

Expression Data (1)

- **Microarray Results**
  - Green: expressed from controlled samples
  - Red: expressed from experimental samples
  - Yellow: expressed in both samples
  - Black: not expressed in either controlled or experimental samples

- **Expression Data**
  - The intensity of the colors is quantified into numeric values
  - Each gene has an expression value on each experimental condition or each time point

<table>
<thead>
<tr>
<th>Gene</th>
<th>Time X</th>
<th>Time Y</th>
<th>Time Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene 1</td>
<td>10</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>Gene 2</td>
<td>10</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>Gene 3</td>
<td>4</td>
<td>8.6</td>
<td>3</td>
</tr>
<tr>
<td>Gene 4</td>
<td>7</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>Gene 5</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>
Expression Data (2)

Data Conversion

Pairwise Distance Matrix

Clustering Expression Data

Cohesiveness and Separability Principles

Co-expression

- Coherent expression patterns of genes across experimental conditions or time points
- More co-expressed genes within the same clusters, and less co-expressed genes from different clusters
Distance / Similarity Measures

➤ Minkowski Distance
  • Given two data, \( X = \{x_1, x_2, \ldots, x_n\} \) and \( Y = \{y_1, y_2, \ldots, y_n\} \), on \( n \) dimensions,
    \[
    d = \left( \sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}
    \]
  • Euclidean distance when \( p=2 \), and Manhattan distance when \( p=1 \)

➤ Pearson Coefficient
  • Evaluates correlation between two data on \( n \) dimensions
    \[
    r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}
    \]
    - co-variance between \( X \) and \( Y \)
    - individual variance of \( X \) and \( Y \)
  • If \( r > 0 \), \( X \) and \( Y \) are positively correlated.
  • If \( r = 0 \), \( X \) and \( Y \) are independent.
  • If \( r < 0 \), \( X \) and \( Y \) are negatively correlated.

Overview

➤ Introduction to Gene Clustering
  • *Partition-Based Clustering Methods*
  • Hierarchical Clustering Methods
  • Validation of Clustering

➤ Gene Classification & Sample Classification
Partition-Based Methods

- Main Idea
  - Constructing the best partition of the data with \( n \) objects into \( k \) clusters

- Issue
  - Finding a partition that optimize the criterion of cluster quality:
    - high intra-class similarity (cohesiveness) and
    - low inter-class similarity (separability)

- Methods
  - Theoretical method: Enumerate exhaustively all possible partitions and select the best one
  - Heuristic method: \( k \)-means, \( k \)-medoids

---

\( k \)-Means

- Main Idea
  - Finding the best partition of \( n \) data objects into \( k \) clusters in which each object belongs to the cluster with the nearest mean

- Process
  1. Partition data objects randomly into \( k \) clusters.
  2. Compute the mean point of the objects in each cluster as a centroid
  3. Assign each object to the nearest centroid and generate \( k \) new clusters
  4. Repeat (2) and (3), until there is no change of the objects in each cluster
Example of $k$-Means

Random partition ($k=2$)

Compute cluster means

Re-assign objects

Compute cluster means

Re-assign objects

Strength and Weakness of $k$-Means

- **Strength**
  - Relatively efficient
    - time complexity $\mathcal{O}(n \text{ objects}, k \text{ clusters}, t \text{ iterations})$

- **Weakness**
  - Need to **specify** $k$, the number of clusters, in advance
  - Sensitive to **noise** and **outliers**
  - Not suitable to detect clusters with non-convex shapes
  - Sometimes fall into local optimum, not identifying global optimum of clusters
Overview

- Introduction to Gene Clustering
  - Partition-Based Clustering Methods
  - Hierarchical Clustering Methods
  - Validation of Clustering
- Gene Classification & Sample Classification

Hierarchical Methods

- Main Idea
  - Decomposing data objects into several levels of nested partitioning (tree of clusters)

  ![Hierarchical Clustering Diagram]

  - Bottom-up approach (agglomerative algorithm)
  - Top-down approach (divisive algorithm)
Agglomerative Algorithm

- **Process**
  - Start with all single-node clusters
  - Iteratively merge the closest (the most similar) clusters
  - Eventually, all nodes belong to one cluster.

