Evaluating Association Rules in Boolean Matrix Factorization

Jan Outrata, Martin Trnecka

DEPARTMENT OF COMPUTER SCIENCE
PALACKÝ UNIVERSITY OLOMOUC

Seminar on Information Science

Paper and Conference

- *Outrata J., Trnecka M.: Evaluating Association Rules in Boolean Matrix Factorization. Workshop on Computational Intelligence and Data Mining, WCIDM 2016, In Proceedings of the 16th ITAT conference, CEUR Workshop Proceedings Vol. 1649, pp. 147–154.*
- 4rd international workshop of Computational Intelligence and Data Mining.
- Tatranské Matliare, Slovakia.
- September 17–18, 2016.

Boolean Matrix Factorization (BMF)

- 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.

Geometry of BMF

- Geometry of factorization → coverage of the entries containing 1s by rectangles.

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

$$\begin{pmatrix} 10111 \\ 01101 \\ 01001 \\ 10110 \end{pmatrix} = \begin{pmatrix} \mathbf{1}\,0\,\mathbf{1}\,\mathbf{1}\,0 \\ 0\,0\,0\,0\,0 \\ 0\,0\,0\,0\,0 \\ \mathbf{1}\,0\,\mathbf{1}\,\mathbf{1}\,0 \end{pmatrix} \vee \begin{pmatrix} 0\,0\,0\,0\,0 \\ 0\,0\,\mathbf{1}\,0\,\mathbf{1} \\ 0\,0\,0\,0\,0 \\ 0\,0\,0\,0\,0 \end{pmatrix} \vee \begin{pmatrix} 0\,0\,0\,0\,0 \\ 0\,\mathbf{1}\,0\,0\,\mathbf{1} \\ 0\,\mathbf{1}\,0\,0\,\mathbf{1} \\ 0\,0\,0\,0\,0 \end{pmatrix}$$

- *Belohlavek R., Vychodil V., Discovery of optimal factors in binary data via a novel method of matrix decomposition, Journal of Computer and System Science 76(1)(2010), 3–20.*

Explanation of Data by Factors

- How large portion of data is explain by factors?
- Distance (error function)

$$E(C, D) = \|C - D\| = \sum_{i,j=1}^{m,n} |C_{ij} - D_{ij}|.$$

- Two components of E

$$\begin{aligned} E(I, A \circ B) &= E_u(I, A \circ B) + E_o(I, A \circ B), \text{ where} \\ E_u(I, A \circ B) &= |\{\langle i, j\rangle \,;\, I_{ij} = 1, (A \circ B)_{ij} = 0\}|, \\ E_o(I, A \circ B) &= |\{\langle i, j\rangle \,;\, I_{ij} = 0, (A \circ B)_{ij} = 1\}|. \end{aligned}$$

- Coverage quality for $A \in \{0,1\}^{n \times l}$ and $B \in \{0,1\}^{l \times m}$

$$c(l) = 1 - E(I, A \circ B)/\|I\|.$$

Two Basic Viewpoint to BMF

- Discrete Basis Problem
  - Given $I \in \{0,1\}^{n \times m}$ and a positive integer k, find $A \in \{0,1\}^{n \times k}$ and $B \in \{0,1\}^{k \times m}$ that minimize $\|I - A \circ B\|$.
  - Emphasizes the importance of the first few (presumably most important) factors.
  - *Miettinen P., Mielikainen T., Gionis A., Das G., Mannila H., The discrete basis problem, IEEE Transactional Knowledge and Data Engineering 20(10)(2008), 1348–1362*
- Approximate Factorization Problem
  - Given I and prescribed error $\varepsilon \geq 0$, find $A \in \{0,1\}^{n \times k}$ and $B \in \{0,1\}^{k \times m}$ with k as small as possible such that $\|I - A \circ B\| \leq \varepsilon$.
  - Emphasizes the need to account for (and thus to explain) a prescribed (presumably reasonably large) portion of data.
  - *Belohlavek R., Trnecka M., From-below approximations in Boolean matrix factorization: Geometry and new algorithm, Journal of Computer and System Science 81(8)(2015), 1678–1697.*

