An Algorithm for the Multi-Relational Boolean Factor Analysis based on Essential Elements

Martin Trnecka, Marketa Trneckova

DEPARTMENT OF COMPUTER SCIENCE
PALACKÝ UNIVERSITY, OLOMOUC

CLA: Concept Lattices and Their Applications
Košice, Slovakia, October 7-10, 2014

Introduction

- 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 Factor analysis, in classic settings, can handle only one input data table.
- Many real-word data sets are more complex than one simple data table.
- Multi-Relational Data = data composed from many tables interconnected via relations between objects or attributes of these data tables.
- Our goal: propose an algorithm form Multi-Relation Boolean Factor Analysis.

Previous Work

- Krmelova M., Trnecka M.: Boolean Factor Analysis of Multi-Relational Data. In: M. Ojeda-Aciego, J. Outrata (Eds.): CLA 2013: Proceedings of the 10th International Conference on Concept Lattices and Their Applications, 2013, pp. 187-198.
- Problem settings: We have two Boolean data tables C_1 and C_2, which are interconnected with relation $\mathcal{R}_{C_1 C_2}$.
- Relation is over objects of first data table C_1 and attributes of second data table C_2, i.e. it is an objects-attributes relation.
- Notion of Multi-Relational Factor, i.e. pair of classic factors from data tables.
- Algorithm for computing Multi-Relational factors is missing!

Satisfyng Relation

- In previous work were introduced three approaches:
  - Narrow approach
  - Wide approach
  - α-approach
- We use the most natural approach $=$ narrow approach.
- Idea of the narrow approach: we connect two factors $F_i^{C_1}$ and $F_j^{C_2}$ if the non-empty set of attributes (if such exist), which are common (in the relation $\mathcal{R}_{C_1C_2}$) to all objects from the first factor $F_i^{C_1}$, is the subset of attributes of the second factor $F_j^{C_2}$.

Naive Algorithm

Table: C_1

	a	b	c	d
1		×	×	×
2	×		×	
3		×		×
4	×	×	×	×

Table: C_2

	e	f	g	h
5	×			×
6		×	×	
7	×	×	×	
8			×	×

Table: $\mathcal{R}_{C_1C_2}$

	e	f	g	h
1		×	×	
2	×		×	
3	×	×		×
4	×	×	×	×

- Factors of data table C_1 are: $F_1^{C_1} = \langle\{1, 4\}, \{b, c, d\}\rangle$, $F_2^{C_1} = \langle\{2, 4\}, \{a, c\}\rangle$, $F_3^{C_1} = \langle\{1, 3, 4\}, \{b, d\}\rangle$ and factors of table C_2 are: $F_1^{C_2} = \langle\{6, 7\}, \{f, g\}\rangle$, $F_2^{C_2} = \langle\{5\}, \{e, h\}\rangle$, $F_3^{C_2} = \langle\{5, 7\}, \{e\}\rangle$, $F_4^{C_2} = \langle\{8\}, \{g, h\}\rangle$.
- These factors can be connected in to two multi-relational factors $\langle F_1^{C_1}, F_1^{C_2}\rangle$ and $\langle F_3^{C_1}, F_1^{C_2}\rangle$.
- Usually is problematic to connect all factors from each data table = small number of connections between them.
- This leads to poor quality multi-relational factors.

Essential Elements

- Notion of the Essential Elements was introduce in: Belohlavek R., Trnecka M.: From-Below Approximations in Boolean Matrix Factorization: Geometry and New Algorithm. `http://arxiv.org/abs/1306.4905`, 2013.
- Essential elements in the Boolean data table are entries in this data table which are sufficient for covering the whole data table by factors (concepts).
- If we take factors which cover all these entries, we automatically cover all entries of the input data table.
- Formally, essential elements in the data table $\langle X, Y, C \rangle$ are defined via minimal intervals in the concept lattice. The entry C_{ij} is essential iff interval bounded by formal concepts $\langle i^{\uparrow\downarrow}, i^{\uparrow} \rangle$ and $\langle j^{\downarrow}, j^{\downarrow\uparrow} \rangle$ is non-empty and minimal w.r.t. $\subseteq$ (if it is not contained in any other interval).
- If the table entry C_{ij} is essential, then interval $\mathcal{I}_{ij}$ represents the set of all formal concepts (factors) which cover this entry.
- It is sufficient take only one arbitrary concept from each interval to create exact Boolean decomposition of $\langle X, Y, C \rangle$.
- Essential part of input data table can easily be constructed.

