Decomposition of Boolean Multi-Relational Data with Graded Relations

Martin Trnecka, Marketa Trneckova

DEPARTMENT OF COMPUTER SCIENCE
PALACKÝ UNIVERSITY OLOMOUC
CZECH REPUBLIC

IEEE International Conference on Intelligent systems IS’16
Sofia, Bulgaria, September 4-6, 2016
Boolean Matrix Decomposition

- Method for analysis of Boolean data.
- A general aim: for a given matrix $I \in \{0, 1\}^{n \times m}$ find matrices $A \in \{0, 1\}^{n \times k}$ and $B \in \{0, 1\}^{k \times m}$ for which $I$ (approximately) equals $A \circ B$.
- $\circ$ is the Boolean matrix product

$$(A \circ B)_{ij} = \max_{l=1}^{k} \min(A_{il}, B_{lj}).$$

$$\begin{pmatrix}
10111 \\
01101 \\
01001 \\
10110
\end{pmatrix}
= \begin{pmatrix}
110 \\
011 \\
001 \\
100
\end{pmatrix} \circ \begin{pmatrix}
10110 \\
00101 \\
01001
\end{pmatrix}$$

- Discovery of $k$ factors that exactly or approximately explain the data.
- Factors = interesting patterns (rectangles) in data.
Limits of Boolean Matrix Decomposition

- Various methods and approaches.
- Classic setting: can handle only one input data matrix.
- Many real-word data sets are more complex than one simple data table.
- **Multi-Relational Data** = data composed from many tables (matrices) interconnected via relations between objects or attributes of these data tables.
Multi-Relation Boolean Matrix Factorization


- Problem settings: Two Boolean data tables $C_1$ and $C_2$ interconnected with binary relation $R_{12}$.

- Multi-Relational Factor = pair of classic factors satisfying relation (several ways).

- Algorithmic issue: how to select these factors.
Simple Example

Table: $C_1$

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>2</td>
<td>×</td>
<td></td>
<td>×</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>×</td>
<td></td>
<td>×</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
</tbody>
</table>

Table: $C_2$

<table>
<thead>
<tr>
<th></th>
<th>e</th>
<th>f</th>
<th>g</th>
<th>h</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>×</td>
<td></td>
<td>×</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>×</td>
<td>×</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>×</td>
<td>×</td>
<td></td>
</tr>
</tbody>
</table>

Table: $R_{C_1 C_2}$

<table>
<thead>
<tr>
<th></th>
<th>e</th>
<th>f</th>
<th>g</th>
<th>h</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>×</td>
<td>×</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>×</td>
<td></td>
<td>×</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>4</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
</tbody>
</table>

Factors of data table $C_1$ are: $F_{C_1}^1 = \langle \{1, 4\}, \{b, c, d\} \rangle$, $F_{C_1}^2 = \langle \{2, 4\}, \{a, c\} \rangle$, $F_{C_1}^3 = \langle \{1, 3, 4\}, \{b, d\} \rangle$ and factors of table $C_2$ are: $F_{C_2}^1 = \langle \{6, 7\}, \{f, g\} \rangle$, $F_{C_2}^2 = \langle \{5\}, \{e, h\} \rangle$, $F_{C_2}^3 = \langle \{5, 7\}, \{e\} \rangle$, $F_{C_2}^4 = \langle \{8\}, \{g, h\} \rangle$. 

<table>
<thead>
<tr>
<th></th>
<th>$F_{C_1}^C_1$</th>
<th>$F_{C_2}^C_1$</th>
<th>$F_{C_2}^C_2$</th>
<th>$F_{C_2}^C_3$</th>
<th>$F_{C_2}^C_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_{C_1}^1$</td>
<td>×</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$F_{C_1}^2$</td>
<td></td>
<td>×</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$F_{C_1}^3$</td>
<td></td>
<td>×</td>
<td>×</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Our Work

- The main advantage of Boolean data is interpretability.
- Considering Boolean data only can be limiting.
- Relation between input matrices is not necessarily of a Boolean nature.
- **Our goal:** Compute for two input Boolean matrices $C_1$ and $C_2$ and relation $R_{12}$ (with grades from some scale $L$) between them, multi-relational factors.
- Multi-relation factor on $C_1$ and $C_2$ is $\langle F_i^{C_1}, F_j^{C_2}, d \rangle$, where $F_i^{C_1} \in F_{C_1}$, $F_j^{C_2} \in F_{C_2}$ ($F_{C_1}$ and $F_{C_2}$ represent sets of classical factors from $C_1$ and $C_2$ respectively) and both are compatible with relation $R_{12}$ in degree $d \in L$.
- We want factors explaining (covering) the largest part of input data.
- We assume that $L$ conforms to the structure of a complete residuated lattice used in Fuzzy logic.
Solution

- Factors = Formal concepts (clear interpretation, geometrical viewpoint).
- We design new BMF algorithm (part of our final algorithm)
- Based on so called “Essential elements”
- Derivate of GreESS algorithm.
- We used calculus over Fuzzy logic and residuated lattices.
Idea of Algorithm (in case of object attribute relation)

- The main issue: how to understand that “factors $F^{C_1}_i \in \mathcal{F}_{C_1}$ and $F^{C_2}_j \in \mathcal{F}_{C_2}$ are compatible in a relation $R_{12}$ in degree $d$”.

- Intuitively: we want all objects from $F^{C_1}_i$ to be compatible with relation $R_{12}$ and also all attributes from $F^{C_2}_j$ to be compatible with this relation.

- “object $x$ is compatible with relation” means: if object $x$ is in $F^{C_1}_i$ then $x$ has all attributes from $F^{C_2}_j$ in relation $R_{12}$.

