International Journal of Approximate Reasoning 96 (2018) 78–94

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Handling noise in Boolean matrix factorization ☆

Radim Belohlavek, Martin Trnecka *

Dept. of Computer Science, Palacký University Olomouc, Czech Republic

ARTICLE INFO

Article history:
Received 12 August 2017
Received in revised form 7 December 2017
Accepted 13 March 2018
Available online 16 March 2018

Keywords:
Boolean matrix factorization
Algorithms
Noise in Boolean data

ABSTRACT

One of the challenges presented by Boolean matrix factorization consists in what became known as the ability to deal with noise in data. In this paper, we critically examine existing considerations regarding noise, reported results regarding various algorithms' ability to deal with noise, and approaches used to evaluate this ability. We argue that the current understanding is underdeveloped in several respects and, in particular, that the present way to assess the ability to handle noise in data is deficient. We provide a new, quantitative way to assess this ability. Our method is based on a common-sense definition of robustness requiring that the factorizations computed from data should not be affected much by varying the noise in the data. We present an experimental evaluation of several algorithms, and compare the results to the observations available in the literature. The experiments reveal important properties of these algorithms as regards handling noise. In addition to providing methodological justification of some properties claimed in the literature without proper justification, they reveal properties which were not reported as well as properties which counter certain claims made in the literature. Importantly, our approach reveals a line separating robust-to-noise from sensitive-to-noise algorithms, which has not been revealed by the previous approaches. We conclude by outlining open problems to which the present considerations and experiments lead.

1. Introduction

1.1. Problem in brief

Boolean matrix factorization (BMF) represents a vigorous research topic in data mining and related fields. One of the challenges faced in BMF, which recently attracted a number of researchers, is the ability to handle noise in Boolean data. Basically, the problem consists in reasonably factorizing data to which noise is added. The noise may be additive (some 0s in the original data are flipped to 1s), subtractive (some 1s are flipped to 0s) or general (both changes occur). The current literature presents various observations regarding the algorithms' ability to handle noise along with experiments used to demonstrate this ability.

We critically examine the existing considerations regarding noise, reported results, and approaches used to evaluate the ability to deal with noise. We point out weaknesses of existing considerations regarding noise and argue that, in several respects, the current understanding of the problem of noise is underdeveloped. Our main contributions are summarized as follows:

☆ Supported by grant No. GA15-17899S of the Czech Science Foundation.
* Corresponding author.
E-mail addresses: radim.belohlavek@acm.com (R. Belohlavek), martin.trnecka@gmail.com (M. Trnecka).

https://doi.org/10.1016/j.ijar.2018.03.006
0888-613X/

- we provide a new way to assess the ability of an algorithm to handle noise; our approach is based on a common-sense requirement of robustness: the factorizations computed by a robust algorithm should not be affected much by adding noise to data;
- we present methodological observations regarding the role of noise in BMF;
- we utilize a similarity measure developed to assess robustness to noise for assessment of another important property of BMF algorithms, namely the ability to recover ground truth in both noise-free and noisy data;
- we present an experimental evaluation of several existing algorithms. The experiments reveal important properties of these algorithms as regards their ability to handle noise. In addition to providing methodological justification of some properties claimed in the literature, they also reveal properties which were not reported as well as properties which counter certain claims made in the literature.

The paper is organized as follows. In the rest of this section, we review basic concepts, notation, and relevant work. In Section 2, we review current understanding of noise in Boolean data (Sections 2.1 and 2.2), present our rationale for the concept of robustness to noise (Section 2.3), critically examine the current approaches to handling noise in BMF (Section 2.4), and present further observations supporting our critical view (Section 2.5). Our new approach to assess robustness to noise, which reflects the rationale provided in Section 2.3, is presented in Section 3.1. The approach is based on measuring similarity of factorizations computed from data and we show in Section 3.2 that the similarity measure may also be used to assess another feature of factorization algorithms, namely, the ability to discover ground truth in data. Section 4 provides experimental evaluation. Conclusions and a list of topics for future research are provided in Section 5.

1.2. Basic concepts and notation

We denote an $n \times m$ Boolean matrix by M and interpret it as an object-attribute matrix (objects and attributes correspond to the matrix rows and columns). That is, the entry M_{ij} corresponding to the row i and the column j is either 1 or 0, indicating whether the object i does or does not have the attribute j, respectively. If objects are transactions and attributes are items, then M_{ij} indicates whether transaction i contains item j and the like. The set of all $n \times m$ Boolean matrices is denoted by $\{0,1\}^{n\times m}$; the ith row vector and the jth column vector of M is denoted by $M_{i\_}$ and $M_{\_j}$, respectively.

The general problem in BMF, of which two important variants—DBP and AFP—are described below, is to find for a given $M \in \{0,1\}^{n\times m}$ matrices $A \in \{0,1\}^{n\times k}$ and $B \in \{0,1\}^{k\times m}$ for which

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

where $\circ$ is the Boolean matrix product, i.e.

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

Decomposing M into $A \circ B$ may be interpreted as a discovery of k *factors that exactly or approximately explain the data*. If one interprets the matrices M, A, and B as the object-attribute, object-factor, and factor-attribute matrices, the factor model (1) says: 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 $M = A \circ B$ exists is called the *Boolean* (or Schein) *rank* of M.

The approximate equality in (1) is commonly approached in BMF via the L_1-norm (Hamming weight) $||\cdot||$ and the induced metric $E(\cdot,\cdot)$, defined for $C, D \in \{0,1\}^{n\times m}$ by

$$||C|| = \textstyle\sum_{i,j=1}^{m,n} |C_{ij}|, \tag{2}$$

$$E(C,D) = ||C - D|| = \textstyle\sum_{i,j=1}^{m,n} |C_{ij} - D_{ij}|. \tag{3}$$

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

- *Discrete Basis Problem* (DBP, [16,17]): Given $M \in \{0,1\}^{n\times m}$ and a positive integer k, find $A \in \{0,1\}^{n\times k}$ and $B \in \{0,1\}^{k\times m}$ that minimize $||M - A \circ B||$.
- *Approximate Factorization Problem* (AFP, [1,2]): Given $M \in \{0,1\}^{n\times m}$ and a 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 $||M - A \circ B|| \leq \varepsilon$.

DBP and AFP mirror two important views: DBP stresses the importance of the first k (presumably most important) factors; AFP stresses the need to account for (and thus to explain) a prescribed portion of data.

Below we employ the following geometric view: Computing an exact or approximate decomposition $M \approx A \circ B$ of $M \in \{0,1\}^{n\times m}$ into $A \in \{0,1\}^{n\times k}$ and $B \in \{0,1\}^{k\times m}$ is tantamount to finding k *rectangles* (called also *tiles*) whose union (max-superposition) exactly or approximately equals M. Here, a rectangle corresponds to a crossproduct $J = C \circ D \in \{0,1\}^{n\times m}$ of some column vector $C \in \{0,1\}^{n\times 1}$ and some row vector $D \in \{0,1\}^{1\times m}$; hence, by permuting the rows and columns, all the 1s in J form a rectangular area (tile). Now, if $M \approx A \circ B$, each factor $l = 1, \ldots k$ of the k factors involved in $A \circ B$ corresponds