**Distance Measures between Clusters**

- **Single-Link Distance:** \[ d(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x, y) \]
- **Complete-Link Distance:** \[ d(C_i, C_j) = \max_{x \in C_i, y \in C_j} d(x, y) \]
- **Average-Link Distance:** \[ d(C_i, C_j) = \frac{1}{n_i n_j} \sum_{x \in C_i} \sum_{y \in C_j} d(x, y) \]
- **Centroid Distance:** \[ d(C_i, C_j) = d(m_i, m_j) \] where \( m_i \) and \( m_j \) are means of \( C_i \) and \( C_j \)
Comparison of Distance Measures

- Single-Link Distance
- Complete-Link Distance
- Average-Link Distance
- Centroid Distance

Divisive Algorithm

- Process
  - Start with one single clusters with all nodes
  - Iteratively divide the farthest (the most dissimilar) clusters
  - Eventually, all clusters have a single node.
Strength and Weakness of Hierarchical Methods

- Strength
  - Not require the number of clusters, $k$, in advance

- Weakness
  - Require the stopping condition
  - Sensitive to noise
  - Not able to undo what was done previously
  - Not scalable, at least $O(n^2)$ where $n$ objects

Overview

- Introduction to Gene Clustering
  - Partition-Based Clustering Methods
  - Hierarchical Clustering Methods
    - Validation of Clustering
  - Gene Classification & Sample Classification
Cluster Validation

- **Definition**
  - Assessing the quality of clustering results

- **Why Validating?**
  - To avoid finding clusters formed by chance
  - To compare clustering algorithms
  - To choose clustering parameters

- **Methods**
  - External index: when "ground truth" is available
  - Internal index: when "ground truth" is unavailable

Internal Index

- **Error Measures**
  - Absolute error = $|x_i - x'_i|$
  - Squared error = $(x_i - x'_i)^2$

- **Sum of Squared Error (SSE)**
  - Measure of cohesiveness by within-cluster sum of squared error
    \[ WSS = \sum_{i} \sum_{x \in C_i} (x - m_i)^2 \]
  - Measure of separability by between-cluster sum of squared error
    \[ BSS = \sum_{i} |C_i| \cdot (m_i - m)^2 \]
  - Relationship between WSS and BSS?
External Index (1)

Notations
- \( N \): the total number of data objects
- \( C = \{C_1, C_2, \ldots, C_n\} \): the set of clusters reported by a clustering algorithm
- \( P = \{P_1, P_2, \ldots, P_m\} \): the set of "ground truth" clusters

Incident Matrix
- \((N \times N)\) matrix
- \( C_{ij} = 1 \) if two data objects \( O_i \) and \( O_j \) belong to the same cluster in \( C \)
- \( C_{ij} = 0 \) otherwise
- \( P_{ij} = 1 \) if \( O_i \) and \( O_j \) belong to the same "ground truth" cluster in \( P \)
- \( P_{ij} = 0 \) otherwise

External Index (2)

Result Categories
- \( SS \): \( C_{ij} = 1 \) and \( P_{ij} = 1 \) (agree)
- \( DD \): \( C_{ij} = 0 \) and \( P_{ij} = 0 \) (agree)
- \( SD \): \( C_{ij} = 1 \) and \( P_{ij} = 0 \) (disagree)
- \( DS \): \( C_{ij} = 0 \) and \( P_{ij} = 1 \) (disagree)

Rand Index
- \( \text{Rand} = \frac{|SS| + |DD|}{|SS| + |DD| + |SD| + |DS|} \)

Jaccard Index
- \( \text{Jaccard Coefficient} = \frac{|SS|}{|SS| + |SD| + |DS|} \)
**f-Measure**

- Recall & Precision
  - Comparison between an output cluster and a ground-truth cluster
  - Let an output cluster \(X\), and a ground-truth cluster \(Y\)
  - Recall (Sensitivity, or True positive rate) = \(\frac{|X \cap Y|}{|Y|}\)
  - Precision (Positive predictive value) = \(\frac{|X \cap Y|}{|X|}\)

