Quest: A Generalized Motif Bicluster Algorithm

Sebastian Kaiser and Friedrich Leisch

Institut für Statistik
Ludwig-Maximilians-Universität München

UseR 2009, 09.07.2009, Rennes, France
Overview

Outline:

I. Introduce Biclustering

II. New Bicluster Algorithm

III. New Developments in the biclust Package

IV. Example

V. Summary and Future Work
I. Biclustering

Why Biclustering?

- Simultaneous clustering of 2 dimensions
- Large datasets where traditional clustering of columns or rows leads to diffuse results
- Only parts of the data influence each other
I. Biclustering

Initial Situation:

Two-Way Dataset

<table>
<thead>
<tr>
<th></th>
<th>$c_1$</th>
<th>$\ldots$</th>
<th>$c_i$</th>
<th>$\ldots$</th>
<th>$c_m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r_1$</td>
<td>$a_{11}$</td>
<td>$\ldots$</td>
<td>$a_{i1}$</td>
<td>$\ldots$</td>
<td>$a_{m1}$</td>
</tr>
<tr>
<td>$r_j$</td>
<td>$a_{1j}$</td>
<td>$\ldots$</td>
<td>$a_{ij}$</td>
<td>$\ldots$</td>
<td>$a_{mj}$</td>
</tr>
<tr>
<td>$r_n$</td>
<td>$a_{1n}$</td>
<td>$\ldots$</td>
<td>$a_{in}$</td>
<td>$\ldots$</td>
<td>$a_{mn}$</td>
</tr>
</tbody>
</table>
I. Biclustering

Goal:

Finding subgroups of rows and columns which are as similar as possible to each other and as different as possible to the rest.
More than one bicluster? Most Bicluster Algorithms are iterative. To find the next bicluster given n-1 found biclusters you have to either

- ignore the n-1 already found biclusters,
- delete rows and/or columns of the found biclusters or
- mask the found biclusters with random values.
II. Bicluster Algorithms: In the Package

Chosen sample of algorithms in order to cover most bicluster outcomes.

**Bimax** (Barkow et al., 2006): Groups with ones in binary matrix

**CC** (Cheng and Church, 2000): Constant values

**Plaid** (Turner et al., 2005): Constant values over rows or columns

**Spectral** (Kluger et al., 2003): Coherent values over rows and columns

**Xmotif** (Murali and Kasif, 2003): Coherent correlation over rows and columns
II. Bicluster Algorithms

Bimax

\[
\begin{array}{cccccc}
1 & * & * & 1 & * & 1 & * \\
* & * & * & * & * & * & * \\
* & * & * & * & * & * & * \\
1 & * & * & 1 & * & 1 & * \\
* & * & * & * & * & * & * \\
1 & * & * & 1 & * & 1 & * \\
* & * & * & * & * & * & * \\
\end{array}
\]

\[
\begin{array}{cccc}
1 & 1 & 1 \\
1 & 1 & 1 \\
1 & 1 & 1 \\
\end{array}
\]

\[
\begin{array}{cccc}
* & * & * & * \\
* & * & * & * \\
* & * & * & * \\
\end{array}
\]

\[
\begin{array}{cccc}
* & * & * & * \\
* & * & * & * \\
* & * & * & * \\
\end{array}
\]

• Finds subgroups of ones in a binary data matrix.

• Suitable if only one kind of outcome is interesting.
II. Bicluster Algorithms

Xmotif

- Finds subgroups of equal outcomes.
- Suitable if equal nominal or ordinal values are wanted.
II. Bicluster Algorithms

Quest (nominal)

- Finds subgroups of equal outcomes over the variables.
- Suitable if equal patterns of nominal or ordinal values are wanted.
II. Bicluster Algorithms

Quest (ordinal)

- Finds subgroups of outcomes inside a given interval or a given size of interval over the variables.

