Clustering
Community Detection in Social Media

Symeon Papadopoulos
CERTH-ITI, 22 June 2011

Customer Relation Management

- Partitioning customers into groups such that customers within a group are similar in some aspects
- A manager can be assigned to a group
- Customized products and services can be developed
What Is Clustering?

- Group data into clusters
  - Similar to one another within the same cluster
  - Dissimilar to the objects in other clusters
  - Unsupervised learning: no predefined classes
Requirements of Clustering

- Scalability
- Ability to deal with various types of attributes
- Discovery of clusters with arbitrary shape
- Minimal requirements for domain knowledge to determine input parameters
Data Matrix

• For memory-based clustering
  – Also called object-by-variable structure
• Represents n objects with p variables (attributes, measures)
  – A relational table

\[
\begin{bmatrix}
  x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\
  \vdots & \ddots & \vdots & \ddots & \vdots \\
  x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\
  \vdots & \ddots & \vdots & \ddots & \vdots \\
  x_{n1} & \cdots & x_{nf} & \cdots & x_{np}
\end{bmatrix}
\]
Dissimilarity Matrix

- For memory-based clustering
  - Also called object-by-object structure
  - Proximities of pairs of objects
  - $d(i, j)$: dissimilarity between objects $i$ and $j$
  - Nonnegative
  - Close to 0: similar

\[
\begin{bmatrix}
0 & & & \\
d(2,1) & 0 & & \\
d(3,1) & d(3,2) & 0 & \\
& & & \\
& & & \\
d(n,1) & d(n,2) & \cdots & 0
\end{bmatrix}
\]
How Good Is Clustering?

• Dissimilarity/similarity depends on distance function
  – Different applications have different functions
• Judgment of clustering quality is typically highly subjective
Types of Data in Clustering

- Interval-scaled variables
- Binary variables
- Nominal, ordinal, and ratio variables
- Variables of mixed types
Interval-valued Variables

• Continuous measurements of a roughly linear scale
  – Weight, height, latitude and longitude coordinates, temperature, etc.

• Effect of measurement units in attributes
  – Smaller unit $\rightarrow$ larger variable range $\rightarrow$ larger effect to the result
  – Standardization + background knowledge
Standardization

- Calculate the mean absolute deviation
  \[ s_f = \frac{1}{n} (|x_{1f} - m_f| + |x_{2f} - m_f| + \ldots + |x_{nf} - m_f|) \]
  \[ m_f = \frac{1}{n} (x_{1f} + x_{2f} + \ldots + x_{nf}) \]

- Calculate the standardized measurement (z-score)
  \[ z_{if} = \frac{x_{if} - m_f}{s_f} \]

- Mean absolute deviation is more robust
  - The effect of outliers is reduced but remains detectable
Similarity and Dissimilarity

- Distances are normally used measures
- Minkowski distance: a generalization
  \[ d(i, j) = q \sqrt{\sum_{i=1}^{p} |x_{i} - x_{j}|^q} \quad (q > 0) \]
  - If \( q = 2 \), \( d \) is Euclidean distance
  - If \( q = 1 \), \( d \) is Manhattan distance
  - If \( q = \infty \), \( d \) is Chebyshev distance
- Weighed distance
  \[ d(i, j) = q \sqrt{\sum_{i=1}^{p} w_i |x_{i} - x_{j}|^q} \quad (q > 0) \]
Manhattan and Chebyshev Distance

Manhattan Distance

Chebyshev Distance

\[ D_{\text{Chebyshev}} = \max_i (p_i - q_i) = \lim_{k \to \infty} \left( \sum_{i=1}^{n} |p_i - q_i|^k \right)^{1/k} \]

When \( n = 2 \), chess-distance

\[ D_{\text{Chess}} = \max (|x_2 - x_1|, |y_2 - y_1|) \]

Picture from Wikipedia

http://brainking.com/images/rules/chess/02.gif
Properties of Minkowski Distance

• Nonnegative: \( d(i,j) \geq 0 \)
• The distance of an object to itself is 0
  \( d(i,i) = 0 \)
• Symmetric: \( d(i,j) = d(j,i) \)
• Triangular inequality
  \( d(i,j) \leq d(i,k) + d(k,j) \)
Binary Variables

- A contingency table for binary data
- Symmetric variable: each state carries the same weight
  - Invariant similarity
- Asymmetric variable: the positive value carries more weight
  - Noninvariant similarity (Jacard)

\[
d(i, j) = \frac{r + s}{q + r + s + t}
\]

<table>
<thead>
<tr>
<th>Object i</th>
<th>1</th>
<th>0</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>q</td>
<td>r</td>
<td>q+r</td>
</tr>
<tr>
<td>0</td>
<td>s</td>
<td>t</td>
<td>s+t</td>
</tr>
<tr>
<td>Sum</td>
<td>q+s</td>
<td>r+t</td>
<td>p</td>
</tr>
</tbody>
</table>
Nominal Variables

• A generalization of the binary variable in that it can take more than 2 states, e.g., Red, yellow, blue, green

• Method 1: simple matching
  – M: # of matches, p: total # of variables

\[
d(i, j) = \frac{p - m}{p}
\]