Idea of Algorithm

Table: C_1

	a	b	c	d
1		×	×	×
2	×		×	
3		×		×
4	×	×	×	×

Table: $Ess(C_1)$

	a	b	c	d
1			×	
2	×			
3		×		×
4				

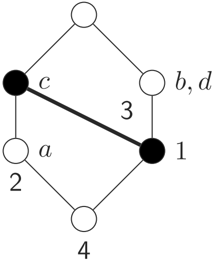
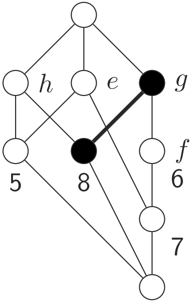


Table: C_2

	e	f	g	h
5	×			×
6		×	×	
7	×	×	×	
8			×	×

Table: $Ess(C_2)$

	e	f	g	h
5	×			×
6		×		
7	×			
8			×	×



Idea of Algorithm

- If we take highlighted intervals, we obtain possibly four connections.
- First highlighted interval contains two concepts $c_1 = \langle\{1,2,4\},\{c\}\rangle$ and $c_2 = \langle\{1,4\},\{b,c,d\}\rangle$. Second consist of concepts $d_1 = \langle\{6,7,8\},\{g\}\rangle$ and $d_2 = \langle\{8\},\{g,h\}\rangle$. Only two connections ($c_1$ with d_1 and c_1 with d_2) satisfy relation $\mathcal{R}_{C_1C_2}$, i.e. can be connected.
- Search space reduction: for two intervals it is not necessary to try all combination of factors. If we are not able to connect concept $\langle A,B\rangle$ from the first interval with concept $\langle C,D\rangle$ from the second interval, we are not able connect $\langle A,B\rangle$ with any concept $\langle E,F\rangle$ from the second interval, where $\langle C,D\rangle \subseteq \langle E,F\rangle$. Also if we are not able to connect concept $\langle A,B\rangle$ from the first interval with concept $\langle E,F\rangle$ from the second interval, we are not able connect any concept $\langle C,D\rangle$ from the first interval, where $\langle C,D\rangle \subseteq \langle A,B\rangle$, with concept $\langle E,F\rangle$.

- Search in intervals is still time consuming.
- Heuristic: take attribute concepts in intervals of the second data table (bottom elements in each interval). In intervals of the first data table take greatest concepts which can be connected via relation (set of common attributes in relation is non-empty).
- The idea behind this heuristic: a bigger set of objects possibly have a smaller set of common attributes in a relation = bigger probability to connect this factor with some factor from the second data table.
- Applying this heuristic on data from the example, we obtain three factors in the first data table, $F_1^{C_1} = \langle\{2,4\},\{a,c\}\rangle$, $F_2^{C_1} = \langle\{1,3,4\},\{c,d\}\rangle$, $F_3^{C_1} = \langle\{1,2,4\},\{c\}\rangle$ and four factors $F_1^{C_2} = \langle\{5\},\{e,h\}\rangle$, $F_2^{C_2} = \langle\{6,7\},\{f,g\}\rangle$, $F_3^{C_2} = \langle\{7\},\{e,f,g\}\rangle$, $F_4^{C_2} = \langle\{8\},\{g,h\}\rangle$ from the second one. Between this factors, there are six connections satisfying the relation.

	$F_1^{C_2}$	$F_2^{C_2}$	$F_3^{C_2}$	$F_4^{C_2}$
$F_1^{C_1}$			×	
$F_2^{C_1}$		×	×	
$F_3^{C_1}$		×	×	×

Final Algorithm for MBMF

Input: Boolean matrices C_1, C_2 and relation $R_{C_1C_2}$ between them and $p \in [0,1]$
Output: set $\mathcal{M}$ of multi-relational factors

```
$E_{C_1} \leftarrow Ess(C_1)$
$E_{C_2} \leftarrow Ess(C_2)$
$U_{C_1} \leftarrow C_1$
$U_{C_2} \leftarrow C_2$
while $(|U_{C_1}| + |U_{C_2}|)/(|C_1| + |C_2|) \geq p$ do
    foreach essential element $(E_{C_1})_{ij}$ do
        compute the best candidate $\langle a, b\rangle$ from interval $\mathcal{I}_{ij}$
    end
    $\langle A, B\rangle \leftarrow$ select one from set of candidates which maximize cover of $C_1$
    select non-empty row $i$ in $E_{C_2}$ for which is $A^{\uparrow_{R_{C_1C_2}}} \subseteq (C_2)_{i\_}^{\downarrow\uparrow_{C_2}}$ and which maximize cover of $C_1$ and $C_2$
    $\langle C, D\rangle \leftarrow \langle (C_2)_{i\_}^{\uparrow\downarrow_{C_2}}, (C_2)_{i\_}^{\uparrow_{C_2}}\rangle$
    if value of cover function for $C_1$ and $C_2$ is equal to zero then
        break
    end
    add $\langle\langle A, B\rangle, \langle C, D\rangle\rangle$ to $\mathcal{M}$
    set $(U_{C_1})_{ij} = 0$ where $i \in A$ and $j \in B$
    set $(U_{C_1})_{ij} = 0$ where $i \in C$ and $j \in D$
end
return $\mathcal{F}$
```

