Pattern Mining and Machine Learning for Demographic Sequences

Dmitry I. Ignatov[1], Ekaterina Mitrofanova[2], Anna Muratova[1], and Danil Gizdatulin[1]

[1]Computer Science Faculty & [2]Institute of Demography
National Research University Higher School of Economics, Moscow, Russia

KESW 2015, Moscow

Outline

- Problem Statement
- Demographic Data
- Methods and Results
    - Decision Trees
        - First lifecourse event prediction
        - Next lifecourse event prediction
        - Rules for ‘gender’ attribute
    - Frequent Patterns
        - Frequent closed sequences
        - Emerging sequences for men and women
- Conclusion and Future Work

Problem Statement

- Analysis of demographic data by means of machine learning and data mining [Blockeel et al, 2001; Billari et al., 2006]

- Demographers questions:
  - What are the differences between demographic behaviour of men and women?
  - What are the typical (frequent) event sequences that appear in lifecourse trajectories?
  - What are the first and the next starting events a particular person may have?
  - And many more…

- Simple DM & ML tools preferably with GUI:
  - Orange http://orange.biolab.si/
  - SPMF http://www.philippe-fournier-viger.com/spmf/
  - Ad-hoc scripts in Python

Demographic Data

The survey «Parents and children, men and women in family and society»
www.socpol.ru/gender/RIDMIZ.shtml

- 4857 people including 1545 men 3312 women
- 11 generations are split in 5 years intervals from 1930 till 1984

gender	education type	locality	religion	how_often	generation	1_event
f	general	town	yes	sev_a_year	9	marriage, sep_par
f	professional	town	yes	sev_a_year	10	work
f	higher	town	yes	sev_a_year	3	work
m	professional	town	yes	never	9	education
f	professional	town	yes	min_once_a_month	3	work, education
m	higher	town	yes	never	7	sep
m	general	town	yes	never	2	work, education
m	higher	town	yes	never	8	sep
f	general	town	yes	min_once_a_month	1	education
m	professional	town	yes	never	8	education
f	professional	town	yes	never	7	education
f	professional	town	yes	sev_a_year	6	education
m	professional	town	yes	never	8	education
m	professional	town	yes	never	4	education
f	general	town	yes	once_a_week	3	education
f	general	town	yes	min_once_a_month	4	education
f	higher	town	yes	sev_a_year	3	work

Comparison of classifiers for the first event prediction

Classifier	Classification Accuracy	F_1	Precision	Recall
First child				
Classification Tree	0.42	n/a	n/a	0.0
kNN	0.39	n/a	0.0	0.0
SVM	0.42	n/a	n/a	0.0
First education				
Classification Tree	–	0.42	0.44	0.39
kNN	–	0.4	0.40	0.40
SVM	–	0.42	0.45	0.39
First marriage				
Classification Tree	–	n/a	0.0	0.0
kNN	–	0.08	0.12	0.06
SVM	–	n/a	n/a	0.0
First partner				
Classification Tree	–	n/a	0.0	0.0
kNN	–	0.10	0.16	0.07
SVM	–	n/a	n/a	0.0
Separation from parents				
Classification Tree	–	0.47	0.41	0.53
kNN	–	0.42	0.41	0.44
SVM	–	0.50	0.40	0.64
First job				
Classification Tree	–	0.45	0.44	0.47
kNN	–	0.42	0.41	0.43
SVM	–	0.40	0.45	0.36

Why Decision Trees?

