ARTICLE IN PRESS

Discrete Applied Mathematics ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Discrete Applied Mathematics

journal homepage: www.elsevier.com/locate/dam

A new algorithm for Boolean matrix factorization which admits overcovering

Radim Belohlavek, Martin Trnecka *

Department Computer Science, Palacký University, Olomouc, Czech Republic

ARTICLE INFO

Article history:
Received 13 February 2016
Received in revised form 13 July 2017
Accepted 18 December 2017
Available online xxxx

Keywords:
Boolean matrix factorization
Formal concept analysis
Algorithms

ABSTRACT

We present a new algorithm for general Boolean matrix factorization. The algorithm is based on two key ideas. First, it utilizes formal concepts of the factorized matrix as crucial components of constructed factors. Second, it performs steps back during the construction of factors to see if some of the already constructed factors may be improved or even eliminated in view of the subsequently added factors. The second idea is inspired by 8M—an old, previously incompletely described and virtually unknown factorization algorithm, which we analyze and describe in detail. We provide experimental evaluation of the new algorithm and compare it to 8M and two other well-known algorithms. The results demonstrate that our algorithm outperforms these algorithms in terms of quality of the decompositions as well in its robustness with respect to small changes in data.

1. Problem description

1.1. Problem in brief

Research in Boolean matrix factorization (BMF), or Boolean matrix decomposition, has resulted in various new methods of analysis and processing of data and has also contributed to our understanding of Boolean (binary, yes/no) data as regards foundational aspects. While the developments of practical methods and theoretical foundations are clearly connected, most of the current BMF methods use limited theoretical insight. Building upon our previous research [4], we developed in our recent paper [3] an efficient BMF algorithm utilizing a better understanding of the geometry of BMF, which we also developed in [3]. The understanding, provided in terms of Galois connections, concept lattices, and other structures underlying formal concept analysis (FCA [11]), as well as the algorithm, are primarily developed for exact Boolean matrix factorizations and employ formal concepts as factors. As such, the constructed factorizations are limited (in that they never commit overcovering, see below). Such limitation presents no restriction when exact factorizations are desired. Moreover, even though computing restricted type of decompositions, the algorithm outperforms the other BMF algorithms also when approximate factorizations are needed with a prescribed precision [3]. Nevertheless, there are situations in which general factorizations are desirable, hence the limitation described above may indeed prove restrictive. In the present paper, we extend our previous approach to BMF and develop a new algorithm that computes general factorizations.

1.2. Basic notions and rationale for computing general BMFs

We denote by I an $n \times m$ Boolean matrix and interpret primarily as an object–attribute incidence matrix (hence the symbol I). That is, the entry I_{ij} corresponding to the row i and the column j is either 1 or 0, indicating that the object i does

* Corresponding author.
E-mail addresses: radim.belohlavek@acm.org (R. Belohlavek), martin.trnecka@gmail.com (M. Trnecka).

https://doi.org/10.1016/j.dam.2017.12.044
0166-218X/

or does not have the attribute j, respectively. The set of all $n \times m$ Boolean matrices is denoted $\{0, 1\}^{n\times m}$. The ith row and jth column vector of I is denoted by $I_{i\_}$ and $I_{\_j}$, respectively. In BMF, one generally attempts to find for a given $I \in \{0, 1\}^{n\times m}$ matrices $A \in \{0, 1\}^{n\times k}$ and $B \in \{0, 1\}^{k\times m}$ for which

$$I \text{ (approximately) equals } A \circ B, \tag{1}$$

where $\circ$ is the Boolean matrix product, i.e. $(A \circ B)_{ij} = \max_{l=1}^{k} \min(A_{il}, B_{lj})$. A decomposition of I into $A \circ B$ may be interpreted as a discovery of k factors that exactly or approximately explain the data: Interpreting I, A, and B as object–attribute, object–factor, and factor–attribute matrices, model (1) reads: The object i has the attribute j if and only if there exists factor l such that l applies to i and j is one of the particular manifestations of l. The least k for which an exact decomposition $I = A \circ B$ exists is called the *Boolean rank* (or Schein rank) of I.

The approximate equality in (1) is commonly assessed in BMF by means of the L_1-norm (Hamming weight in case of Boolean matrices) $\|\cdot\|$ and the corresponding metric $E(\cdot, \cdot)$, defined for $C, D \in \{0, 1\}^{n\times m}$ by

$$\|C\| = \sum_{i,j=1}^{m,n} |C_{ij}| \quad \text{and} \quad E(C, D) = \|C - D\| = \sum_{i,j=1}^{m,n} |C_{ij} - D_{ij}|. \tag{2}$$

The following particular variants of the BMF problem, relevant to this paper, are considered in the literature.

