Cluster Analysis in Data Mining

Part II

Andrew Kusiak
Intelligent Systems Laboratory
2139 Seamans Center
The University of Iowa
Iowa City, IA 52242 - 1527
andrew-kusiak@uiowa.edu
http://www.icaen.uiowa.edu/~ankusiak
Tel. 319-335 5934

Cluster Analysis

Decomposition

Aggregation

Grouping
Models

- Matrix formulation
- Mathematical programming formulation
- Graph formulation

Cluster Representation 1

Mutually exclusive regions
Cluster Representation 2

Overlapping regions

Cluster Representation 3

Hierarchy
Cluster Representation 4

<table>
<thead>
<tr>
<th>Clusters</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>.6</td>
<td>.3</td>
<td>.1</td>
</tr>
<tr>
<td>B</td>
<td>.5</td>
<td>.2</td>
<td>.3</td>
</tr>
<tr>
<td>C</td>
<td>.6</td>
<td>.3</td>
<td>.1</td>
</tr>
<tr>
<td>D</td>
<td>.3</td>
<td>.6</td>
<td>.1</td>
</tr>
<tr>
<td>E</td>
<td>.2</td>
<td>.1</td>
<td>.7</td>
</tr>
<tr>
<td>F</td>
<td>.1</td>
<td>.2</td>
<td>.7</td>
</tr>
<tr>
<td>G</td>
<td>.2</td>
<td>.6</td>
<td>.2</td>
</tr>
<tr>
<td>H</td>
<td>.3</td>
<td>.6</td>
<td>.1</td>
</tr>
</tbody>
</table>

Fuzzy clusters

Two types of data

- Object with decisions
- Object without decisions
Clustering Data with Decisions

Method 1

• Ignore decisions
• Treat the objects as data without decisions

Method 2

• Group objects according to the decisions
• Treat each group of objects with decisions as a separate data set

Solving the Clustering Problem: Binary Matrix Formulation

• Similarity coefficient methods
• Sorting based algorithms
• Bond energy algorithm
• Cost-based method
• Cluster identification algorithm
• Extended cluster identification algorithm
Clustering Algorithms

Desired Properties

• Low computational time complexity
• Structure enhancement
• Enhancement of human interaction and simplification of interface with software
• Generalizability

Similarity Coefficient Method

\[
S_{ij} = \frac{\sum_{k=1}^{n} \delta_1(a_{ik}, a_{jk})}{n} \frac{\sum_{k=1}^{n} \delta_2(a_{ik}, a_{jk})}{n}
\]

where \( \delta_1(a_{ik}, a_{jk}) = \begin{cases} 
1 & \text{if } a_{ik} = a_{jk} \\
0 & \text{otherwise}
\end{cases} \)

Generalization to qualitative features
\[ \delta_2 (a_{ik}, a_{jk}) = \begin{cases} 
0 & \text{if } a_{ik} = a_{jk} \\
1 & \text{otherwise} 
\end{cases} \]

\[ n = \text{the number of features} \]

---

**Example**

Consider

\[
\begin{array}{cccccc}
1 & 2 & 3 & 4 & 5 \\
\text{GO-1} & \{ & 1 & \text{Yes} & 1 & 1 \\
& 2 & \text{Yes} & 1 \\
\text{GO-2} & \{ & 3 & 1 & 1 & 1 \\
& 4 & 1 & 1 \\
\end{array}
\]

---
The University of Iowa
Intelligent Systems Laboratory

Tree Representing Similarity Coefficients

Object number

Similarity (%)
Cluster Identification Algorithm

Decomposition Problem

Decompose an object-feature incidence matrix into mutually separable submatrices (groups of objects and groups of features) with the minimum number of overlapping objects (or features) subject to the following constraints:

Constraint C1: Empty groups of objects or features are not allowed.
Constraint C2: The number of objects in a group does not exceed an upper limit, b (or alternatively, the number of features in a group does not exceed, d).
Cluster Identification Algorithm

Step 0. Set iteration number \( k = 1 \).

Step 1. Select row \( i \) of incidence matrix \([aij](k)\) and draw a horizontal line \( hi \) through it (\([aij](k)\) is read: matrix \([aij]\) at iteration \( k \)).

Step 2. For each entry of \( * \) crossed by the horizontal line \( hi \) draw a vertical line \( vj \).