Our Work

- Association rules form a ground of the ASSO algorithm.
- *Miettinen P., Mielikainen T., Gionis A., Das G., Mannila H., The discrete basis problem, IEEE Transactional Knowledge and Data Engineering 20(10)(2008), 1348–1362*
- Confidence parameter influences the quality of factorization.
- Can other type of association rules improve ASSO?
- Can be used association rules in other BMF algorithms?
- GRECOND algorithm.
- *Belohlavek R., Vychodil V., Discovery of optimal factors in binary data via a novel method of matrix decomposition, Journal of Computer and System Science 76(1)(2010), 3–20.*

Association Rules in GUHA

- GUHA (General Unary Hypothesis Automaton)
- For Boolean data association rule (over a given set of attributes) is an expression

$$i \approx j$$

  where i and j are attributes.
- GUHA general association rule is an expression $\varphi \approx \psi$ where φ and ψ are arbitrary complex logical formulas above the attributes.
- Four-fold table 4ft(i, j, I)

$$\langle a, b, c, d \rangle = \langle fr(i \wedge j), fr(i \wedge \neg j), fr(\neg i \wedge j), fr(\neg i \wedge \neg j) \rangle$$

I	j	$\neg j$
i	$a = fr(i \wedge j)$	$b = fr(i \wedge \neg j)$
$\neg i$	$c = fr(\neg i \wedge j)$	$d = fr(\neg i \wedge \neg j)$.

(Generalized) Quantifiers

- Function q which assigns to any four-fold table 4ft(i, j, I) a logical value 0 or 1 defines a so-called (generalized, GUHA) quantifier.
- Logical and statistical viewpoints
- Interpret different types of association rules (with different meaning of the association $\approx$ between attributes)

(Generalized) Quantifiers

- *founded (p-)implication*, $\Rightarrow_p$ (for $\approx$)

$$q(a,b,c,d) = \begin{cases} 1 \text{ if } \frac{a}{a+b} \geq p, \\ 0 \text{ otherwise.} \end{cases}$$

- Used in ASSO.
- *double founded implication*, $\Leftrightarrow_p$

$$q(a,b,c,d) = \begin{cases} 1 \text{ if } \frac{a}{a+b+c} \geq p, \\ 0 \text{ otherwise.} \end{cases}$$

- Meaning: the number of objects having in I both i and j is at least $100 \cdot p\%$ of the number of objects having i or j.

(Generalized) Quantifiers

- *founded equivalence*, $\equiv_p$

$$q(a,b,c,d) = \begin{cases} 1 \text{ if } \frac{a+d}{a+b+c+d} \geq p, \\ 0 \text{ otherwise.} \end{cases}$$

- Meaning: At least $100 \cdot p\%$ among all objects in I have the same attributes.
- *E-equivalence*, $\sim_\delta^E$

$$q(a,b,c,d) = \begin{cases} 1 \text{ if } \max\left(\frac{b}{a+b}, \frac{c}{c+d}\right) < \delta, \\ 0 \text{ otherwise.} \end{cases}$$

- *negative Jaccard distance*

$$q(a,b,c,d) = \begin{cases} 1 \text{ if } \frac{b+c}{b+c+d} \geq p, \\ 0 \text{ otherwise.} \end{cases}$$

- Our new quantifier resembling Jaccard distance dissimilarity measure used in data mining.
- Meaning: at least $100 \cdot p\%$ objects have i or j among the objects not having i or j.

Modified Asso algorithm

Input: A Boolean matrix $I \in \{0,1\}^{n \times m}$, a positive integer k, a threshold value $\tau \in (0,1]$, real-valued weights w^+, w^- and a quantifier q_τ (with parameter τ) interpreting $i \approx j$
Output: Boolean matrices $A \in \{0,1\}^{n \times k}$ and $B \in \{0,1\}^{k \times m}$

for $i = 1, \ldots, m$ **do**
  for $j = 1, \ldots, m$ **do**
    $Q_{ij} = q_\tau(a, b, c, d)$
  end
end
$A \leftarrow$ empty $n \times k$ Boolean matrix
$B \leftarrow$ empty $k \times m$ Boolean matrix
for $l = 1, \ldots, k$ **do**
  $(Q_{i\_}, e) \leftarrow \arg\max_{Q_{i\_},\ e \in \{0,1\}^{n \times 1}} cover(\begin{bmatrix} B \\ Q_{i\_} \end{bmatrix}, [A\ e], I, w^+, w^-)$
  $A \leftarrow [A\ e]$, $B \leftarrow \begin{bmatrix} B \\ Q_{i\_} \end{bmatrix}$
end
return A and B