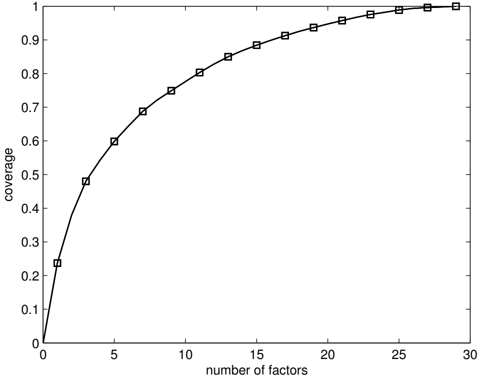
Remarks

- In each step we connect factors, which cover the biggest part of still uncovered part of data tables C_1 and C_2.
- Firstly, we obtain multi-relational factor $\langle F_2^{C_1}, F_2^{C_2} \rangle$ which covers 50 percent of the data. Then we obtain factor $\langle F_3^{C_1}, F_4^{C_2} \rangle$ which covers together with first factor 75 percent of the data and last we obtain factor $\langle F_1^{C_1}, F_3^{C_2} \rangle$.
- All these factors cover 90 percent of the data. By adding other factors we do not obtain better coverage of input data. These three factors cover the same part of input data as six connections from previous table.
- Multi-relational factors are not always able to explain the whole data. This is due to nature of data. Simply there is no information how to connect some classic factors, e.g. in the example no set of objects from C_1 has in $\mathcal{R}_{C_1 C_2}$ a set of common attributes equal to $\{e, h\}$ (or only $\{e\}$ or only $\{h\}$). From this reason we are not able to connect any factor from C_1 with factor $F_1^{C_2}$.

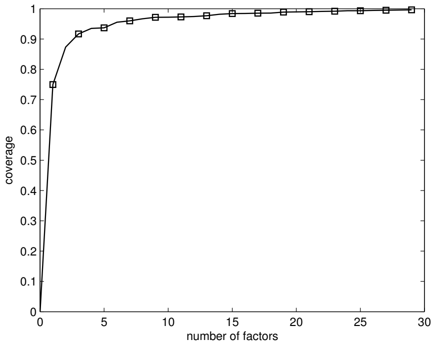
MovieLens Dataset

- `http://grouplens.org/datasets/movielens/`
- Two data tables that represent a set of users and their attributes (e.g. gender, age, sex, occupation) and a set of movies and their attributes (e.g. genre).
- Relation between data tables (contains 1000209 anonymous ratings of 3952 movies made by 6040 MovieLens users who joined to MovieLens in 2000).
- Each user has at least 20 ratings.
- Ratings are made on a 5-star scale (values 1-5, 1 means, that user does not like a movie and 5 means that he likes a movie).
- We convert the ordinal relation in to binary one and we make restriction to 3000 users (users, who rate movies the most).
- We use three different scaling:
  - User rates a movie.
  - User does not like a movie (he rates movie with 1-2 stars).
  - User likes a movie (rates 4-5).

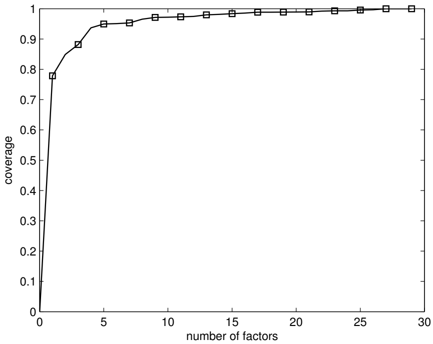
Cumulative Coverage of Input ("User rates a movie")



Coverage of Input Data Tables (“User rates a movie")



(a) Coverage of Users data table



(b) Coverage of Movies data table

Results

The most important factors are:

- Males rate new movies (movies from 1991 to 2000).
- Young adult users (ages 25-34) rate drama movies.
- Females rate comedy movies.
- Youth users (18-24) rate action movies.

Another interesting factors are:

- Old users (from category 56+) rate movies from their childhood (movies from 1941 to 1950).
- Users in age range 50-55 rate children's movies. Users in this age usually have grand children.
- K-12 students rate animation movies.

Reconstruction Error

- In case of MovieLens we are able to reconstruct input data tables almost wholly for each three relations.
- Q: Can we reconstruct relation between data tables?
- A: Yes, we can.
- Multi-relational factor carry also information about the relation between data tables. So we can reconstruct it, but with some error. This error is a result of choosing the narrow approach.
- Reconstruction error of relation is interesting information and can be minimize if we take this error into account in phase of computing coverage.
- In other words we want maximal coverage with minimal relation reconstruction error.

Conclusion and Future Research

- We present new algorithm for multi-relational Boolean matrix factorization.
- The most important factors (factors which explain the biggest portion of data) are computed first.
- Algorithm is applicable for usually large data.

A future research shall include the following topics:

- Generalization of the algorithm for ordinal data,
- Construction of algorithm which takes into account reconstruction error of the relation between data tables.
- Test the potential of this method in recommendation systems.
- Create not crisp operator for connecting classic factors into multi-relational factors.

Thank you