Information Sciences 459 (2018) 71–85

Contents lists available at ScienceDirect

Information Sciences

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

Toward quality assessment of Boolean matrix factorizations

Radim Belohlavek, Jan Outrata, Martin Trnecka*

Department of Computer Science, Palacký University Olomouc, Czech Republic

ARTICLE INFO

Article history:
Received 4 December 2017
Revised 26 March 2018
Accepted 2 May 2018
Available online 9 May 2018

Keywords:
Boolean matrix
Factorization
Algorithm
Quality assessment

ABSTRACT

Boolean matrix factorization has become an important direction in data analysis. In this paper, we examine the question of how to assess the quality of Boolean matrix factorization algorithms. We critically examine the current approaches, and argue that little attention has been paid to this problem so far and that a systematic approach to it is missing. We regard quality assessment of factorization algorithms as a multifaceted problem, identify major views with respect to which quality needs to be assessed, and present various observations on the available algorithms in this regard. Due to its primary importance, we concentrate on the quality of collections of factors computed from data, present a method to assess this quality, and evaluate this method by experiments.

1. Introduction

1.1. Problem setting

During the last decade or so, Boolean matrix factorization (BMF, also Boolean matrix decomposition) became a thoroughly explored direction in data analysis. Most of the existing research contributions focused on development of new approaches and algorithms. Performance of the developed algorithms has been observed, but its evaluation has remained on intuitive grounds or has been conducted in a simple way using the matrix distance function (we review the current approaches below). No systematic treatment of quality assessment of BMF algorithms has been pursued so far and this important topic thus remains underdeveloped. The primary aim of the present paper is to examine quality assessment of BMF algorithms in detail.

Throughout this paper we use the following notation. The set of all $n \times m$ Boolean matrices shall be denoted by $\{0, 1\}^{n \times m}$ and the particular matrices by I, J, etc. These matrices are primarily interpreted as describing a relationship between n objects and m attributes. In particular, the entry I_{ij} corresponding to ith row and jth column indicates whether the object i has (in which case $I_{ij} = 1$) or does not have ($I_{ij} = 0$) the attribute j. The ith row and jth column of I are denoted by $I_{i\_}$ and $I_{\_j}$, respectively.

The basic problem in BMF, several variants of which are considered in the literature (see below), consists in finding for a given Boolean matrix $I \in \{0, 1\}^{n \times m}$ matrices $A \in \{0, 1\}^{n \times k}$ and $B \in \{0, 1\}^{k \times m}$ for which k is small and

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

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

https://doi.org/10.1016/j.ins.2018.05.016
0020-0255/

In such a decomposition, k represents the number of factors (i.e. discovered hidden variables, hence the term "factorization") and $\circ$ represents the Boolean matrix product, which is defined by the formula

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

The thus introduced factor model provides us with a natural interpretation. A decomposition of a Boolean matrix I into $A \circ B$ corresponds to discovery of k factors which explain, exactly or approximately, the data represented by I. In particular, the model given by (1) is described as follows: the matrices A (the object-factor matrix) and B (the factor-attribute matrix) explain the object-attribute matrix I as follows: the object i has the attribute j iff among the factors there is a factor l such that l applies to i and j is one of the particular manifestations of l.

The approximate equality $\approx$ in (1) is measured using the well-known matrix norm (L_1-norm) $||\cdot||$ defined by

$$||C|| = \sum_{i,j=1}^{m,n} |C_{ij}|,$$

and the corresponding metric E, which is given for Boolean matrices $C, D \in \{0, 1\}^{n \times m}$ by

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

1.2. Illustrative example

Consider the following 5×5 Boolean matrix I:

$$\begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

One may verify that I may be (exactly) factorized by the 5×4 and 4×5 matrices A and B, i.e. $I = A \circ B$, as follows:

$$\begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix} \circ \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \end{pmatrix}$$

Since the inner dimension in the matrix product is 4, the decomposition involves 4 factors. These factors correspond to 4 rectangles in that rectangle $l \in \{1, 2, 3, 4\}$ is represented by the product $A_{\_l} \circ B_{l\_}$ of the lth column of A and the lth row of B. The 4 rectangles involved in our factorization are:

$$A_{\_1} \circ B_{1\_} = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, A_{\_2} \circ B_{2\_} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix},$$

$$A_{\_3} \circ B_{3\_} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, A_{\_4} \circ B_{4\_} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

and I equals the $\vee$-superposition (i.e. max-superposition) of these rectangles, i.e. $I = A_{\_1} \circ B_{1\_} \vee A_{\_2} \circ B_{2\_} \vee A_{\_3} \circ B_{3\_} \vee A_{\_4} \circ B_{4\_}$.

If errors are acceptable, one may ask whether there exists an approximate factorization $I \approx A \circ B$ for which the error $E(I, A \circ B)$ is not too large. The following is such an approximate factorization involving two factors:

$$\begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix} \approx \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 0 & 1 \end{pmatrix} \circ \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$