Journal of Computer and System Sciences 81 (2015) 1678–1697

Contents lists available at ScienceDirect

Journal of Computer and System Sciences

www.elsevier.com/locate/jcss

From-below approximations in Boolean matrix factorization: Geometry and new algorithm

Radim Belohlavek *, Martin Trnecka

Data Analysis and Modeling Lab, Dept. Computer Science, Palacký University, Czech Republic

ARTICLE INFO

Article history:
Received 10 March 2014
Received in revised form 31 January 2015
Accepted 31 March 2015
Available online 11 June 2015

Keywords:
Boolean matrix
Matrix decomposition
Closure structures
Concept lattice
Approximation algorithm

ABSTRACT

We present new results on Boolean matrix factorization and a new algorithm based on these results. The results emphasize the significance of factorizations that provide from-below approximations of the input matrix. While the previously proposed algorithms do not consider the possibly different significance of different matrix entries, our results help measure such significance and suggest where to focus when computing factors. An experimental evaluation of the new algorithm on both synthetic and real data demonstrates its good performance in terms of good coverage by the first few factors as well as a small number of factors needed for an almost exact decomposition and indicates that the algorithm outperforms the available ones in these terms. We also propose future research topics.

1. Introduction

Boolean matrix factorization (BMF, called also Boolean matrix decomposition) is becoming an established method for analysis and preprocessing of data. The existing BMF methods are based on various types of heuristics because the main computational problems involved are known to be provably hard. The heuristics employed, however, use only a limited theoretical insight regarding BMF. This paper attempts to show that a better understanding of the geometry of Boolean data results in a better understanding of BMF, theoretically justified heuristics, and better algorithms.

In particular, we present new results in BMF derived from examining the closure and order-theoretic structures related to Boolean data, namely the lattice of all fixpoints of the Galois connections associated with the input matrix (so-called concept lattice). This viewpoint makes explicit the essence of BMF as a covering problem and emphasizes one type of factorizations we call from-below factorizations. Such factorizations and related notions were examined in some previous papers. While all the existing BMF methods consider the entries containing 1s in the input matrix essentially equally important, we propose to differentiate their role. In particular, we examine the entries that are essential for BMF in that their coverage by factors guarantees exact decomposition of the input matrix I by these factors. Crucial in our approach are intervals in the concept lattice associated with I. We show that every such interval contains just the factors covering a certain block full of 1s in I and that the intervals form reasonable subspaces for the search of factors. We present a new BMF algorithm which is based on these results and computes from-below factorizations. It turns out from experimental evaluation on both synthetic and real data that the new algorithm outperforms the existing BMF algorithms. Moreover, we clarify some connections between the existing approaches to BMF and argue that the closure and order-theoretic structures utilized in this paper provide a natural geometric view and represent a useful framework for theoretical analysis of the various BMF problems.

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

http://dx.doi.org/10.1016/j.jcss.2015.06.002
0022-0000/

2. Preliminaries and related work

2.1. Notation and basic notions

Throughout this paper, we denote by I an $n \times m$ Boolean matrix, interpreted primarily as an object–attribute incidence (hence the symbol I) matrix, i.e. the entry I_{ij} corresponding to the row i and the column j is either 1 or 0, indicating that the object i does or does not have the attribute j. The set of all $n \times m$ Boolean matrices is denoted by $\{0,1\}^{n\times m}$. The ith row and jth column vector of I is denoted by $I_{i\_}$ and $I_{\_j}$, respectively. A general aim in BMF is to find for a given $I \in \{0,1\}^{n\times m}$ (and possibly other given parameters) 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 the object–attribute, object–factor, and factor–attribute matrices, the 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* (Schein rank) of k and is denoted by $\mathrm{rank}_{\mathrm{B}}(I)$.

Recall that the L_1-norm (Hamming weight in case of Boolean matrices) $||\cdot||$ and the corresponding metric $E(\cdot,\cdot)$ are 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 variants of the BMF problem, relevant to this paper, are considered in the literature.

- *Discrete Basis Problem* (DBP, [20]):
  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 two problems reflect two important views on BMF. The first one emphasizes the importance of the first k (presumably most important) factors. The second one emphasizes the need to account for (and thus to explain) a prescribed portion of data, which is specified by ε.

2.2. Related work

Matrix decompositions represent an extensive subject whose coverage is beyond the scope of this paper. A good overview from BMF viewpoint is found e.g. in [20]. Except for the area of Boolean matrix theory itself, see e.g. [12], relevant results are traditionally presented in the literature on binary relations and graph theory, see e.g. [6,27]. These results may be translated to the results on Boolean matrices due to various correspondences, such as those connecting Boolean matrices, bipartite graphs, and binary relations, and pertain mostly to combinatorial and computational complexity questions. An important related area is formal concept analysis (FCA) [10], in which Boolean matrices are represented by so-called formal contexts, i.e. binary relations between objects and attributes. FCA provides solid lattice-theoretical foundations which we utilize.

Decompositions of Boolean matrices using methods designed originally for real-valued data and various modifications of these methods appear in a number of papers. [29] compares several approaches to assessment of dimensionality of Boolean data and concludes that a major problem with applying to Boolean data the methods designed originally for real-valued data is the lack of interpretability. Similar observations were presented by other authors. Among the first works on applications of BMF in data analysis are [24,25], in which the authors have already been aware of the provable computational difficulty (NP-hardness) of the decomposition problem due to NP-hardness of the set basis problem [28]. The interest in BMF in data mining is primarily due to the work of Miettinen et al. In particular, the DBP, the corresponding complexity results, and the Asso algorithm discussed below appeared in [20]. In [11], they authors examine "tiling" of Boolean data and various related problems, their complexity, and algorithms. Tiling is closely related to BMF as it corresponds to the from-below factorizations we investigate in this paper and is discussed in more detail in Section 5.1. In [4], we showed that formal concepts (i.e. fixpoints of Galois connections) are natural factors of Boolean matrices, proved their optimality for exact factorizations, described transformations between attribute and factor spaces, and proposed two BMF algorithms discussed below. In [31], the authors investigate the problem of summarizing transactional databases by so-called hyperrectangles, examine the computational complexity of the problems involved, provide the Hyper algorithm and discuss related problems. The summarizations involved may be rephrased as Boolean matrix decompositions and this approach is discussed in more detail in Sections 3 and 5. Directly relevant to our paper is also [16], where the authors propose the PaNDa algorithm discussed in Section 5. In particular, we use the algorithms proposed in [4,11,16,20,31] in the experimental evaluation of our new algorithm. Further work