- Similarly for attributes.

- For two factors $\langle A, B \rangle$ and $\langle C, D \rangle$:

\[
d = \left( \bigwedge_{x \in A} \left( x \rightarrow \bigwedge_{y \in D} R_{12}(x, y) \right) \right) \otimes \left( \bigwedge_{y \in D} \left( y \rightarrow \bigwedge_{x \in A} R_{12}(x, y) \right) \right).
\]
Algorithm

**Input:** Boolean matrices $C_1$, $C_2$ and relation $R_{12}$.

**Output:** Set $F$ of multi-relational factors.

1: $F_{C_1} \leftarrow$ Boolean factors of $C_1$
2: $F_{C_2} \leftarrow$ Boolean factors of $C_2$
3: $U_{C_1} \leftarrow C_1$
4: $U_{C_2} \leftarrow C_2$
5: **foreach** $\langle A, B \rangle \in F_{C_1}$ **do**
6: compute set of all candidates $F_{\langle A,B \rangle} \subseteq F_{C_2}$ which
are compatible in $R_{12}$ with $\langle A, B \rangle$ in degree $d > 0$
7: **end for**
8: **while** exist $\langle A, B \rangle$ and $\langle C, D \rangle \in F_{\langle A,B \rangle}$ which can be connected and improve coverage **do**
9: **select** $\langle A, B \rangle$ and corresponding $\langle C, D \rangle \in F_{\langle A,B \rangle}$ that
cover the biggest parts of $U_{C_1}$ and $U_{C_2}$
10: **add** $\langle \langle A, B \rangle, \langle C, D \rangle, d \rangle$ **to** $F$
11: **remove** all entries in $\langle A, B \rangle$ **from** $U_{C_1}$
12: **remove** all entries in $\langle C, D \rangle$ **from** $U_{C_2}$
13: **remove** $\langle C, D \rangle$ **from** $F_{\langle A,B \rangle}$
14: **end while**
Experimental Evaluation on Synthetic Data

- Quality of factorization.
- The main factor: density of relational matrix.
- To eliminate influence of input matrices $C_1$ and $C_2$, we fixed them. $C_1$ has a size $1000 \times 500$ and approximate density of ones 25% and $C_2$ has a size $500 \times 1000$ and the same density.
- Relational matrix has a size $500 \times 500$. Grades of this matrix are from the scale

$$L = \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}.$$ 

- We wanted to demonstrate that the number of zeros in this relation plays a crucial role. We used 10 different sets of relational matrices with different distribution of grades.
- Each set contains 1000 of such relations.
### Results

**Table: Results for synthetic data**

<table>
<thead>
<tr>
<th>Set</th>
<th>average percent of zeros</th>
<th>average coverage of $C_1$</th>
<th>average coverage of $C_2$</th>
<th>average total coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set 1</td>
<td>89%</td>
<td>65%</td>
<td>58%</td>
<td>62%</td>
</tr>
<tr>
<td>Set 2</td>
<td>81%</td>
<td>75%</td>
<td>69%</td>
<td>72%</td>
</tr>
<tr>
<td>Set 3</td>
<td>72%</td>
<td>85%</td>
<td>79%</td>
<td>82%</td>
</tr>
<tr>
<td>Set 4</td>
<td>61%</td>
<td>93%</td>
<td>90%</td>
<td>91%</td>
</tr>
<tr>
<td>Set 5</td>
<td>52%</td>
<td>95%</td>
<td>93%</td>
<td>94%</td>
</tr>
<tr>
<td>Set 6</td>
<td>39%</td>
<td>99%</td>
<td>98%</td>
<td>98%</td>
</tr>
<tr>
<td>Set 7</td>
<td>28%</td>
<td>99.8%</td>
<td>99.6%</td>
<td>99.7%</td>
</tr>
<tr>
<td>Set 8</td>
<td>20%</td>
<td>99.9%</td>
<td>99.9%</td>
<td>99.9%</td>
</tr>
<tr>
<td>Set 9</td>
<td>15%</td>
<td>99.9%</td>
<td>100%</td>
<td>99.9%</td>
</tr>
<tr>
<td>Set 10</td>
<td>10%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>
Experimental Evaluation on Real Data

- MovieLens dataset.
- [http://grouplens.org/datasets/movielens/](http://grouplens.org/datasets/movielens/)
- Two data tables that represent a set of users and their attributes (e.g. gender, age, occupation) and a set of movies and their attributes (e.g. genre).
- Ratings are made on a 5-star scale (values 1-5, 1 means, that user does not like a movie and 5 means that he likes a movie).
- We used 10M version of MovieLens dataset
- We chose users that rate the most and films that are rated the most.
- Ratings were normalized to \([0, 1]\) interval.
- By our algorithm we obtained 46 multi-relational factors.
- These factors cover 98 percent of input data tables.
Cumulative Coverage

Figure: Cumulative coverage of User and Movie data tables
Interpretation of Obtained Factors

- College female students rated action, sci-fi and thriller movies from 1980s with at least three stars.
- Females students of elementary school rated new comedy films with at least three stars.
- College males students rated action, adventure and fantasy movies with at least four stars.
- Middle aged males rated new drama films at with at least three stars.
- Late forties females working as academics or educators rated films from 1970s with five stars.
- Females in the age of 25-34 rated children, animated and comedy movies with four stars.
Conclusion

- We extend a problem of multi-relational Boolean matrix decomposition toward a more general case.
- We proposed a new algorithm for this general case.
- Our new approach is tailored for multi-relational data that contains a relation with degrees from some scale.
- We used calculus over Fuzzy logic to solve a problem how to connect factors into multi-relational factors.
Thank you