Boolean Matrix Decomposition by Formal Concept Sampling

Petr Osicka, Martin Trnecka

Department of Computer Science, Palacky University Olomouc, Czech Republic

Boolean Matrix Decomposition

- method for analysis of Boolean data
- 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

Formal Concept Analysis

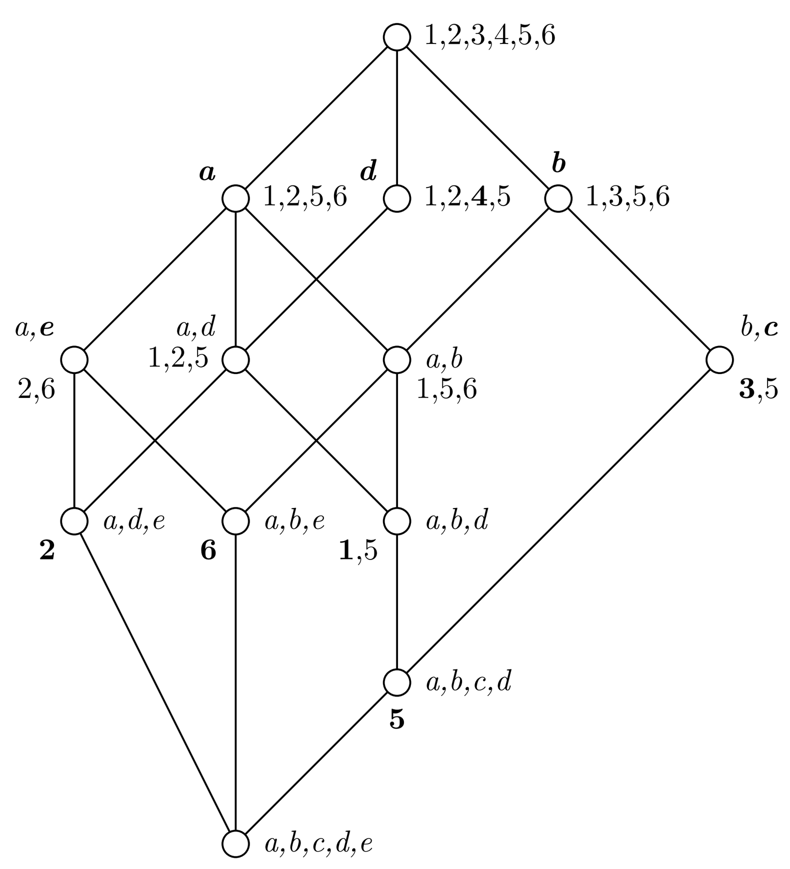
- $n \times m$ matrix $I \to$ formal context $(\{1,\dots,n\},\{1,\dots,m\},J)$ with $(x,y) \in J$ iff $I_{xy} = 1$
- formal context induces a pair of operators ${}^{\uparrow} : 2^X \to 2^Y$ and ${}^{\downarrow} : 2^Y \to 2^X$ defined by

$$D^{\uparrow} = \{y \in \{1,\dots,m\} \mid \text{for all } x \in D : (x,y) \in J\},$$
$$E^{\downarrow} = \{x \in \{1,\dots,n\} \mid \text{for all } y \in E : (x,y) \in J\}$$

- formal concept $\langle D,E\rangle$, where $D \subseteq \{1,\dots,n\}$, $E \subseteq \{1,\dots,m\}$, $D^{\uparrow} = E$, $E^{\downarrow} = D$
- set of all formal concepts $\mathcal{B}(I)$ can be ordered by $\preceq$ defined by $\langle D,E\rangle \preceq \langle D',E'\rangle$ iff $D \subseteq D'$ (or equivalently $E' \subseteq E$), moreover forms complete lattice

Formal Concepts as Factors

$$\begin{array}{cc} & \begin{array}{ccccc} a & b & c & d & e \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} & \begin{pmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \end{pmatrix} \end{array}$$



GreCon Algorithm

- the simplest algorithm for BMD
- Belohlavek, R., Vychodil, V.: Discovery of optimal factors in binary data via a novel method of matrix decomposition. *J. Comput. Syst. Sci.* 76, 1 (2010), 3–20.
- reduction BMD to Set-Cover problem.

```
procedure GreCon(I)
    U ← I
    F ← ∅
    Compute the set B(I) of all formal concepts of I
    while there is a 1 in U do
        find ⟨D, E⟩ ∈ B(I) that covers the most 1s in U
        add ⟨D, E⟩ to F
        for (i, j) ∈ D × E do
            U_ij ← 0
    t ← 1
    for ⟨D, E⟩ ∈ F do
        set A_t to the characteristic vector of D
        set B_t to the characteristic vector of E
        t ← t + 1
    return A, B
```

