biclust - A Toolbox for Bicluster Analysis in R

Sebastian Kaiser and Friedrich Leisch

Institut für Statistik
Ludwig-Maximilians-Universität München

Overview

Outline:

Introduce Biclustering

The biclust - Package

Examples

Future Work
Why Biclustering?

- Simultaneous clustering of 2 dimensions
- Large datasets where clustering leads to diffuse results
- Only parts of the data influence each other
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>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</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>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</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>
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.

\[
\begin{bmatrix}
A & * & * & A & * & A & * \\
* & * & * & * & * & * & * \\
* & * & * & * & * & * & * \\
A & * & * & A & * & A & * \\
* & * & * & * & * & * & * \\
A & * & * & A & * & A & * \\
* & * & * & * & * & * & * \\
\end{bmatrix}
\Rightarrow
\begin{bmatrix}
A & A & A & * & * & * & * \\
A & A & A & * & * & * & * \\
A & A & A & * & * & * & * \\
* & * & * & * & * & * & * \\
* & * & * & * & * & * & * \\
* & * & * & * & * & * & * \\
* & * & * & * & * & * & * \\
\end{bmatrix}
\]
More than one bicluster?

Most Bicluster Algorithms are iterativ. To find the next bicluster given n-1 found bicluster you have to either

- ignore the n-1 already found bicluster,

- delete rows and/or columns of the found bicluster or

- mask the found bicluster with random values.
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

**Xmotifs** *(Murali and Kasif (2003))*: Coherent correlation over rows and columns
The biclust - Package

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(), Bubbleplot()

Validation: jaccardind(), clusterVariance(), ...
Validation: Jaccard

Jaccard index for biclustering

- Compare two bicluster results \((\text{Biclust}_1, \text{Biclust}_2)\).

- Percentage of datapoints in the same cluster.

- Only datapoints who are clustered in at least one of the results.

\[ \text{JacInd}(\text{Biclust}_1, \text{Biclust}_2) = \frac{|BICP_1 \cap BICP_2|}{|BICP_1| + |BICP_2| - |BICP_1 \cap BICP_2|} \]

- where \(BICP\) are the point combinations in a cluster in result \(\text{Biclust}_I\)
Example

**BicatYeast**

- Subsample of the *Saccharomyces Cerevisiae* organism (Yeast)
- Used to present bicluster algorithms by Barkow et al. (2006)
- Microarray data: 419 genes, 80 experiments
Example: BicatYeast

```r
> data(BicatYeast)
> x<-discretize(BicatYeast)
> Xmotif<-biclust(x, method=BCXmotifs(), number=50, alpha=0.05,
+ nd=20, ns=20, sd=5)
> Xmotif

An object of class Biclust

call:
  biclust(x = x, method = BCXmotifs(), number = 50, alpha = 0.05)

Number of Clusters found: 15

First Cluster size:
  Number of Rows: 175
  Number of Columns: 6
```
Example: BicatYeast

> parallelCoordinates(x=BicatYeast, bicResult=Xmotif, number=6)
Example: BicatYeast

> parallelCoordinates(x=BicatYeast, bicResult=Xmotif, number=1, + geneTitle=TRUE)

Bicluster 1 (genes= 157 ; conditions= 6 )
Example: BicatYeast

<table>
<thead>
<tr>
<th></th>
<th>BCPlaid</th>
<th>BCXmotifs</th>
<th>BCCC</th>
<th>BCSpect.</th>
<th>BCBimax</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCPlaid</td>
<td>1.0000</td>
<td>0.0007</td>
<td>0.0116</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>BCXmotifs</td>
<td>0.0007</td>
<td>1.0000</td>
<td>0.1789</td>
<td>0.0935</td>
<td>0.0000</td>
</tr>
<tr>
<td>BCCC</td>
<td>0.0116</td>
<td>0.1789</td>
<td>1.0000</td>
<td>0.0898</td>
<td>0.0036</td>
</tr>
<tr>
<td>BCSpectral</td>
<td>0.0000</td>
<td>0.0935</td>
<td>0.0898</td>
<td>1.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>BCBimax</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0036</td>
<td>0.0000</td>
<td>1.0000</td>
</tr>
</tbody>
</table>
Example: BicatYeast (Jaccard)

Adapted Jaccard Index (Boxplot)

![Boxplot of Jaccard Index for different candidates](image)

- BCPlaid
- BCXmotifs
- BCCC
- BCSpectral
- BCBimax

The boxplot shows the distribution of 1 - Jaccard Index values for each candidate, with outliers indicated by individual points.
Example: BicatYeast (Jaccard)

Adapted Jaccard Index (Beplot I)

Podium
Example: BicatYeast (Jaccard)

Adapted Jaccard Index (Beplot II)
Australian Tourism Survey

- Survey of the Faculty of Commerce, University of Wollongong
- Questions on activities during the holidays
- 1003 people, 56 question blocks a about 15 questions
- Using a sample of 2 blocks (30 questions)
Example: Tourism Data

```r
> x<-AUSTourismSurvey
> Xmotif<-biclust(x, method=BCXmotifs(), number=10, alpha=0.05, + ns=50, nd=50, sd=5)
> Xmotif

An object of class Biclust
call:
biclust(x, method=BCXmotifs(), number=10, alpha=0.05)
Number of Clusters found: 7
First Cluster size:
  Number of Rows: 218
  Number of Columns: 20
```
Example: Tourism Data

```r
> parallelCoordinates( x=AUSTourismsurvey, bicResult=Xmotif, bicluster=1)
```
Example: Tourism Data

> drawHeatmap( x=AUSTourismsurvey, bicResult=Xmotif, bicluster=1)
Future Work

• Benchmark algorithms on more difficult scenarios.

• Compare algorithms in different application fields.

• Develop rules to choose bicluster algorithm.

• Develop a model based bicluster algorithm.

• Statistical tests on bicluster results.
Acknowledgments

The package biclust is a joint work with Microarray Analysis and Visualization Effort, University of Salamanca, Spain, especially Rodrigo Santamaria.

The benchplot is the work of Manuel Eugster, working group computational statistics, LMU Munich.

See

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

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

and