Boolean Factor Analysis of Multi-Relational Data

Martin Trnecka, Marketa Trneckova

DEPARTMENT OF COMPUTER SCIENCE

PALACKÝ UNIVERSITY OLOMOUC

Motivation

- The Boolean factor analysis (BFA) is an established method for analysis and preprocessing of Boolean data.
- The basic task in the BFA: find new variables (factors) which explain or describe original single input data.
- Finding factors is an important step for understanding and managing data.
- Boolean nature of data is in this case beneficial especially from the standpoint of interpretability of the results.
- BFA is suitable for single input Boolean data table with just one relation between objects and attributes.
- Many real-world data sets are more complex than a simple data table.
- We propose new approach to the BFA, which is tailored for multi-relational data.

Multi-Relational Data

- Usually, they are composed from many data tables, which are interconnected by relations.
- Relations are crucial.
- Represent additional information about the relationship between data tables.
- This information is important for understanding data as a whole.
- Example: Social networks, Dating agency database.

Related Work

- Hacene M. R., Huchard M., Napoli A., Valtechev P.: Relational concept analysis: mining concept lattices from multi-relational data.
- Our approach is different from the RCA!
- Iteratively merge data tables into one.
- All formal concepts of one data table are used as additional attributes for the merged data table.
- Our approach delivers more informative results than a simple use of BMF on merged data table

Boolean Factor Analysis

- Consider an $n \times m$ object-attribute matrix C with entries $C_{ij} \in \{0, 1\}$ expressing whether an object i has an attribute j or not.
- The goal of the BMF is to find decomposition

$$C = A \circ B$$

of C into a product of an $n \times k$ object-factor matrix A over $\{0, 1\}$, a $k \times m$ factor-attribute matrix B over $\{0, 1\}$.

- The product $\circ$ in (4) is a Boolean matrix product, defined by

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

where $\bigvee$ denotes maximum (truth function of logical disjunction) and $\cdot$ is the usual product (truth function of logical conjunction). For example the following matrix can be decomposed into two Boolean matrices with $k < m$.

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

Boolean Factor Analysis via FCA

- An optimal decomposition of the Boolean matrix can be found via FCA
- Factors are represented by formal concepts.
- The aim is to decompose the matrix C into a product $A_{\mathcal{F}} \circ B_{\mathcal{F}}$.

$$\mathcal{F} = \{\langle A_1, B_1\rangle, \ldots, \langle A_k, B_k\rangle\} \subseteq \mathcal{B}(X, Y, C),$$

  where $\mathcal{B}(X, Y, C)$ represents set of all formal concepts of context $\langle X, Y, C\rangle$. Denote by $A_{\mathcal{F}}$ and $B_{\mathcal{F}}$ the $n \times k$ and $k \times m$ binary matrices defined by

$$(A_{\mathcal{F}})_{il} = \begin{cases} 1 \text{ if } i \in A_l \\ 0 \text{ if } i \notin A_l \end{cases} \qquad (B_{\mathcal{F}})_{lj} = \begin{cases} 1 \text{ if } j \in B_l \\ 0 \text{ if } j \notin B_l \end{cases}$$

  for $l = 1, \ldots, k$. In other words, $A_{\mathcal{F}}$ is composed from characteristic vectors A_l. Similarly for $B_{\mathcal{F}}$. The set of factors is a set $\mathcal{F}$ of formal concepts of $\langle X, Y, C\rangle$, for which holds $C = A_{\mathcal{F}} \circ B_{\mathcal{F}}$. For every C such a set always exists.
- Because a factor can be seen as a formal concept, we can consider the intent part (denoted by $intent(F)$) and the extent part (denoted by $extent(F)$) of the factor F.

Boolean Factor Analysis of Multi-Relational Data

- Our settings: We have two Boolean data tables C_1 and C_2, which are interconnected with relation $\mathcal{R}_{C_1C_2}$.
- This relation is over the objects of first data table C_1 and the attributes of second data table C_2, i.e. it is an objects-attributes relation.
- In general, we can also define an objects-objects relation or an attributes-attributes relation.
- Our goal: is to find factors, which explain the original data and which take into account the relation $\mathcal{R}_{C_1C_2}$ between data tables.

Definition

Relation factor (pair factor) on data tables C_1 and C_2 is a pair $\left\langle F_1^i, F_2^j \right\rangle$, where $F_1^i \in \mathcal{F}_1$ and $F_2^j \in \mathcal{F}_2$ ($\mathcal{F}_i$ denotes set of factors of data table C_i) and satisfying relation $\mathcal{R}_{C_1C_2}$.

- There are several ways how to define the meaning of "satisfying relation" from Definition.

Narrow Approach

- F_1^i and F_2^j form pair factor $\langle F_1^i, F_2^j \rangle$ if holds:

$$\bigcap_{k \in extent(F_1^i)} \mathcal{R}_k \neq \emptyset \text{ and } \bigcap_{k \in extent(F_1^i)} \mathcal{R}_k \subseteq intent(F_2^j),$$

  where $\mathcal{R}_k$ is a set of attributes, which are in relation with an object k.
- This definition holds for an object-attribute relation, other types of relations can be defined in similar way.