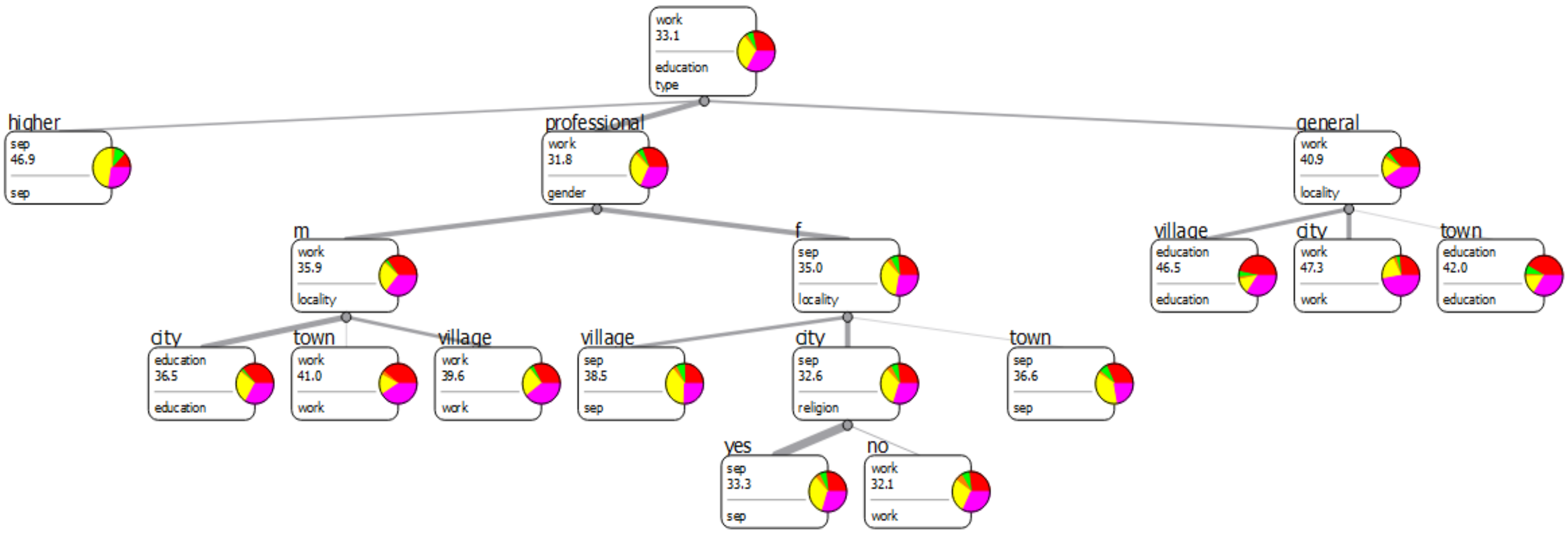
- Not a black-box approach
- Simple if-then rule based representation
- However…

There is a peculiarity of the method. Consider two if-then rules from a decision tree $\mathcal{T}$ in binary classification task with two classes $\{+,-\}$:

$$r_1 : a_1 = value_1, a_2 = value_2, \ldots, a_n = value_n \rightarrow class = +$$

$$r_2 : a_1 = value_1, a_2 = value_2, \ldots, a_n = value_n^* \rightarrow class = -$$

The first event prediction by Decision Trees



The tree is built in **Orange**

The first event mainly depends on education type

Feature Encoding Schemes

General attributes:

- Gender
- Generation
- Type of education
- Locality
- Religious views
- Religious activity

Events encoding:

- Binary encoding (0 - absence, 1 - presence)
- Time-base encoding (age in months)
- Pairs of events ordered by precedence relations (<, >, =, n/a)

The anonymised datasets for each experiment are freely available in CSV files: http://bit.ly/KESW2015seqdem

Influence of feature encoding on the next event prediction

Encoding type	Classification Accuracy	
	Imbalanced data	Balanced data
Binary (BE)	0.8498	**0.8780 (*)**
Time-based (TE)	0.3516	**0.3591**
Pairs of events (PE)	**0.7076**	0.7013
BE+ TE	0.7293 (~)	**0.7459**
BE + PE	0.8407	**0.8438**
TE + PE	**0.5465**	0.4959
BE + TE + PE	0.7295 (~)	**0.7503**

(*) means the best result, (~) means almost equivalent results

The confusion matrix for the next event prediction

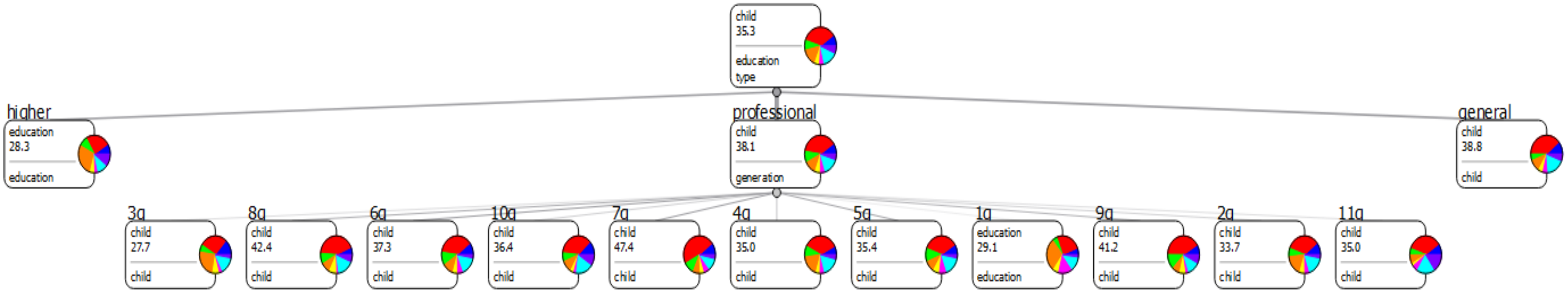
	br	child	div	education	marriage	partner	sep	work	
br	583	63	0	1	17	0	7	2	**673**
child	11	2371	0	7	0	6	42	3	**2440**
div	142	53	397	0	0	0	8	1	**601**
education	0	0	0	1041	0	0	0	10	**1051**
marriage	59	79	0	2	177	1	36	1	**355**
partner	0	42	101	0	26	142	14	2	**327**
sep	0	28	0	8	0	0	975	5	**1016**
work	0	19	0	34	0	0	12	375	**440**
	795	**2655**	**498**	**1093**	**220**	**149**	**1094**	**399**	**6903**

