Handling Noise in Boolean Matrix Factorization

Radim Belohlavek, Martin Trnecka
Department of Computer Science, Palacky University Olomouc, Czech Republic

Boolean Matrix Factorization

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

\[
(A \otimes B)_{ij} = \max_{l=1}^{k} \min_{m=1}^{n} (A_{im}, B_{lj})
\]

- discovery of \( k \) factors that exactly or approximately explain the data
- factors = interesting patterns (rectangles) in data

Current Understanding of Noise and Its Role in BMF

What is Noise in Boolean Data?

- noise in Boolean data → distortion of data, i.e. flipping some data entries of true data
- subtractive, additive noise, general noise

Is Noise Always a Reasonable Assumption?

- term “noise” → strange
- noise = random and mostly small fluctuations in data
- Boolean data → “complete change” term “error”
- many real datasets do not contain noise because they simply contain verified truth
- there exist applications of BMF, in which presence of noise would be counterintuitive or even damaging (e.g. role mining problem)
- which levels of noise are realistic (some works consider 40% noise)

A Rationale for Robustness to Noise

- algorithms do not committing “overcover error” are not able to discover these factors

A Critique of Current Approach

- current experiments ↦ robustness to noise
- wrong approach
- mix of three distinctive terms: coverage quality, robustness to noise and ground truth

Revisiting Datasets from Gupta at al.’s Paper

- moreover our observations are different from those reported in literature

New Way to Assess Robustness to Noise

Similarity of Factorizations

- factor = rectangular area in data, i.e. the Cartesian product \( C \times D \) for some \( C \subseteq \{1, \ldots, n\} \) and \( D \subseteq \{1, \ldots, m\} \)
- for two sets \( F \) and \( F' \) of factors we define

\[
sim(F, F') = \min \left( \frac{\sum_{c \in F} \sup_{s \in D} \{s(C) \cap s(C') \} }{|F|}, \frac{\sum_{c \in F} \sup_{s \in D} \{s(C) \cap s(C') \} }{|F'|} \right)
\]

where

\[
s(C) = \{s \in D \mid s \neq \emptyset \wedge s(C) \neq \emptyset \}
\]

- assess capability to discover ground truth

Experimental evaluation

- synthetic and real data

Robustness to Noise

Table 1: Robustness to noise (Domino dataset, general noise)

<table>
<thead>
<tr>
<th>Noise (%)</th>
<th>AISO</th>
<th>GreConD</th>
<th>PaNDa</th>
<th>Nassau</th>
<th>GreConD+</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.894</td>
<td>0.895</td>
<td>0.913</td>
<td>0.924</td>
<td>0.947</td>
</tr>
<tr>
<td>0.5</td>
<td>0.904</td>
<td>0.901</td>
<td>0.923</td>
<td>0.935</td>
<td>0.958</td>
</tr>
<tr>
<td>1</td>
<td>0.927</td>
<td>0.928</td>
<td>0.946</td>
<td>0.960</td>
<td>0.974</td>
</tr>
<tr>
<td>2</td>
<td>0.956</td>
<td>0.957</td>
<td>0.975</td>
<td>0.987</td>
<td>0.993</td>
</tr>
<tr>
<td>5</td>
<td>0.989</td>
<td>0.990</td>
<td>0.999</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Table 2: Recovery of ground truth on synthetic data 500 × 250 with \( k = 5 \)

<table>
<thead>
<tr>
<th>Noise Change (%)</th>
<th>AISO</th>
<th>GreConD</th>
<th>PaNDa</th>
<th>Nassau</th>
<th>GreConD+</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.682</td>
<td>0.681</td>
<td>0.681</td>
<td>0.680</td>
<td>0.685</td>
</tr>
<tr>
<td>0.5</td>
<td>0.686</td>
<td>0.685</td>
<td>0.685</td>
<td>0.684</td>
<td>0.689</td>
</tr>
<tr>
<td>1</td>
<td>0.700</td>
<td>0.699</td>
<td>0.700</td>
<td>0.700</td>
<td>0.700</td>
</tr>
<tr>
<td>2</td>
<td>0.720</td>
<td>0.719</td>
<td>0.720</td>
<td>0.720</td>
<td>0.720</td>
</tr>
<tr>
<td>5</td>
<td>0.760</td>
<td>0.760</td>
<td>0.760</td>
<td>0.760</td>
<td>0.760</td>
</tr>
</tbody>
</table>

Conclusion

- new methodological ground
- clear separation of algorithms

Supported by grant No. GA17-17899S of the Czech Science Foundation.