• Method 2: use a large number of binary variables
  – Creating a new binary variable for each of the M nominal states
Ordinal Variables

- An ordinal variable can be discrete or continuous
- Order is important, e.g., rank
- Can be treated like interval-scaled
  - Replace $x_{if}$ by their rank
  - Map the range of each variable onto $[0, 1]$ by replacing the i-th object in the f-th variable by
    \[ z_{if} = \frac{r_{if} - 1}{M_f - 1} \]
  - Compute the dissimilarity using methods for interval-scaled variables

\[ r_{if} \in \{1, ..., M_f \} \]
Ratio-scaled Variables

- Ratio-scaled variable: a positive measurement on a nonlinear scale
  - E.g., approximately at exponential scale, such as $Ae^{Bt}$
- Treat them like interval-scaled variables?
  - Not a good choice: the scale can be distorted!
- Apply logarithmic transformation, $y_{if} = \log(x_{if})$
- Treat them as continuous ordinal data, treat their rank as interval-scaled
Variables of Mixed Types

- A database may contain all the six types of variables
  - Symmetric binary, asymmetric binary, nominal, ordinal, interval and ratio

- One may use a weighted formula to combine their effects

\[ d(i, j) = \frac{\sum_{f=1}^{P} \delta_{ij}(f) d_{ij}(f)}{\sum_{f=1}^{P} \delta_{ij}(f)} \]
Clustering Methods

- K-means and partitioning methods
- Hierarchical clustering
- Density-based clustering
- Grid-based clustering
- Pattern-based clustering
- Other clustering methods
Partitioning Algorithms: Ideas

• Partition n objects into k clusters
  – Optimize the chosen partitioning criterion
• Global optimal: examine all possible partitions
  – \( (k^n-(k-1)^n-\ldots-1) \) possible partitions, too expensive!
• Heuristic methods: k-means and k-medoids
  – K-means: a cluster is represented by the center
  – K-medoids or PAM (partition around medoids): each cluster is represented by one of the objects in the cluster
K-means

• Arbitrarily choose k objects as the initial cluster centers

• Until no change, do

  – (Re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster

  – Update the cluster means, i.e., calculate the mean value of the objects for each cluster
K-Means: Example

K=2
Arbitrarily choose K object as initial cluster center

Assign each object to the most similar center

Update the cluster means

reassign

Update the cluster means

reassign

reassign

Update the cluster means
Pros and Cons of K-means

- Relatively efficient: $O(tkn)$
  - $n$: # objects, $k$: # clusters, $t$: # iterations; $k, t << n$.
- Often terminate at a local optimum
- Applicable only when mean is defined
  - What about categorical data?
- Need to specify the number of clusters
- Unable to handle noisy data and outliers
- Unsuitable to discover non-convex clusters
Variations of the K-means

• Aspects of variations
  – Selection of the initial k means
  – Dissimilarity calculations
  – Strategies to calculate cluster means

• Handling categorical data: k-modes
  – Use mode instead of mean
    • Mode: the most frequent item(s)
  – A mixture of categorical and numerical data: k-prototype method

• EM (expectation maximization): assign a probability of an object to a cluster (will be discussed later)
A Problem of K-means

• Sensitive to outliers
  – Outlier: objects with extremely large values
  • May substantially distort the distribution of the data

• K-medoids: the most centrally located object in a cluster
PAM: A K-medoids Method

- PAM: partitioning around Medoids
- Arbitrarily choose k objects as the initial medoids
- Until no change, do
  - (Re)assign each object to the cluster to which the nearest medoid
  - Randomly select a non-medoid object o’, compute the total cost, S, of swapping medoid o with o’
  - If S < 0 then swap o with o’ to form the new set of k medoids
Swapping Cost

- Measure whether o’ is better than o as a medoid
- Use the squared-error criterion
  \[ E = \sum_{i=1}^{k} \sum_{p \in C_i} d(p, o_i)^2 \]
  - Compute \( E_{o'} - E_o \)
  - Negative: swapping brings benefit
PAM: Example

Do loop

Until no change

Total Cost = 26

Randomly select a nonmedoid object, \( O_{\text{random}} \)

Assign each remaining object to nearest medoids

Assign each remaining object to nearest medoids

Swapping \( O \) and \( O_{\text{random}} \)

If quality is improved.

K=2
Pros and Cons of PAM

• PAM is more robust than k-means in the presence of noise and outliers
  – Medoids are less influenced by outliers
• PAM is efficient for small data sets but does not scale well for large data sets
  – $O(k(n-k)^2)$ for each iteration
Careful Initialization: K-means++

- Select one center uniformly at random from the data sets.
- For each object $p$ that is not the chosen center, choose the object as a new center with probability proportional to $\text{dist}(p)^2$, where $\text{dist}(p)$ is the distance from $p$ to the closest center that has already been chosen.
- Repeat the above step until $k$ centers are selected.
To-Do List

• Read Chapters 10.1 and 10.2
• Find out how to use k-means in WEKA
• (for graduate students only) find out how to use k-means in SPARK MLlib