– binary encoding for balanced dataset
– "br" means "break up" and "div" means "divorce" events

Examples of rules for the next event prediction

Premise (path in the tree)	Conclusion (leaf)	Confidence
Education and child birth	Separation from parents	93.9%
Education, separation from parents, child birth	First job	98.9%
Male, child birth, education, partner, separation from parents, and first job	Marriage	83.2%
Female, child birth, education, partner, separation from parents, and first job	Break-up	54.6%
Child birth, education, marriage, partner, separation from parents, and first job	Break-up	78.1%
First job, separation from parents, education, marriage, and child birth	Divorce	72.9%
Female, education, separation from parents, and first job	Child birth	78.1%
Education, separation from parents, marriage	Child birth	95.7%
Education (general or professional), partner, separation from parents, and first job	Child birth	60.5% or 54.5% resp.
Education	First job	90.3%
Education and First job	Separation from parents	76.7%

The next event prediction based on general attributes



A decision tree diagram built only for general attributes.
The next event is influenced by the type of education.

Classification accuracy of different encoding schemes for gender prediction

Encoding scheme	Unbalanced data	Balanced data
	Classification Accuracy	Classification Accuracy
Binary	**0.6838(*)**	0.5824
Time-based	**0.6827**	0.6758
Pairwise	**0.6817**	0.5896
Binary and time-based	**0.6842(~)**	0.6647
Binary and pairwise	**0.6815**	0.5923
Time-based and pairwise	**0.6827**	0.6743
BE, TE, and PE	0.6842(~)	**0.6915(*)**

(~) means very close results, and (*) means the best result in the column.

Examples of rules for prediction of target attribute «gender»

Men:

Antecedent	Confidence
First job after 19.9 years, marriage in 20.6-22.4, education before 20.7, break up after 27.6, divorce before 30.5	65.9%
First job after 19.9, marriage in 20.6-22.4, break-up before 27.6	61.1%
First job before 17.2, marriage in 20.6-22.4, break-up before 27.6	61.3%
First job after 21, marriage after 29.5	70.2%

Women:

Antecedent	Confidence
First job in 18.2-19.9, marriage in 20.6-22.4, break-up after 27.6, divorce after 30.5	71.9%
First job in 18.2-19.9, marriage in 20.6-22.4, break-up after 27.6, divorce before 30.5	70.9%
First job in 17.2-19.9, marriage in 20.6-22.4, break-up before 27.6	62.8%
First job in 17.7-21, marriage after 29.5	62.8%

Sequence Mining

[Zaki & Meira, 2014; Agrawal & Srikant, 1995]

- **A sequence** is an ordered set of elements (events)

$$\langle e_1 e_2 e_3 \ldots e_l \rangle$$

- **The support** of sequence *r* in database $D = \{s_1, s_2, \ldots, s_N\}$ is the number of sequences in *D* that contain *r*

$$\sup(r) = \#\{s_i \in D \mid r \textit{ is subsequence of } s_i\}$$

- **The relative support of** *r* is the fraction of sequences that contain *r*

$$rsup(r) = \frac{\sup(r)}{N}$$

- **A sequence** *r* is **frequent** in a database *D if* $\sup(r) \geq minsup$, where *minsup* is a minimal support threshold

- **A frequent closed sequence** is a sequence such that there is no any its supersequence with the same support

Emerging Sequences

- **An emerging sequence** is a sequence such that its support growths drastically in the transition from one database to another one

- **The growth rate** of a sequence *s* in the transition from databases D_1 to D_2:

$$GrowthRate(s) = \begin{cases} 0, \text{ if } sup_1(s) = 0 \text{ and } sup_2(s) = 0 \\ \infty, \text{ if } sup_1(s) = 0 \text{ and } sup_2(s) \neq 0 \\ \frac{sup_2(s)}{sup_1(s)}, \text{ otherwise} \end{cases} \qquad (1)$$