- Suitable if similar patterns of ordinal or continuous values are wanted.
II. Bicluster Algorithms

## Quest (continuous)

<table>
<thead>
<tr>
<th></th>
<th>74</th>
<th>0.23</th>
<th>0.35</th>
<th>80.5</th>
<th>0.27</th>
<th>77</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>80.5</td>
<td>0.23</td>
<td>0.35</td>
<td>80.5</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td></td>
<td>*</td>
<td>*</td>
<td>80.5</td>
<td>*</td>
<td>77</td>
</tr>
<tr>
<td></td>
<td></td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>77</td>
<td>0.27</td>
<td>0.35</td>
<td>77</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td></td>
<td>*</td>
<td>77</td>
<td>*</td>
<td>77</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

⇒

<table>
<thead>
<tr>
<th>74</th>
<th>0.23</th>
<th>0.35</th>
<th>80.5</th>
<th>0.27</th>
<th>77</th>
</tr>
</thead>
<tbody>
<tr>
<td>74</td>
<td>0.23</td>
<td>0.35</td>
<td>80.5</td>
<td>0.27</td>
<td>77</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| *   |      |      |      |      |    *
|     |      |      |      |      |    *
|     |      |      |      |      |    *
| 80.5| 0.35 | 0.27 | 77   | 0.27 | 77  |
| 80.5| 0.35 | 0.27 | 77   | 0.27 | 77  |
|     |      |      |      |      |     |
|    *|      |      |      |      |     |
|     |      |      |      |      |     |
| 80.5| 0.35 | 0.27 | 77   | 0.27 | 77  |
| 80.5| 0.35 | 0.27 | 77   | 0.27 | 77  |
|     |      |      |      |      |     |
|     |      |      |      |      |     |

- Finds subgroups of outcomes having a high likelihood for a joint normal distribution over the variables.

- Suitable if similar patterns of continuous values are wanted.

- Expandable on other distributions.
Function: biclust

The main function of the package is

biclust(data, method=BCxxx(), number, ...)

with:

data: The preprocessed data matrix
method: The algorithm used (E.g. BCCC() for CC)
number: The maximum number of bicluster to search for
... : Additional parameters of the algorithms

Returns an object of class Biclust for uniform treatment.
Additional methods

Preprocessing: discretize(), binarize(), ...

Visualization: parallelCoordinates(), drawHeatmap(), plotclust(), ...

Validation: jaccardind(), clusterVariance(), ...
III. The biclust - Package: Visualizations

Bicluster 2
(rows= 10 ; columns= 5 )

Cluster 1 Size: 9

Cluster 2 Size: 10

Cluster 3 Size: 10

Bicluster 2 (size 10 x 5 )
III. The biclust - Package: biclustmember()

biclustmember(Biclust, data, number, ...)

BiCluster Membership Graph

<table>
<thead>
<tr>
<th>Variable 15</th>
<th>Variable 14</th>
<th>Variable 13</th>
<th>Variable 12</th>
<th>Variable 11</th>
<th>Variable 10</th>
<th>Variable 9</th>
<th>Variable 8</th>
<th>Variable 7</th>
<th>Variable 6</th>
<th>Variable 5</th>
<th>Variable 4</th>
<th>Variable 3</th>
<th>Variable 2</th>
<th>Variable 1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CL. 1        CL. 2        CL. 3

Variable 15  Variable 15  Variable 15
Variable 14  Variable 14  Variable 14
Variable 13  Variable 13  Variable 13
Variable 12  Variable 12  Variable 12
Variable 11  Variable 11  Variable 11
Variable 10  Variable 10  Variable 10
Variable 9   Variable 9   Variable 9
Variable 8   Variable 8   Variable 8
Variable 7   Variable 7   Variable 7
Variable 6   Variable 6   Variable 6
Variable 5   Variable 5   Variable 5
Variable 4   Variable 4   Variable 4
Variable 3   Variable 3   Variable 3
Variable 2   Variable 2   Variable 2
Variable 1   Variable 1   Variable 1
III. The biclust - Package: biclustbarchart()

barchart(Biclust, data, number, ...)