Wide Approach

- F_1^i and F_2^j form pair factor $\langle F_1^i, F_2^j \rangle$ if holds:

$$\left(\left(\bigcap_{k \in extent(F_1^i)} \mathcal{R}_k \right) \cap intent(F_1^j) \right) \neq \emptyset.$$

- This definition holds for an object-attribute relation, other types of relations can be defined in similar way.

α-approach

- For any $\alpha \in [0, 1]$, F_1^i and F_2^j form pair factor $\langle F_1^i, F_2^j \rangle$ if holds:

$$\frac{\left|\left(\bigcap_{k \in extent(F_1^i)} \mathcal{R}_k\right) \cap intent(F_2^j)\right|}{\left|\bigcap_{k \in extent(F_1^i)} \mathcal{R}_k\right|} \geq \alpha.$$

- This definition holds for an object-attribute relation, other types of relations can be defined in similar way.
- It is obvious, that for $\alpha = 0$ and replacing $\geq$ by $>$, we get the wide approach and for $\alpha = 1$, we get the narrow one.

Lemma

For $\alpha_1 > \alpha_2$ holds, that a set of relation factors counted by α_1 is a subset of a set of relation factors obtained with α_2.

Simple Example

Let us have two data tables C_W and C_M. C_W represents women and their characteristics and C_M represents men and their characteristics.

Table : C_W

	athlete	*undergraduate*	*wants kids*	*is attractive*
Abby		×	×	×
Becky	×		×	
Claire		×		×
Daphne	×	×	×	×

Table : C_M

	athlete	*undergraduate*	*wants kids*	*is attractive*
Adam	×			×
Ben		×	×	
Carl	×	×	×	
Dave			×	×

Table : $\mathcal{R}_{C_W C_M}$

	athlete	*undergraduate*	*wants kids*	*is attractive*
Abby		×	×	
Becky	×		×	
Claire	×	×		×
Daphne	×	×	×	×

Moreover, we consider relation $\mathcal{R}_{C_W C_M}$ between the objects of first the data table and the attributes of the second data table. In this case, it could be a relation with meaning "woman looking for a man with the characteristics".

Factors Obtained via GreConD Algorithm

Factors of data table C_W are:

- $F_1^W = \langle\{\text{Abby, Daphne}\}, \{\textit{undergraduate, wants kids, is attractive}\}\rangle$
- $F_2^W = \langle\{\text{Becky, Daphne}\}, \{\textit{athlete, wants kids}\}\rangle$
- $F_3^W = \langle\{\text{Abby, Claire, Daphne}\}, \{\textit{undergraduate, is attractive}\}\rangle$

Factors of data table C_M are:

- $F_1^M = \langle\{\text{Ben, Carl}\}, \{\textit{undergraduate, wants kids}\}\rangle$
- $F_2^M = \langle\{\text{Adam}\}, \{\textit{athlete, is attractive}\}\rangle$
- $F_3^M = \langle\{\text{Adam, Carl}\}, \{\textit{athlete}\}\rangle$
- $F_4^M = \langle\{\text{Dave}\}, \{\textit{wants kids, is attractive}\}\rangle$

Joint Factors Into Relational Factors

- We use so far unused relation $\mathcal{R}_{C_W C_M}$, between C_W and C_M to joint factors of C_W with factors of C_M into relational factors. For the above defined approaches we get results which are shown below. We write it as binary relations, i.e F_W^i and F_M^j belongs to relational factor $\langle F_W^i, F_M^j \rangle$ iff F_W^i and F_M^j are in relation:

Narrow approach

	F_M^1	F_M^2	F_M^3	F_M^4
F_W^1	$\times$			
F_W^2				
F_W^3	$\times$			

Wide approach

	F_M^1	F_M^2	F_M^3	F_M^4
F_W^1	$\times$			$\times$
F_W^2	$\times$	$\times$	$\times$	$\times$
F_W^3	$\times$			

0.6-approach

	F_M^1	F_M^2	F_M^3	F_M^4
F_W^1	$\times$			
F_W^2		$\times$		
F_W^3	$\times$			

0.5-approach

	F_M^1	F_M^2	F_M^3	F_M^4
F_W^1	$\times$			$\times$
F_W^2		$\times$		
F_W^3	$\times$			

Interpretation

The relational factor in form $\langle F_W^i, F_M^j \rangle$ can be interpreted in the following ways:

- Women, who belong to extent of F_W^i like men who belong to extent of F_M^j. Specifically in this example, we can interpret factor $\langle F_W^1, F_M^1 \rangle$, that Abby and Daphne should like Ben and Carl.
- Women, who belong to extent of F_W^i like men with characteristic in intent of F_M^j. Specifically in this example, we can interpret factor $\langle F_W^1, F_M^1 \rangle$, that Abby and Daphne should like undergraduate men, who want kids.
- Women, with characteristic from intent F_W^i like men who belong to extent F_M^j. Specifically in this example, we can interpret factor $\langle F_W^1, F_M^1 \rangle$, that undergraduate, attractive women, who want kids should like Ben and Carl.
- Women, with characteristic from intent F_W^i like men with characteristic in intent of F_M^j. Specifically in this example, we can interpret factor $\langle F_W^1, F_M^1 \rangle$, that undergraduate, attractive women, who want kids should like undergraduate men, who want kids.

