How to assess quality of BMF algorithms?

Radim Belohlavek, Jan Outrata, Martin Trnecka

DEPARTMENT OF COMPUTER SCIENCE
PALACKÝ UNIVERSITY OLOMOUC
CZECH REPUBLIC

IEEE International Conference on Intelligent systems IS'16
Sofia, Bulgaria, September 4-6, 2016

Motivation

- Boolean matrix factorization (BMF).
- Method for analysis of Boolean data.
- Various algorithms (more than 25).
- How to assess their quality?

Boolean Matrix Factorization

- A 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 in data.

Computational Complexity

- Basic feature of each algorithm.
- We prefer algorithm with the smaller complexity.
- Big O notation (hides several issues).
- Better way: relative time complexity.
- "One algorithm is three-times faster than other."
- Time (and space) complexity is not critical issue for the most of current algorithms.
- Runable on ordinar PC.

Approximation Factor

- Optimization version of the basic decomposition problem is NP-hard.
- No polynomial time algorithm (computing exact solution) exists.
- Based on heuristic - approximation factor.
- Recent results on approximability: basic decomposition problem is NP-hard to approximate within factor $n^{1-\epsilon}$.
- Lower bound is not encouraging.
- Current algorithm produce much better results.

Quality of Factors

1. Geometry of factorization
2. Interpretability of individual factors
   - Knowledge discovery view
3. Quality of a set of extracted factors
   - Reduction of dimensionality
   - Explanatory view

Explanation of Data by Factors

- Distance (error function)

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

- Two components of E

$$\begin{aligned} E(I, A \circ B) &= E_u(I, A \circ B) + E_o(I, A \circ B), \text{ where} \\ E_u(I, A \circ B) &= |\{\langle i, j\rangle \,;\, I_{ij} = 1, (A \circ B)_{ij} = 0\}|, \\ E_o(I, A \circ B) &= |\{\langle i, j\rangle \,;\, I_{ij} = 0, (A \circ B)_{ij} = 1\}|. \end{aligned}$$

- Coverage quality

$$c(l) = 1 - E(I, A \circ B)/||I||.$$

Two Basic Viewpoint to BMF

- Discrete Basis Problem
  - 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 $||I - A \circ B||$.
  - Emphasizes the importance of the first few (presumably most important) factors.
  - *Miettinen P., Mielikainen T., Gionis A., Das G., Mannila H., The discrete basis problem, IEEE Transactional Knowledge and Data Engineering 20(10)(2008), 1348–1362*
- Approximate Factorization Problem
  - 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 $||I - A \circ B|| \leq \varepsilon$.
  - Emphasizes the need to account for (and thus to explain) a prescribed (presumably reasonably large) portion of data.
  - *Belohlavek R., Trnecka M., From-below approximations in Boolean matrix factorization: Geometry and new algorithm, Journal of Computer and System Science 81(8)(2015), 1678–1697.*

Quality Measure

- $w_l = l/k$ for the DBP view
- $w_l = 1 + (E(I, A \circ B) - \varepsilon)/(||I|| - \varepsilon)$ for AFP view
- $w_l = (l/k + 1 + (E(I, A \circ B) - \varepsilon)/(||I|| - \varepsilon))/2$ combined view.

$$q = 1 - \left(\sum_{j=0}^{l} w_j \frac{E(I, A \circ B)}{||I||} \right) / \left(\sum_{j=0}^{l} w_j \right).$$

- Reflect natural requirement for a good decomposition.

Interpretation

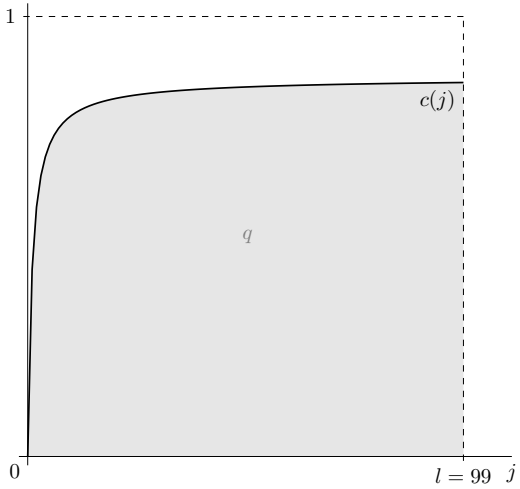

Figure: Measure of quality of BMF algorithm

Experimental Evaluation

- Asso—*Miettinen P., Mielikainen T., Gionis A., Das G., Mannila H., The discrete basis problem, IEEE Transactional Knowledge and Data Engineering 20(10)(2008), 1348–1362.*
- GreConD—*Belohlavek R., Vychodil V., Discovery of optimal factors in binary data via a novel method of matrix decomposition, Journal of Computer and System Science 76(1)(2010), 3–20.*
- NaiveCol—*Ene A. et al., Fast exact and heuristic methods for role minimization problems. Proc. SACMAT 2008, pp. 1–10.*
- Panda—*Lucchese C., Orlando S., Perego R., Mining top-K patterns from binary datasets in presence of noise, SIAM DM 2010, pp. 165–176.*
- Hyper—*Xiang Y., Jin R., Fuhry D., Dragan F. F., Summarizing transactional databases with overlapped hyperrectangles, Data Mining and Knowledge Discovery 23(2011), 215–251*
- GreEss—*Belohlavek R., Trnecka M., From-below approximations in Boolean matrix factorization: Geometry and new algorithm, Journal of Computer and System Science 81(8)(2015), 1678–1697.*

Results

Table: Numbers of factors and coverage quality

Dataset		Asso	GreConD	NaiveCol	Hyper	PaNDa	GreEss
Mushroom	$c = 80\,\%$	19	29	32	42	NA	31
	$c = 90\,\%$	34	46	47	57	NA	47
	$c = 95\,\%$	50	62	62	70	NA	61
	$c = 100\,\%$	NA	120	110	123	NA	105
	$k = 10$	0.556	0.582	0.512	0.285	0.346	0.546
	$k = 20$	0.652	0.715	0.674	0.502	0.346	0.696
	$k = 30$	0.720	0.812	0.789	0.664	0.346	0.793
	$k = 40$	0.765	0.873	0.862	0.780	0.346	0.865

Results

Table: BMF algorithm quality

Dataset		Asso	GreConD	NaiveCol	Hyper	PaNDa	GreEss
Mushroom	$q_{0.8}$	0.622	0.740	0.729	0.657	0.344	0.733
	$q_{0.9}$	0.695	0.801	0.786	0.709	0.344	0.794
	$q_{0.95}$	0.725	0.827	0.810	0.728	0.344	0.819
	q_{1}	0.745	0.844	0.827	0.749	0.344	0.835
	q_{10}	0.556	0.582	0.511	0.285	0.346	0.545
	q_{20}	0.650	0.712	0.671	0.498	0.346	0.693
	q_{30}	0.715	0.805	0.781	0.654	0.346	0.786
	q_{40}	0.756	0.861	0.848	0.760	0.346	0.851
	$q_{10,0.9}$	0.764	0.876	0.863	0.798	0.344	0.870
	$q_{20,0.8}$	0.763	0.874	0.860	0.792	0.344	0.867

Conclusion

- We point out an important problem in BMF: assessment of quality of BMF algorithms.
- We identify key aspects of such assessment.
- We propose quantitative ways how to measure quality of BMF algorithms.

Thank you