Machine learning and data mining 2

Factor analysis

Radim Bělohlávek

DEPARTMENT OF COMPUTER SCIENCE
PALACKÝ UNIVERSITY, OLOMOUC

Literature

- L. Eldén: Matrix Methods in Data Mining and Pattern Recognition. SIAM, 2007.
- D. Skillicorn, Understanding Complex Datasets. Data Mining with Matrix Decompositions. Chapman & Hall/CRC, Boca Raton, FL, 2007.
- For Boolean factor analysis these slides and papers on the website.

Factor analysis

- very broad term, understood differently in different contexts
- "classical" factor analysis goes back to the early 20th century
- psychology
- nowadays many methods classified as factor analytic methods:
  - classical factor analysis (term not uniquely used)
  - principal component analysis
  - independent compoent analysis
  - Boolean factor analysis

Our visit to factor analysis

- principal component analysis
- Boolean factor analysis

Boolean factor analysis

Sources

- [Miea08] P. Miettinen, T. Mielikäinen, A. Gionis, G. Das, H. Mannila, The discrete basis problem, IEEE Trans. Knowl. Data Eng. 20 (10) (2008) 1348–1362 (see webpage)
- [BeVy10] R. Belohlavek, V. Vychodil, Discovery of optimal factors in binary data via a novel method of matrix decomposition, J. Comput. Syst. Sci. 76 (1) (2010) 3–20 (see webpage)
- [BeTr15] R. Belohlavek, M. Trnecka, From-below approximations in Boolean matrix factorization: Geometry and new algorithm, J. Comput. Syst. Sci. 81 (2015) 1678–1697 (see webpage)
- how to study: follow slides first and the links to the topics in the papers

Problem description

- basic description including interpretation of factors: [BeVy10] 3–4
- problems AFP and DBP: [BeTr15] sec. 2.1
- illustrative example: [BeVy10] sec. 2.2
- NP-hardness of the exact factorization problem: [BeVy10] sec. 2.4. 2

  (several variants of the problem are also NP-hard)
- applications
  - knowledge discovery aspect factors = most important concepts/clusters in data

    (explain the whole data or a large portion of data)
  - reduction of dimensionality aspect

    object may be represented in the less dimensional space of factors, rather than in the space of attributes: [BeVy10] sec. 2.3

    Belohlavek, R., Outrata, J., Trnecka, M.: Impact of Boolean factorization as preprocessing methods for classification of Boolean data. Annals of Mathematics and Artificial Intelligence 72(1)(2014), 3–22.

    `http://belohlavek.inf.upol.cz/publications/BeOuTr_Ibfpmcbd.pdf`

Factorization using formal concepts and GreConD

- preliminaries in formal concept analysis: [BeVy10] sec. 1.4
- formal concepts are universal and optimal factors: [BeVy10] sec. 2.1
- algorithm GreConD [BeVy10] sec. 2.5
  - primarily for exact decomposition probem and generally for AFP
  - Algorithm 1 in [BeVy10] just for reference
  - Algorithm 2 in [BeVy10] is nowadays known as GreConD
- experimental evaluation (student needs to know what is being observed)

  [BeVy10] sec. 3

  [BeTr15] sec. 5

- extensions (just for reference):
  - several extensions of GreConD, such as GreEss described in [BeTr15]

    (some ideas in lecture)
  - extension from binary data to ordinal data (papers on my website)

    theory: Belohlavek R.: Optimal decompositions of matrices with entries from residuated lattices. Journal of Logic and Computation 22(6)(2012), 1405-1425.

    examples: Belohlavek, R., Krmelova, M.: Factor analysis of ordinal data via decomposition of matrices with grades. Annals of Mathematics and Artificial Intelligence 72(1)(2014), 23-44.

Asso

- primarily for DBP
- presented in [Miea08] P. Miettinen, T. Mielikäinen, A. Gionis, G. Das, H. Mannila, The discrete basis problem, IEEE Trans. Knowl. Data Eng. 20 (10) (2008) 1348–1362 (see webpage)
- We provide a self-contained presentation since the description in the above paper is somehwat ambiguous.