Step 3. For each entry of \( * \) crossed-once by the vertical line \( vj \) draw a horizontal line \( hk \).

Step 4. Repeat steps 2 and 3 until there are no more crossed-once entries of \( * \) in \([aij](k)\). All crossed-twice entries \( * \) in \([aij](k)\) form row cluster \( RC-k \) and column cluster \( CC-k \).

Step 5. Transform the incidence matrix \([aij](k)\) into \([aij](k+1)\) by removing rows and columns corresponding to the horizontal and vertical lines drawn in steps 1 through 4.

Step 6. If matrix \([aij](k+1)\) = 0 (where 0 denotes a matrix with all empty elements), stop; otherwise set \( k = k + 1 \) and go to step 1.

Example 1

<table>
<thead>
<tr>
<th>Incidence matrix</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2 3 4 5 6 7 8</td>
<td>1 2 3 4 5 6 7 8</td>
</tr>
<tr>
<td>1 [ * * * ]</td>
<td>1 [ * * * ]</td>
</tr>
<tr>
<td>2 [ * ]</td>
<td>2 [ * ]</td>
</tr>
<tr>
<td>3 [ * ]</td>
<td>3 [ * ]</td>
</tr>
<tr>
<td>4 [ * ]</td>
<td>4 [ * ]</td>
</tr>
<tr>
<td>5 [ * ]</td>
<td>5 [ * ]</td>
</tr>
<tr>
<td>6 [ * ]</td>
<td>6 [ * ]</td>
</tr>
<tr>
<td>7 [ * * * * ]</td>
<td>7 [ * * * * ]</td>
</tr>
</tbody>
</table>

The University of Iowa
Intelligent Systems Laboratory

Page 10
Next Step

Delete all double-crossed elements
Resultant Matrix

\[
\begin{bmatrix}
1 & 4 & 6 & 7 \\
2 & * & * \\
3 & * & * \\
4 & * & * \\
6 & * \\
\end{bmatrix}
\]
Final decomposition result

<table>
<thead>
<tr>
<th></th>
<th>CC-1</th>
<th>CC-2</th>
<th>CC-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>4</td>
<td>7</td>
</tr>
</tbody>
</table>

**RC-1**

1  
5  
7

**RC-2**

2  
4

**RC-3**

3  
6
Three types of matrices

(a) decomposable matrix
(b) non-decomposable matrix with overlapping features
(c) non-decomposable matrix with objects

Extended CI Algorithm

Step 0. Set iteration number $k = 1$.
Step 1. Select those objects (rows of matrix $[a_{ij}]^{(k)}$) that based on the user's expertise, are potential candidates for inclusion in machine cell $GO-k$. Draw a horizontal line $h_i$ through each row of matrix $[a_{ij}]^{(k)}$ corresponding to these objects. In the absence of the user's expertise any machine can be selected.
Step 2. For each column in $[a_{ij}]^{(k)}$ corresponding to entry 1, single-crossed by any of the horizontal lines $h_i$, draw a vertical line $v_j$.
Step 3. For each row in $[a_{ij}]^{(k)}$ corresponding to the entry 1, single-crossed by the vertical line $v_j$, drawn in Step 2, draw a horizontal line $h_i$. Based on the objects corresponding to all the horizontal lines drawn in Step 1 and Step 3, a temporary machine cell $GO-k$, is formed.
If the user's expertise indicates that some of the objects cannot be included in the temporary group GO-k, erase the corresponding horizontal lines in the matrix \( [a_{ij}]^{(k)} \). Removal of these horizontal lines results in group GO\(^-\)k.

Delete from matrix \( [a_{ij}]^{(k)} \) features (columns) that are contained in at least one of the objects already included in GO-k.

Place these features on a separate list.

Draw a vertical line \( v_j \) through each single-crossed entry \( 1 \) in \( [a_{ij}]^{(k)} \) which does not involve any other objects than those included in GO-k.

**Step 4.** For all the double-crossed entries \( 1 \) in \( [a_{ij}]^{(k)} \), form a cluster GO-k and a group GF-k.

Step 5. Transform the incidence matrix \( [a_{ij}]^{(k)} \) into \( [a_{ij}]^{(k+1)} \) by removing all the rows and columns included in GO-k and GF-k, respectively.

