Handling Noise in Boolean Matrix Factorization

Radim Belohlavek, Martin Trnecka

DEPARTMENT OF COMPUTER SCIENCE
PALACKÝ UNIVERSITY OLOMOUC

26th International Joint Conference on Artificial Intelligence (IJCAI-17)
August 19-25, 2017

Outline

Boolean Matrix Factorization (BMF)

- 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

What is Noise in Boolean Data?

- noise in Boolean data = distortion of data (i.e. flipping some data entries of true data)
- subtractive noise $1 \to 0$
- aditive noise $0 \to 1$
- general noise $1 \to 0$ and $0 \to 1$

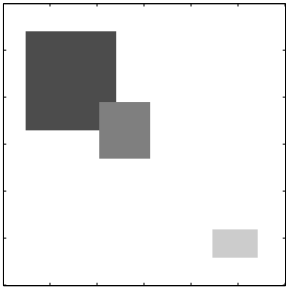
Main Points of Our Work

- we point out weaknesses of noise in BMF
- ability of BMF algorithms to deal with noise
- current understanding is underdevelopment
- missing important aspects
- new way to assess the ability of an algorithm to handle noise
- experimental evaluations

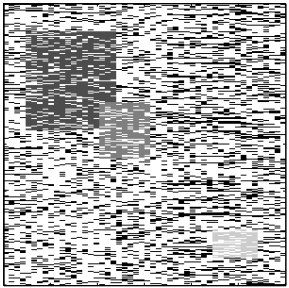
Is Noise Always a Reasonable Assumption?

- term "noise" → strange
- noise = random and mostly small fluctuations in data
- Boolean data → "complete change"
- term "error"
- many real datasets do not contain noise because they simply contain verified truth
- there exist applications of BMF, in which presence of noise would be counterintuitive or even damaging (e.g. role mining problem)
- which levels of noise are realistic (some works consider 40% noise)

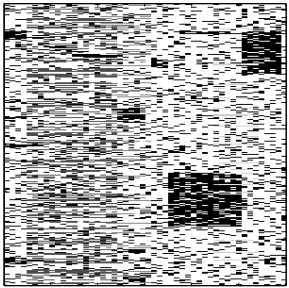
Example of 40% noise



(a) Original data



(b) 40% Noise added



(c) Permuting

A Rationale for Robustness to Noise

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

Observed data

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

Hypothetical true data

- algorithms do not committing "overcover error" are not able to discover these factors

A Critique of Current Approach

- current experiments → robustness to noise
- wrong approach
- mix of three distinctive terms: coverage quality, robustness to noise and ground truth

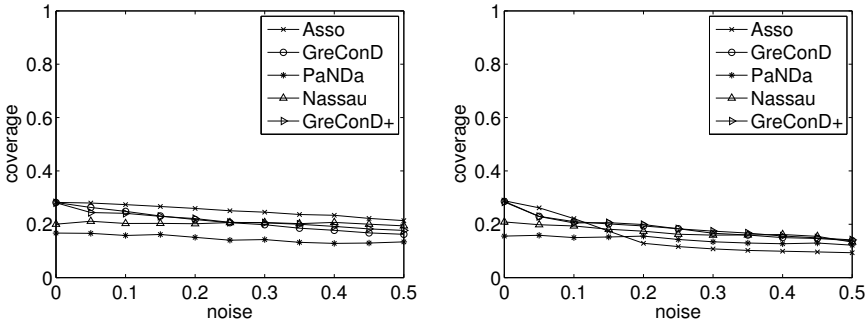
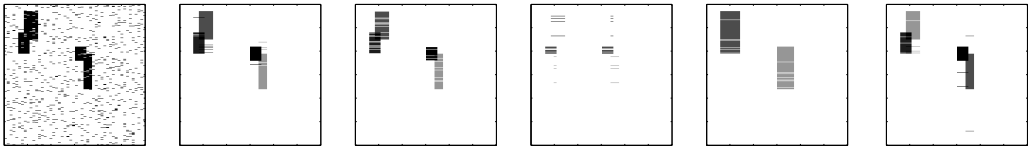


Figure 1: Current experiments to assess robustness to noise

- wrong view: small decrease in coverage quality indicates robustness to noise

A Critique of Current Approach

- ignores: more noise → the factors may have changed → lack of robustness
- no information if a particular algorithm found the factors used to generate the data
- moreover our observations are different from those reported in literature
- R. Gupta, G. Fang, B. Field, M. Steinbach, and V. Kumar. Quantitative evaluation of approximate frequent pattern mining algorithms. *In Proc. of the 14th ACM SIGKDD (KDD'08)*, pp. 301–309, 2008.