Notes to the problem and algorithm:

- Given a Boolean matrix $C \in \{0,1\}^{n \times m}$, desired number k of factors, and parameters
  - threshold value $\tau \in (0,1]$ (to be used for computing association matrix),
  - real-valued weights w^+ and w^- (penalty for overcovering nad undercovering)

  the algorithm looks for matrices $S \in \{0,1\}^{n \times k}$ and $B \in \{0,1\}^{k \times m}$ such that

$$||C - S \circ B|| \text{ is small.}$$

- Idea:
  - compute from C and τ the association matrix $A \in \{0,1\}^{m \times m}$
  - for $i = 1, \ldots, k$: compute factors $\langle e_i, A_{i*} \rangle$
    i.e. $e_i \in \{0,1\}^{n \times 1}$ and $A_{i*} \in \{0,1\}^{n \times 1}$
  - rows of B are A_{i*}s (rows of association matrix)
  - columns of S are e_is (detemined to maximize coverage, see below)

- Association matrix $A \in \{0,1\}^{m \times m}$:
  - $$A_{ij} = \begin{cases} 1 & \text{if confidence of association rule } \{i\} \Rightarrow \{j\} \ \geq \ \tau \\ 0 & \text{otherwise.} \end{cases}$$
  - meaning of (validity of) association rule $\{i\} \Rightarrow \{j\}$:
    if an object o has attribute i then o has also attribute j
    in terms of the object-attribute matrix C: if $C_{oi} = 1$ then $C_{oj} = 1$ for $o = 1, \ldots, n$
  - confidence of $\{i\} \Rightarrow \{j\} = \frac{|\{o \ ; \ C_{oi}=1 \text{ and } C_{oj}=1\}|}{|\{o \ ; \ C_{oi}=1\}|}$
    i.e. confidence is the proportion of objects having also j from objects having i

- Example of association matrix
  - let

$$C = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix} \quad \text{and} \quad \tau = 0.5.$$

  - denoting the attributes 1, 2, and 3, we have

$$\text{confidence of } \{1\} \Rightarrow \{2\} = \frac{|\{o\ ;\ C_{o1} = 1 \text{ and } C_{o2} = 1\}|}{|\{o\ ;\ C_{o1} = 1\}|} = \frac{1}{2} \geq \tau, \text{ hence } A_{12} = 1.$$

  - the corresponding association matrix:

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

- The notions of coverage:
  - consider C and candidates $\langle S_{*1}, B_{1*} \rangle$ for a first factor

$$C = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix}, \quad S_{*1} = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \end{pmatrix}, \quad B_{1*} = \begin{pmatrix} 1 & 1 & 0 \end{pmatrix}$$

  - one may consider an approximate decomposition using this first factor:

$$C \approx S_{*1} \circ B_{1*}, \text{ i.e. } \begin{pmatrix} 0 & 0 & 0 \\ \mathbf{1} & \mathbf{1} & 0 \\ \mathbf{1} & \underline{0} & 0 \\ 0 & \mathit{1} & \mathit{1} \end{pmatrix} \approx \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \end{pmatrix} \circ \begin{pmatrix} 1 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

    in which the bold, italicized, and underlined entries of C are covered, uncovered, and overcovered
  - rule: the next computed factor should cover a large number of yet uncovered 1s in C and should overcover a small number of 0s in C