Formal Concept Sampling

- Metropolis-Hastings algorithm
- Boley, M., Gartner, T., Grosskreutz, H.: Formal Concept Sampling for Counting and Threshold-Free Local Pattern Mining. In Proceedings of the SIAM International Conference on Data Mining, SDM 2010, 1770–188, 2010.
- algorithm moves from one formal concept to another formal concept in $\mathcal{B}(I)$
- f assigning to a formal concept a nonzero value
- formal concept $\langle D,E\rangle$ is drawn with probability $\frac{f(\langle D,E\rangle)}{\sum f(\langle D',E'\rangle)}$
- the logarithmic number iterations is sufficient

New Algorithm

- modification of GreCon
- f takes into account the number covered 1s

Experimental evaluation

- sets of 1000×500 matrices with various properties
- real-world data
- GreCon is outperformed by the probabilistic algorithm
- increase in the number trials increases the coverage quality
- sparser data usually require less trials

Synthetic Data (Selected Results)

	5	10	15	20	25	30	35
GreCon	0.299	0.511	0.686	0.822	0.923	1.000	1.000
25 trials	0.283	0.499	0.675	0.813	0.921	0.990	1.000
50 trials	0.293	0.510	0.685	0.824	0.932	0.999	1.000
75 trials	0.296	0.510	0.686	0.825	0.932	0.999	1.000
100 trials	0.299	0.511	0.687	0.826	0.933	1.000	1.000

Table 1: Coverage quality of the first l rectangles on Set A

	5	10	15	20	25	30	35
GreCon	0.262	0.462	0.620	0.755	0.877	1.000	1.000
25 trials	0.198	0.363	0.509	0.646	0.751	0.827	0.896
50 trials	0.240	0.434	0.598	0.729	0.837	0.921	0.980
75 trials	0.253	0.449	0.614	0.753	0.872	0.954	0.996
100 trials	0.254	0.460	0.630	0.771	0.889	0.977	1.000

Table 3: Coverage quality of the first l rectangles on Set C

Real Data (Selected Results)

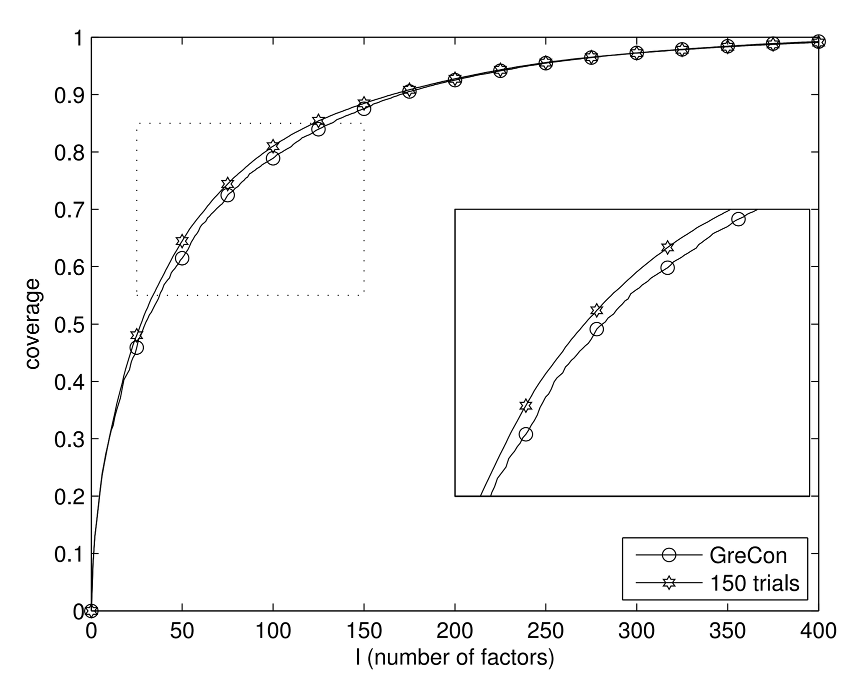


Figure 1: Coverage quality of the first l factors on DNA dataset

Conclusion

- new (probabilistic) BMD algorithm
- new directions in BMD algorithm design

Supported by grant No. GA15-17899S of the Czech Science Foundation and by ECOP project no. CZ.1.07/2.3.00/20.0059.