New Way to Assess Robustness to Noise

- factor = rectangular area in data, i.e. the Cartesian product $C \times D$ for some $C \subseteq \{1, \ldots, n\}$ and $D \subseteq \{1, \ldots, m\}$
- for two sets $\mathcal{F}$ and $\mathcal{F}'$ of factors we define

$$sim(\mathcal{F}, \mathcal{F}') = \min\left(\frac{\sum_{c \in \mathcal{F}} \max_{c' \in \mathcal{F}'} s(c, c')}{|\mathcal{F}|}, \frac{\sum_{c' \in \mathcal{F}'} \max_{c \in \mathcal{F}} s(c, c')}{|\mathcal{F}'|}\right)$$

where

$$s(\langle C_1, D_1 \rangle, \langle C_2, D_2 \rangle) = \frac{|C_1 \times D_1 \cap C_2 \times D_2|}{|C_1 \times D_1 \cup C_2 \times D_2|}$$

- assess capability to discover ground truth

Experimental evaluation

- selected algorithms: 8M, Tiling, Asso, GreConD, PaNDa, Hyper, GreEss, Nassau, GreConD+
- real data → how the original factors are change
- synthetic data → how the ground truth is change
- revision of current experiments

Robustness to Noise

k	Noise (%)	8M	Tiling	Asso	GreConD	PaNDa	Hyper	GreEss	Nassau	GreConD+
5	0.1	0.985	0.228	1.000	0.228	1.000	0.063	0.239	0.722	1.000
	0.5	0.984	0.210	0.998	0.210	0.998	0.063	0.220	0.358	0.998
	1	0.789	0.186	0.998	0.187	0.992	0.063	0.195	0.261	0.998
	2	0.834	0.161	0.998	0.161	0.995	0.063	0.167	0.306	0.998
	5	0.760	0.111	0.997	0.112	0.880	0.063	0.115	0.236	0.982
10	0.1	1.000	0.459	1.000	0.459	0.833	0.127	0.481	0.777	1.000
	0.5	0.976	0.415	1.000	0.415	0.813	0.123	0.434	0.548	1.000
	1	0.893	0.373	0.996	0.375	0.764	0.115	0.389	0.584	0.999
	2	0.963	0.315	0.995	0.313	0.802	0.108	0.321	0.442	0.995
	5	0.789	0.223	0.946	0.224	0.616	0.104	0.229	0.442	0.980
15	0.1	0.995	0.825	0.994	0.904	0.833	0.248	0.924	0.802	0.999
	0.5	0.954	0.759	0.987	0.826	0.778	0.241	0.830	0.685	0.994
	1	0.908	0.625	0.933	0.687	0.831	0.225	0.686	0.519	0.997
	2	0.888	0.542	0.893	0.573	0.751	0.211	0.573	0.432	0.972
	5	0.742	0.366	0.863	0.360	0.725	0.192	0.355	0.421	0.921

Table 1: Robustness to noise (Domino dataset, general noise)

Recovery of Ground Truth

Noise	Change (%)	8M	Tiling	Asso	GreConD	PaNDa	Hyper	GreEss	Nassau	GreConD+
Aditive	0.1	0.887	0.726	0.974	0.726	0.728	0.015	0.726	0.520	0.932
	0.5	0.854	0.482	0.974	0.482	0.697	0.014	0.482	0.612	0.928
	1	0.791	0.353	0.973	0.356	0.592	0.014	0.356	0.688	0.927
	2	0.788	0.257	0.973	0.260	0.631	0.013	0.260	0.747	0.929
	5	0.760	0.146	0.964	0.151	0.579	0.012	0.151	0.867	0.922
Subtractive	0.1	0.632	0.683	0.991	0.683	0.650	0.085	0.715	0.732	0.914
	0.5	0.593	0.443	0.991	0.441	0.673	0.084	0.461	0.637	0.895
	1	0.555	0.288	0.983	0.290	0.660	0.084	0.326	0.704	0.858
	2	0.533	0.209	0.956	0.213	0.682	0.084	0.238	0.677	0.821
	5	0.591	0.105	0.855	0.105	0.555	0.083	0.104	0.537	0.665

Table 2: Recovery of ground truth on synthetic data 500 × 250 with $k = 5$

Conclusion

- we show that something is wrong
- new methodological ground
- clear separation of algorithms

Thank you

come visit my poster #3368