- **The contribution** of a sequence to a particular class:

$$score(s, C_i) = \sum_{e \sqsubseteq s} \frac{GrowthRate(e)}{GrowthRate(e) + 1} \cdot sup_i(e), \qquad (2)$$

where *e* is a subsequence of *s*

The idea is based on [Mill, 1843; Finn, 1983; Dong & Li, 1999]

Frequent closed sequences

- An SPMF output for frequent closed sequences mining

gender	educati	generat	locality	religion	how_ofte	event1	event2	event3	event4	suppor
						education				4857
						work				4812
						sep_from_par				4723
						child				4399
						marriage				4201
				yes		education				4024
				yes		work				3985
				yes		sep_from_par				3908
						work	child			3828
				yes		child				3646
						marriage	child			3568
				yes		marriage				3494
						work	marriage	child		2762
				yes		work	marriage	child		2296
						education	marriage	child		2183
f						work	marriage	child		1819
				yes		education	marriage	child		1818
						sep_from_par	marriage	child		1800
						education	work	marriage	child	1091
				yes		education	work	marriage	child	906
						sep_from_par	work	marriage	child	822
f						education	work	marriage	child	717

Emerging Sequence Mining

- Implemented in Python 2.7

Input: two datasets for men and women with age indication for demographic events

1. Transformation of events to sequences (80:20 test-to-training ratio)
2. Passing the test set to SPMF and finding frequent sequences.
3. Finding emerging sequences (classification rules) and their contributions to classes.
4. Defining the class of a rule by its contribution and then contribution normalisation.
5. Accuracy calculation for the test set.

Output: files with classification rules for men and women

- minsup = 0.005; for each class we use 3312 sequences after oversampling.
- The best classification accuracy (0.936) has been reached at minimal growth rate 1.0, with 577 rules for men and 1164 for women, and 3 non-covered objects.

Emerging sequences

- The list of emerging sequences for **men** (with their class contribution):

$\langle\{education\},\{separation\},\{work\},\{marriage\}\rangle$, 0.0124
$\langle\{separation, education\},\{work\},\{partner\},\{children\}\rangle$, 0.0079
$\langle\{education\},\{separation\},\{work\},\{marriage\},\{children\}\rangle$, 0.0074
$\langle\{education\},\{separation\},\{partner\},\{marriage\},\{children\}\rangle$, 0.0065
$\langle\{work\},\{education\},\{marriage, partner\},\{divorce, break\text{-}up\}\rangle$, 0.0057
$\langle\{divorce, break\text{-}up\},\{children\}\rangle$, 0.0055
$\langle\{work\},\{divorce, break\text{-}up\},\{children\}\rangle$, 0.0055
$\langle\{education\},\{marriage\},\{work, children\}\rangle$, 0.005
$\langle\{partner\},\{divorce, break\text{-}up\},\{children\}\rangle$, 0.005
$\langle\{marriage\},\{divorce, break\text{-}up\},\{children\}\rangle$, 0.005
$\langle\{education\},\{partner\},\{divorce\},\{children\}\rangle$, 0.005

- The list of emerging sequences for **women:**

$\langle\{partner, education\},\{children\},\{break\text{-}up\}\rangle$, 0.0147
$\langle\{separation\},\{children\},\{work\},\{education\}\rangle$, 0.0121
$\langle\{separation, partner\},\{marriage\},\{education\}\rangle$, 0.0106
$\langle\{work, education, marriage\},\{separation\}\rangle$, 0.0102
$\langle\{work, partner, education\},\{break\text{-}up\}\rangle$, 0.0098
$\langle\{separation, partner\},\{children\},\{work\}\rangle$, 0.0092
$\langle\{partner, education\},\{marriage\},\{break\text{-}up\}\rangle$, 0.008
$\langle\{work\},\{partner, education\},\{break\text{-}up\}\rangle$, 0.008
$\langle\{work, partner, education\},\{children\},\{break\text{-}up\}\rangle$, 0.008
$\langle\{work, partner\},\{children\},\{divorce\}\rangle$, 0.008
$\langle\{separation, partner, education\},\{break\text{-}up\}\rangle$, 0.0072

Conclusion and Future Work

- We have shown that decision trees and sequence mining could become the tools of choice for demographers.

- Machine learning and data mining tools can help in finding regularities and dependencies that are hidden in voluminous demographic datasets. However, these methods need to be properly tuned and adapted to the domain needs.

- In the near future we are planning:
  - to implement emerging prefix-string mining and learning to deal with sequences without gaps
  - to use different rule-based techniques that are able to cope with unbalanced multi-class data (Cerf et al., 2013)
  - to apply Pattern Structures to demographic sequence mining (Kuznetsov et al., 2013).
  - and many more ideas and tricks…

Thank you.
Questions?