Population mean:  -- Segmentwise means:  --- in bicluster
outside bicluster
IV. Example: Tourism Survey

Australian Tourism Survey

- Survey conducted by researchers from the Faculty of Commerce, University of Wollongong

- Data collected from a nationally representative online Internet panel

- Questions about travel and unpaid help behavior

- 1003 people, 56 blocks of question à about 5 to 51 questions (around 600 questions)
IV. Example: Tourism Survey I

**Activity questions:** Questions on activities participants did during their vacation.

```r
> bimaxres <- biclust(x = activity, method = BCBimax(), number = 50,
+                      mrow = 50, mcol = 4)
> bimaxres

An object of class Biclust

call:
biclust(x = activity, method = BCBimax(), number = 50, mrow = 50, mcol = 4)

Number of Clusters found: 11

First 5 Cluster sizes:
BC 1 BC 2 BC 3 BC 4 BC 5

Number of Rows: "74" "59" "55" "50" "75"
Number of Columns: "11" "10" "9" "8" "7"
```
IV. Example: Tourism Survey I

biclustmember(res=bimaxres,data=activity,number=1,...)

Result Biclustering on Activity Questions
IV. Example: Tourism Survey I

Motivation questions: Questions on motivations for unpaid help weighted with importance.

```r
> questres <- biclust(x = motivation, method = BCQuestord(), d = 2, ns = 500, + nd = 500, sd = 1, alpha = 0.05, number = 10)
> questres

An object of class Biclust

call:
  biclust(x = motivation, method = BCQuestord(), ns = 500, nd = 500, sd = 1, alpha = 0.05, number = 10)

Number of Clusters found: 10

First 5 Cluster sizes:
        BC 1  BC 2  BC 3  BC 4  BC 5
Number of Rows: "76" "69" "77" "59" "57"
Number of Columns: "12" "6" "4" "5" "3"
```
IV. Example: Tourism Survey II

biclustmember(res=questres, data=motivation, number=1,...)

Result Biclustering on Motivation Questions

<table>
<thead>
<tr>
<th></th>
<th>CL. 1</th>
<th>CL. 2</th>
<th>CL. 3</th>
<th>CL. 4</th>
<th>CL. 5</th>
<th>CL. 6</th>
<th>CL. 7</th>
<th>CL. 8</th>
<th>CL. 9</th>
<th>CL. 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>BxElearnskills</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>xEimproveenv</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3xEfeelvalued</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3xEhelpethnic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>xEperspective</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BxEassistorg</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BxEmindoff</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BxEhelpthose</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BxEsocialise</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BxEnjoy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BxEgiveback</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>:supportcause</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BxElearnskills</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BxEimproveenv</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BxEfeelvalued</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BxEhelpethnic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BxEperspective</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BxEassistorg</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BxEmindoff</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BxEhelpthose</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BxEsocialise</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BxEnjoy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BxEgiveback</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BxEsupportcause</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
barchart(res=questres, data=motivation, number=1,...)

Result Biclustering on Motivation Questions

Population mean: Segmentwise means: 
in bicluster outside bicluster
V. Summary and Future Work

Summary

- New bicluster algorithm to deal with nominal, ordinal and continuous data
- New developments in the biclust package
- Example on tourism data

Future Work

- Simultaneous clustering of nominal, ordinal and continuous data (Questionaire)
- Fully model based biclustering
Market segmentation is a joint work with Sara Dolnicar from the School of Management and Marketing of the University of Wollongong in Australia.

The package biclust is a joint work with Microarray Analysis and Visualization Effort, University of Salamanca, Spain, especially Rodrigo Santamaria.
biclust - A Toolbox for Bicluster Analysis in R,

BICLUSTERING: Overcoming data dimensionality problems in market segmentation,

Links:

http://cran.r-project.org/package=biclust/ official release

http://r-forge.r-project.org/projects/biclust/ newest developments

http://www.statistik.lmu.de/~kaiser/bicluster.html Papers and Links