Modified GreConD algorithm

Input: A Boolean matrix $I \in \{0,1\}^{n \times m}$ and a prescribed error $\varepsilon \geq 0$
Output: Boolean matrices $A \in \{0,1\}^{n \times k}$ and $B \in \{0,1\}^{k \times m}$

$Q \leftarrow$ empty $m \times m$ Boolean matrix
for $i = 1, \ldots, m$ **do**
  for $j = 1, \ldots, m$ **do**
    if $i \Rightarrow_1 j$ is *true* in I **then**
      $Q_{ij} = 1$
    end
  end
end
$A \leftarrow$ empty $n \times k$ Boolean matrix
$B \leftarrow$ empty $k \times m$ Boolean matrix
while $\|I - A \circ B\| > \varepsilon$ **do**
  $D \leftarrow \arg\max_{Q_{i\_}} cover(Q_{i\_}, I, A, B)$
  $V \leftarrow cover(D, I, A, B)$
  while there is j such that $D_j = 0$ and $cover(D + [j], I, A, B) > V$ **do**
    $j \leftarrow \arg\max_{j, D_j = 0} cover(D + [j], I, A, B)$
    $D \leftarrow (D + [j])^{\downarrow\uparrow}$
    $V \leftarrow cover(D, I, A, B)$
  end
  $A \leftarrow [A \; D^{\downarrow}]$, $B \leftarrow \begin{bmatrix} B \\ D \end{bmatrix}$
end

Experimental Evaluation

- Synthetic data
  1000 of randonly generated datasets (500 rows and 250 columns).

Dataset	k	dens A	dens B	dens I
Set C1	40	0.07	0.04	0.10
Set C2	40	0.07	0.06	0.15
Set C3	40	0.11	0.05	0.20

Table: Synthetic data

- Real data

Dataset	Size	$\|\|I\|\|$
DNA	4590×392	26527
Mushroom	8124×119	186852
Zoo	101×28	862