- \(f\)-Measure
  - Harmonic mean of Recall and Precision
  - \(f\)-measure = \(2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}\)

**Statistical \(p\)-Value**

- \(p\)-value of Hyper-Geometric Distribution
  - Let the set of all data objects, \(N\)
  - Let an output cluster \(X\), and a ground-truth cluster \(Y\)
  - Probability that at least \(k\) data objects in \(X\) are included in \(Y\)
  - \(\sum_{i=k}^{\min\{|Y|, |X|\}} \binom{|X|}{i} \binom{|N|-|Y|}{|X|-i} \binom{|Y|}{i} \binom{|N|-|Y|}{|X|-i}^{-1} \) where \(k = |X \cap Y|\)

- A low \(p\)-value indicates it is less probable that the cluster \(X\) is produced by chance
  - \(- \log(p)\) is usually used for clustering evaluation
Overview

- Introduction to Gene Clustering
  - Partition-Based Clustering Methods
  - Hierarchical Clustering Methods
  - Validation of Clustering
- Gene Classification & Sample Classification

Supervised vs. Unsupervised Learning

- Supervised Learning
  - Called classification
  - Training data (observations, measurement, etc.) are given
  - Training data include class labels predefined
  - Find rules or models of class labels of training data
  - New data are classified based on the rules or models

- Unsupervised Learning
  - Called clustering
  - No training data are given
  - New data are classified without any training data
Classification vs. Prediction

- **Classification**
  - Training class labels in attributes of a training data set
  - Predicts class labels of a new data set based on the rules or models of class labels of the training data set

- **Prediction**
  - Modeling continuous-valued functions for a data set
  - Predicts unknown or missing values in the data set

Gene Classification (1)

- **Definition**
  - Finding the functions of unknown genes by training expression data of known (functionally characterized) genes

- **Expression Data**
  - Expression levels (numeric values) for each gene across time points
  - Time-series expression data for each gene (Gene-Time expression data)
Gene Classification (2)

- Assumption
  - Genes having the same functions are co-expressed
    (co-expression: coherent expression patterns of genes across time points)

- Process
  1. Finding coherent patterns of time-series expression profiles for each class → Training step
  2. Matching the time-series expression profile of an unknown gene to the patterns → Predicting step

Time-Series Expression Data

- Examples

![Graphs showing expression patterns over time](image-url)
Sample Classification

- Definition
  - Finding the class (disease) of unknown samples by training expression data of known samples

- Expression Data
  - Expression levels (numeric values) for each gene across samples at a certain time point
  - Gene-Sample expression data

- Process
  1. Informative gene selection in a training data set
  2. Class prediction of unknown samples

Informative Gene Selection (1)

- Purpose
  - Eliminating irrelevant genes
  - Selecting significant (informative) genes for class prediction

- Examples
  - gene 1
  - gene 2
  - ideal case
Informative Gene Selection (2)

- Statistical Approach
  - Selects genes expressed more differently between 2 classes
  - Computes statistical information (mean, standard deviation) of expression values of the samples in each class
  - Measures the correlation of expressions between 2 classes for each gene
    \[ P(g) = \frac{\mu_1(g) - \mu_2(g)}{\sigma_1(g) + \sigma_2(g)} \]
  - Ranks genes by the correlation metric

Informative Gene Selection (3)

- Example
Class Prediction

- Weighted Voting Approach
  - For each unknown sample, each informative gene votes for either class 1 or class 2, based on whether its expression value in the sample is closer to $\mu_1$ or $\mu_2$.
  - Computes the weighted vote of a gene $g$ by $w(g)v(g)$
    
    \[
    w(g) = P(g) \\
    v(g) = |x(g) - (\mu_1(g) + \mu_2(g))/2|
    \]
    
    where $x(g)$ is the expression value of $g$.

Advanced Topics

- Sample Classification in Time-Series Data
  - Classifying samples using time-series expression data for each gene (gene-sample-time expression data)

- Bi-Clustering
  - Simultaneous clustering of both genes and conditions (or time points)
Questions?

- Lecture Slides are found on the Course Website,
  web.ecs.baylor.edu/faculty/cho/3360