Interpretation

- Interpretation of the relation between F_W^i and F_M^j is driven by used approach.
- If we obtain factor $\langle F_W^i, F_M^j \rangle$ by narrow approach, we can interpret relation between F_W^i and F_M^j: "women who belong to F_W^i, like men from F_M^j completely". For example factor $\langle F_W^1, F_M^1 \rangle$ can be interpreted: "All undergraduate attractive women, who want kids, wants undergraduate men, who want kids."
- If we obtain factor $\langle F_W^i, F_M^j \rangle$ by wide approach, we can interpret the relation between F_W^i and F_j^M: "women who belong to F_W^i, like something about the men from F_M^j". For example $\langle F_W^2, F_M^1 \rangle$ can be interpreted: "All athlete woman, who want kids, like undergraduate men or man, who want kids."
- If we get $\langle F_W^i, F_M^j \rangle$ by α-approach with value α, we interpret the relation between F_W^i and F_M^j as: "women from F_W^i, like men from F_M^j enough", where α determines measurement of tolerance.

Remark

- Not all factors from data tables C_W or C_M must be present in any relational factor.
- In this case, we can add these simple factors to the set of relational factors and consider two types of factors. This factors are not pair factors, but classical factors from C_W or C_M. Of course this depends on a particular application.

Another Approach

- Simpler approach to multi-relational data factorization is such, that we do factorization of the relation $\mathcal{R}_{C_1C_2}$. This is correct because we can imagine the relation between data tables C_1 and C_2 as another data table.
- For each factor, we take the extent of this factor and compute concept in C_1, which contains this extent. Similarly for intents of factors and concepts in C_2. For example one of the factors of $\mathcal{R}_{C_WC_M}$ from example is:

  $$\langle\{\text{Becky, Daphne}\}, \{\textit{athlete, wants kids}\}\rangle.$$

  Relational factor computed from this factor will be

  $$\Big\langle\langle\{\text{Becky, Daphne}\}, \{\textit{athlete, wants kids}\}\rangle,\\ \langle\{\text{Carl}\}, \{\textit{athlete, undergraduate, wants kids}\}\rangle\Big\rangle.$$

- This approach seems to be better in terms of that we get pair of concepts for every factors, but we do not get an exact decomposition of data tables C_1 and C_2. Moreover this approach can not be extended to n-ary relations.

n-tuple relational factors

- Above approaches (Narrow, Wide, α-approach) can be generalized for more than two data tables.
- In this generalization, we do not get factor pairs, but generally factor n-tuples.

Definition

Relation factor on data tables C_1, $C_2, \ldots C_n$ is a n-tuple $\left\langle F_1^{i_1}, F_2^{i_2}, \ldots F_n^{i_n} \right\rangle$, where $F_j^{i_j} \in \mathcal{F}_j$ where $j \in \{1, \ldots, n\}$ ($\mathcal{F}_j$ denotes set of factors of data table C_j) and satisfying relations $\mathcal{R}_{C_l C_{l+1}}$ or $\mathcal{R}_{C_{l+1} C_l}$ for $l \in \{1, \ldots, n-1\}$.

Example

- Data table C_P represents people and their characteristic, C_R represents restaurants and their characteristics and C_C represents which ingredients are included in national cuisines.
- Relation $\mathcal{R}_{C_P C_C}$ represents relationship "person likes ingredients" and relation $\mathcal{R}_{C_R C_C}$ represents relationship "restaurant cooks national cuisine".
- One of the relational factors, which we get by 0.5-approach, is $\langle F_P^1, F_C^{11}, F_R^3 \rangle$ and could be interpreted as "men would enjoy eating in luxury restaurants where the meals are cheap". Another factor is $\langle F_P^3, F_C^2, F_R^1 \rangle$ and could be interpreted as "women enjoy eating in ordinal cheap restaurants".
- We can represent the relational factors via graph (n-partite).

Conclusion

- In this work we present the new approach to BMF of multi-relational data, i.e. data which are composed from many data tables and relations between them.
- This approach, as opposed from to BMF, takes into account the relations and uses these relations to connect factors from individual data tables into one complex factor, which delivers more information than the simple factors.

Future Research

- Generalization multi-relational Boolean factorization for ordinal data, especially data over residuated lattices
- Design an effective algorithm for computing relational factors.
- Develop new approaches for connecting factors which utilize statistical methods and last but not least drive factor selection in the second data table
- Using information about factors in the first one and relation between them, for obtaining more relevant data

Thank you.