Table: Real data

Results C1

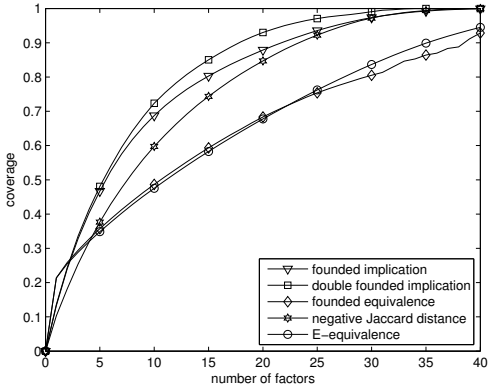

Figure: Coverage for synthetic dataset C_1

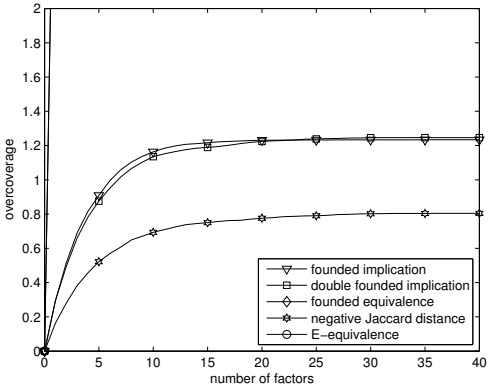

Figure: Overcoverage for synthetic dataset C_1

Results C2

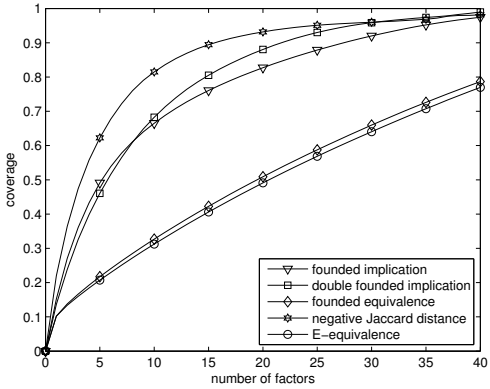

Figure: Coverage for synthetic dataset C_2

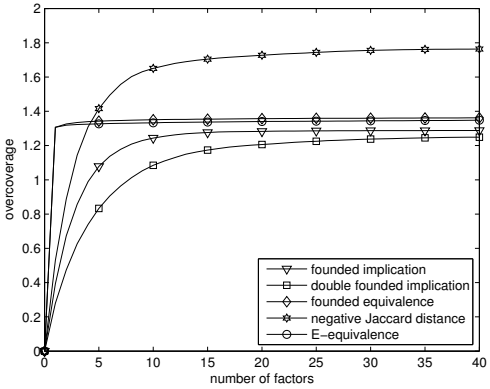

Figure: Overcoverage for synthetic dataset C_2

Results Mushroom

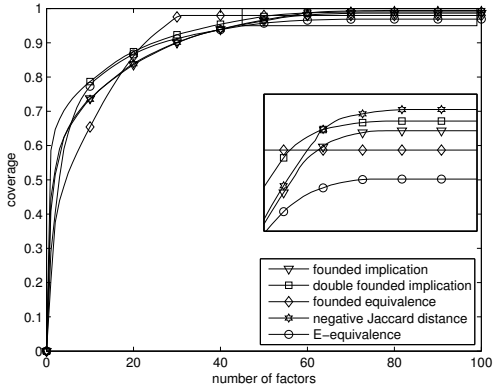

Figure: Coverage for Mushroom dataset

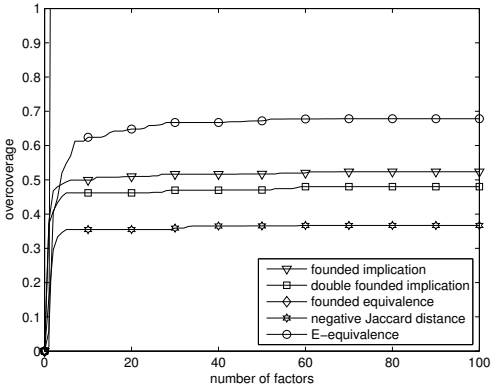

Figure: Overcoverage for Mushroom dataset

Results GreConD

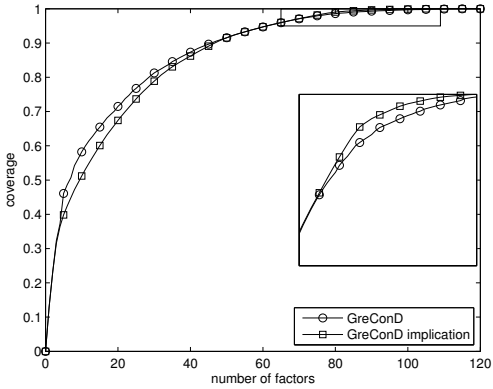

Figure: Original and modified GreConD on Mushroom dataset

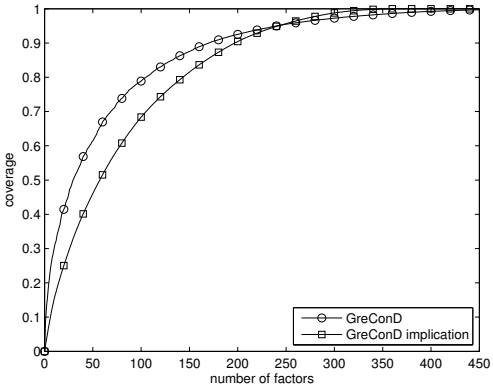

Figure: Original and modified GreConD on DNA dataset

General Remarks

- Time complexity.
- Modification of GreConD is slightly faster than original.
- Modification of Asso is equally fast as the original.
- Time (and space) complexity is not critical issue (for the most of current algorithms)
- Implementation in MATLAB.
- Runable on ordinar PC.

Conclusions

- We evaluated the use of various types of (general) association rules from the GUHA knowledge discovery method in the Boolean matrix factorization (BMF).
- We modify ASSO and GRECOND (not based on association rules).
- Our modified algorithms outperform, for some types of rules, the original ones.
- The most promissing results: founded implication and (our new) negative Jaccard distance quantifiers.