Step 6. If matrix \( [a_{ij}]^{(k+1)} = 0 \) (where 0 denotes a matrix with all elements equal to zero), stop; otherwise set \( k = k + 1 \) and go to step 1.
Example

Extended CI Algorithm

Constraints:
Max |GO| = 4
Objects 1 and 4 in the cluster

<table>
<thead>
<tr>
<th>Feature number</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2 3 4 5 6 7 8 9 10 11</td>
</tr>
<tr>
<td>1 1 1 1 1 1 1 1 1 1 1</td>
</tr>
<tr>
<td>1 1 1 1 1 1 1 1 1 1 1</td>
</tr>
<tr>
<td>1 1 1 1 1 1 1 1 1 1 1</td>
</tr>
<tr>
<td>1 1 1 1 1 1 1 1 1 1 1</td>
</tr>
</tbody>
</table>

Step 0. Set iteration number $k = 1$.
Step 1. Since user's expertise indicates that objects 1 and 4 should be included cluster GO-1, two horizontal lines $h_1$ and $h_4$ are drawn.

<table>
<thead>
<tr>
<th>Feature number</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2 3 4 5 6 7 8 9 10 11</td>
</tr>
<tr>
<td>1 1 1 1 1 1 1 1 1 1 1</td>
</tr>
<tr>
<td>1 1 1 1 1 1 1 1 1 1 1</td>
</tr>
<tr>
<td>1 1 1 1 1 1 1 1 1 1 1</td>
</tr>
<tr>
<td>1 1 1 1 1 1 1 1 1 1 1</td>
</tr>
</tbody>
</table>
Step 2. For columns 1, 2, 3, 6, and 7 crossed by the horizontal lines h1 and h4, five vertical lines, v1, v2, v3, v6, and v7 are drawn.

Step 3. Three horizontal lines, h2, h6, and h7 are drawn through rows 2, 5, and 7 corresponding of the single-crossed elements 1.
• A temporary machine cell GO'-1 with objects \{1, 2, 4, 6, 7\} is formed.
• objects 2 and 6 are excluded from in GO'-1 as they include less double-crossed (committed) elements than object 7. The horizontal lines h2 and h6 are erased.

\[
\begin{array}{ccccccc}
 & 2 & 3 & 5 & 6 & 7 & 8 & 10 & 11 \\
\hline
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
2 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
3 &  &  &  &  &  &  &  &  \\
4 &  &  &  &  &  &  &  &  \\
5 &  &  &  &  &  &  &  &  \\
6 &  &  &  &  &  &  &  &  \\
7 &  &  &  &  &  &  &  &  \\
\end{array}
\]

Step 4. The double-crossed entries 1 of matrix indicate:
• Object cluster GO-1 = \{1, 4, 7\} and
• Feature cluster GF-1 = \{2, 3, 6, 7\}

Step 5. Matrix is transformed into the matrix below.

\[
\begin{array}{cccc}
5 & 8 & 10 & 11 \\
\hline
2 & 1 & 1 & 1 \\
3 &  &  &  \\
5 & 1 & 1 & 1 \\
6 & 1 & 1 & 1 \\
\end{array}
\]
Step 6.
Set $k = k + 1 = 2$ and go to Step 1.
The second iteration ($k = 2$) results in:
- Object cluster $GO-2 = \{2, 3, 5, 6\}$ and
- Feature cluster $GF-2 = \{5, 8, 10, 11\}$

The final result

```
0 0 0 0 1 0 1 1 1 1 0 0
0 0 0 0 0 1 0 0 0 1 0 0
0 0 0 0 0 0 1 0 1 1 0 0
0 0 0 0 0 0 0 1 1 1 0 0
0 0 0 0 0 0 0 0 1 1 0 0
```

The University of Iowa
Intelligent Systems Laboratory
The p -Median Model

Notation

\[ m = \text{number of objects} \]
\[ n = \text{number of features} \]
\[ p = \text{number of clusters} \]
\[ x_{ij} = \begin{cases} 
1 & \text{if feature } i \text{ belongs to cluster } j \\
0 & \text{otherwise} 
\end{cases} \]
\[ d_{ij} = \text{distance between objects } i \text{ and } j \]