- *Discrete Basis Problem* (DBP, [21]):
  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 $\|I - A \circ B\|$.
- *Approximate Factorization Problem* (AFP, [4]):
  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 $\|I - A \circ B\| \leq \varepsilon$.

These problems reflect two important views of BMF: DBP emphasizes the importance of the first few (presumably most important) factors; AFP emphasizes the need to account for (and thus to explain) a prescribed portion of data.

A useful view of BMF is provided in terms of rectangles [4,3]: $J \in \{0, 1\}^{n\times m}$ is called *rectangular* (a rectangle, for short) if $J = C \circ D$ for some $C \in \{0, 1\}^{n\times 1}$ (column) and $D \in \{0, 1\}^{1\times m}$ (row); this implies that upon suitable permutations of columns and rows, the 1s in J form a rectangular area. We say that J (or, the pair $\langle C, D\rangle$ for which $J = C \circ D$) *covers* $\langle i, j\rangle$ if $J_{ij} = 1$ (equivalently, $C_i = 1$ and $D_j = 1$). For matrices J_1 and J_2, we put

$$J_1 \leq J_2 \ (J_1 \text{ is contained in } J_2) \text{ iff } \quad (J_1)_{ij} \leq (J_2)_{ij} \text{ for every } i, j. \tag{3}$$

The following observation shows that a Boolean matrix product may be considered as a $\vee$-superposition of (or a coverage by) rectangles (see e.g. [3]):

Observation 1. *The following conditions are equivalent for any $I \in \{0, 1\}^{n\times m}$.*

(a) *$I = A \circ B$ for some $A \in \{0, 1\}^{n\times k}$ and $B \in \{0, 1\}^{k\times m}$.*
(b) *There exist rectangles $J_1, \ldots, J_k \in \{0, 1\}^{n\times m}$ such that $I = J_1 \vee \cdots \vee J_k$, i.e. $I_{ij} = \max_{l=1}^{k} (J_l)_{ij}$.*
(c) *There exist rectangles $J_1, \ldots, J_k \in \{0, 1\}^{n\times m}$ such that $I_{ij} = 1$ if and only if $\langle i, j\rangle$ is covered by some J_l.*

In particular, if A and B are the matrices from Observation 1(a), then one may put $J_l = A_{\_l} \circ B_{l\_}$ ($l = 1, \ldots, k$), i.e. J_l is the product of the lth column of A and the lth row of B, to obtain the rectangles in (b) and (c). Conversely, if $J_1 = C_1 \circ D_1, \ldots, J_k = C_k \circ D_k$ are the rectangles in (b) or (c) then the matrices A and B in which the lth column and lth row are C_l and D_l, respectively, satisfy (a). As a result, computing an exact factorization of I with a small number k of factors is equivalent to computing k rectangles contained in I that cover all the 1s in I. Since maximal rectangles in I correspond to formal concepts of I [11], the above observation led to the employment of formal concepts as factors in [4,3]. Clearly, one may utilize rectangles in I to cover not necessarily all 1s in I and thus to solve AFP. Even though such approach to AFP, as demonstrated by the algorithms in [4,3], is considerably successful, the resulting approximate factorizations $I \approx A \circ B$, which are called *from-below approximations* of I in [3], are restricted: While it may happen that $I_{ij} = 1$ and $(A \circ B)_{ij} = 0$ (undercovering), it never happens that $I_{ij} = 0$ and $(A \circ B)_{ij} = 1$ (overcovering).

That the lack of possible overcovering may be severely limiting is apparent from the following examples. Consider first the matrices I in Fig. 1a and J in Fig. 1b. Let I represent the observed data. One clearly recognizes three rectangles in I, the union of which forms the gray area, even though some of the entries inside the area contain 0 rather than 1. A natural view of I is that it results from the true data, represented by J, due to error. For instance, there might be insufficient evidence for the presence of some attributes on some objects, i.e. for the presence of 1s in certain entries, which is a plausible explanation of this kind of situation. From this viewpoint, one is interested in discovering from the observed data I the three factors behind the true data J, i.e. in view of the above observation in discovering from I the 10×3 and 3×10 matrices A and B for which $A \circ B = J$. But even if I represented true data, one may be interested in the decomposition into the above A and B because it

Please cite this article in press as: R. Belohlavek, M. Trnecka, A new algorithm for Boolean matrix factorization which admits overcovering, Discrete Applied Mathematics (2018), https://doi.org/10.1016/j.dam.2017.12.044.