M. Krmelova, M. Trnecka: Boolean Factor Analysis of Multi-Relational Data

Department of Computer Science, Palacky University, Olomouc (17. listopadu 12, CZ–77146 Olomouc, Czech Republic)

Introduction

The Boolean factor analysis is an established method for analysis and preprocessing of Boolean data. In the basic setting, this method is designed for finding factors, new variables, which may explain or describe the original input data. Many real-world data sets are more complex than a simple data table. For example almost every web database is composed from many data tables and relations between them. We present a new approach to the Boolean factor analysis, which is tailored for multi-relational data.

Our Basic 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.

BFA of Multi-Relation Data

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 $F_i^1 \in \mathcal{F}_1$ and $F_2^j \in \mathcal{F}_2$ ($\mathcal{F}_i$ denotes set of factors of data table C_i) and satisfying relation $\mathcal{R}_{C_1C_2}$.

Meaning of Satisfying Relation

- There are several ways how to define the meaning of "satisfying relation".
- We propose three definition of this meaning.
- This definition holds for an object-attribute relation, other types of relations can be defined in similar way.

Definition of "satisfying relation"

- 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.

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

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

- 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:

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

Simple Example

Data tables C_W represents women and their characteristics and C_M represents men and their characteristics. $\mathcal{R}_{C_WC_M}$ represent relation with meaning "woman looking for a man with the characteristics".

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

C_W

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

C_M

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

$\mathcal{R}_{C_WC_M}$

Data Analysis

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:

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

We have two sets of factors (formal concepts), first set $\mathcal{F}_W = \{F_W^1, F_W^2, F_W^3\}$ factorising data table C_W and $\mathcal{F}_M = \{F_M^1, F_M^2, F_M^3\}$ factorising data table C_M.

We use so far unused relation $\mathcal{R}_{C_WC_M}$, between C_W and C_M to joint factors of C_W with factors of C_M into relational factors. 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:

	F_M^1	F_M^2	F_M^3	F_M^4
F_W^1	×			
F_W^2				
F_W^3	×			

Narrow approach

	F_M^1	F_M^2	F_M^3	F_M^4
F_W^1	×			×
F_W^2	×	×	×	×
F_W^3	×			

Wide approach

	F_M^1	F_M^2	F_M^3	F_M^4
F_W^1	×			
F_W^2		×		
F_W^3	×			

0.6-approach

	F_M^1	F_M^2	F_M^3	F_M^4
F_W^1	×			×
F_W^2		×		
F_W^3	×			

0.5-approach

Interpretation of Results

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. We can interpret factor $\langle F_W^1, F_M^1 \rangle$ from our example, 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. We can interpret factor $\langle F_W^1, F_M^1 \rangle$ from our example, 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. 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. 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.

Generalization

- Our approaches can be generalized for more than two data tables. In this generalization, we do not get factor pairs, but generally factor n-tuples.
- 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 $\mathcal{R}_{C_lC_{l+1}}$ or $\mathcal{R}_{C_{l+1}C_l}$ for $l \in \{1, \ldots, n-1\}$.

Conclusion

We present the new approach to BMF of multi-relational data. This approach 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.

Acknowledgment We acknowledge support by the Operational Program Education for Competitiveness Project No. CZ.1.07/2.3.00/20.0060 co-financed by the European Social Fund and Czech Ministry of Education.

For more details see the full paper.

International Center for Information and Uncertainty
SSIU 2014: International Spring School and Workshop "Information and Uncertainty"
http://mcin.upol.cz/SSWIU-2014/

european social fund in the czech republic
EUROPEAN UNION
MINISTRY OF EDUCATION, YOUTH AND SPORTS
OP Education for Competitiveness
INVESTMENTS IN EDUCATION DEVELOPMENT