- Computing $\langle e_i, A_{i*} \rangle$:
  - for each row A_{l*} of A, compute the best column e_l in that the improvement c_l in coverage (see below) is the highest possible;
  - best means that adding A_{l*} as a new row to B and e_l as a new column to S has the highest improvement in coverage
  - select as $\langle e_i, A_{i*} \rangle$ the one for which $c_i = \max\{c_1, \ldots, c_m\}$
  - improvement in coverage c_l of a candidate factor $\langle e_i, A_{i*} \rangle$:

$$\begin{aligned} c_l \quad = \quad & w^+ \cdot \# \text{ previously uncovered 1s in } C \text{ that are covered by } \langle e_i, A_{i*} \rangle \\ & -w^- \cdot \# \text{ previously not overcovered 0s in } C \text{ that are overcovered by } \langle e_i, A_{i*} \rangle \end{aligned}$$

  - selection of best e_l for the given A_{l*} is not exponential even though there is $|2^n|$ possible candidates for e_l
    namely, e_l is selected coordinate by coordinate: select the best $(e_l)_1$, ..., select the best $(e_l)_n$
  - justification: e_l is best iff if each $(e_l)_i$ is best in terms of coverage

Asso pseudocode

Asso(C, k, τ, w^+, w^-)
input: $C \in \{0,1\}^{n \times m}$, $k \geq 1$, $\tau \in (0,1]$, $w^+, w^- \geq 0$
output: $S \in \{0,1\}^{n \times k}$ and $B \in \{0,1\}^{k \times m}$

1. **for** $i = 1, \ldots, m$ **do** [compute association matrix]
2. **for** $i = 1, \ldots, m$ **do**
3. **if** confidence of $\{i\} \Rightarrow \{j\} \geq \tau$ **then** $A_{ij} \leftarrow 1$ **else** $A_{ij} \leftarrow 0$
4. $S \leftarrow$ empty matrix, $B \leftarrow$ empty matrix
5. **for** $l = 1, \ldots, k$ **do** [compute all factors]
6. **for** $j = 1, \ldots, m$ **do** [search possible new factor]
7. $\langle c_i, e_i \rangle \leftarrow \text{Cover}(A_{j*}, C, S, B, w^+, w^-)$
8. **select** i for which c_i is maximal [pick new factor]
9. add e_i as a new column to S and add A_{j*} as a new row to B [add new factor]
10. **return** S and B

Cover pseudocode

Cover(a, C, S, B, w^+, w^-)
input: $a \in \{0,1\}^{1\times m}$ (row of association matrix A), $C \in \{0,1\}^{n\times m}$, Boolean matrices S and B (computed so far by Asso), $w^+, w^- \geq 0$
output: $S \in \{0,1\}^{n\times k}$ and $B \in \{0,1\}^{k\times m}$

1. $uncovered \leftarrow 1_{n\times m}$
2. $e \leftarrow 0_{n\times 1}$ [initial value of best counterpart to a]
3. **foreach** $\langle i, j\rangle$ **do**
4. **if** $(S \circ B)_{ij} = 1$ **then** $uncovered_{ij} \leftarrow 0$ [already covered or overcovered entries]
5. **foreach** row C_{i*} of C **do** [determine best value of e_i]
6. $partial\text{-}cover_i \leftarrow w^+ \sum_{j=1}^{m} a_j \cdot uncovered_{ij} - w^- \sum_{j=1}^{m}(1 - a_j) \cdot uncovered_{ij}$
7. **if** $partial\text{-}cover_i < 0$ **then** $partial\text{-}cover_i = 0$ **else** $e_j = 1$
8. $cover = \sum_{i=1}^{n} partial\text{-}cover_i$
9. **return** $\langle cover, e\rangle$

Example in practice lesson.

Comparison of GreConD and Asso:

- GeConD utilizes formal concepts as factors, hence does not commit any overcover error
- Asso does commit overcover error
- Pros and cons:
  - Asso may construct more general decompositions compared to GreConD (disadvantage for GreConD).
  - One overcover error is committed by a computed factor, it cannot be removed by adding further factors (disadvantage for Asso).
  - For AFP, GreConD performs much better.
  - For DBP, Asso performs slightly better (but this depends on the number k of required factors).
  - See [BeTr15] and practice lesson for comparatory experiments.
- Several variants improving the basic GreConD and Asso (some details in lecture).