Model

\[ \min \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij} x_{ij} \]
subject to:

\[ \sum_{j=1}^{n} x_{ij} = 1 \text{ for all } i = 1, \ldots, n \]
\[ \sum_{j=1}^{n} x_{jj} = p \]
\[ x_{ij} \leq x_{jj} \text{ for all } i = 1, \ldots, n, \quad j = 1, \ldots, n \]
\[ x_{ij} = 0,1 \text{ for all } i = 1, \ldots, n, \quad j = 1, \ldots, n \]
Example

Hamming Distance

\[ d_{ij} = \sum_{k=1}^{m} d(aki, akj) \]

Two objects

\[ \mathbf{O}_1 = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} \quad \mathbf{O}_5 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} \]

Hamming distance

\[ d_{15} = 1 + 0 + 1 + 1 + 0 = 4 \]

Decision variable

\[ X_{ij} \]

Object (or feature) number

Object (or feature) number

Cluster number

Note: Distances can be computed for objects or features.
Example

Given

\[ p = 2 \]

\[ m \]

\[ d_{ij} = \sum_{k=1}^{m} d(aki, akj) \]

<table>
<thead>
<tr>
<th>Feature number</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
[d_{ij}] = \begin{bmatrix}
0 & 4 & 0 & 4 & 4 \\
4 & 0 & 4 & 0 & 1 \\
0 & 4 & 0 & 4 & 3 \\
4 & 0 & 4 & 0 & 1 \\
4 & 1 & 3 & 1 & 0
\end{bmatrix}
\]

Input file

\[ \text{MIN } 4X12+4X14+3X15+4X21+4X23+1X25+4X32+4X34+3X35+4X41+4X43+1X45+3X51+1X52+3X53+1X54 \]

\[ \text{SUBJECT TO } \sum_{j=1}^{n} x_{ij} = 1 \quad \text{for all } i = 1, ..., n \]

\[ X11+X12+X13+X14+X15=1 \]
\[ X21+X22+X23+X24+X25=1 \]
\[ X31+X32+X33+X34+X35=1 \]
\[ X41+X42+X43+X44+X45=1 \]
\[ X51+X52+X53+X54+X55=1 \]
\[ \text{Feature number} \]

\[
[d_{ij}] = \begin{bmatrix}
0 & 4 & 0 & 4 & 4 \\
2 & 4 & 0 & 4 & 0 \\
3 & 0 & 4 & 0 & 4 \\
4 & 4 & 0 & 4 & 0 \\
5 & 4 & 1 & 3 & 1
\end{bmatrix}
\]
Constraint \( x_{ij} \leq x_{jj} \)

\[
\begin{align*}
X_{21} - X_{11} & \leq 0 \\
X_{31} - X_{11} & \leq 0 \\
X_{41} - X_{11} & \leq 0 \\
X_{51} - X_{11} & \leq 0 \\
X_{12} - X_{22} & \leq 0 \\
X_{32} - X_{22} & \leq 0 \\
X_{42} - X_{22} & \leq 0 \\
X_{52} - X_{22} & \leq 0 \\
X_{13} - X_{33} & < 0 \\
X_{23} - X_{33} & \leq 0 \\
X_{43} - X_{33} & \leq 0 \\
X_{53} - X_{33} & \leq 0 \\
X_{14} - X_{44} & \leq 0 \\
X_{24} - X_{44} & \leq 0 \\
X_{34} - X_{44} & \leq 0 \\
X_{54} - X_{44} & \leq 0 \\
X_{15} - X_{55} & \leq 0 \\
X_{25} - X_{55} & \leq 0 \\
X_{35} - X_{55} & \leq 0 \\
X_{45} - X_{55} & \leq 0 \\
X_{55} & \leq 0 \\
\end{align*}
\]

\( END \)

\( \text{INTEGER 25} \)

Solution \( x_{11} = 1, \ x_{31} = 1 \)
\( x_{24} = 1, \ x_{44} = 1, \ x_{54} = 1 \)

Based on the definition of \( x_{ij} \), two groups of features are formed:

\[
\begin{align*}
\text{GF-1} & = \{1, 3\} & \text{GF-2} & = \{2, 4, 5\} \\
\text{GO-1} & = \{2, 4\} & \text{GO-2} & = \{1, 3\} \\
\end{align*}
\]

The University of Iowa
Intelligent Systems Laboratory
Solution: $x_{13}=1$, $x_{23} = 1$
$x_{35} = 1$, $x_{45} = 1$, $x_{55} = 1$