Pattern Mining and Machine Learning for Demographic Sequences

Dmitry I. Ignatov(✉), Ekaterina Mitrofanova, Anna Muratova,
and Danil Gizdatullin

National Research University Higher School of Economics, Moscow, Russia
dignatov@hse.ru
http://www.hse.ru

Abstract. In this paper, we present the results of our first studies in application of pattern mining and machine learning techniques to analysis of demographic sequences in Russia based on data of 11 generations from 1930 to 1984. The main goal is not prediction and data mining methods themselves but rather extraction of interesting patterns and knowledge acquisition from substantial datasets of demographic data. We use decision trees as techniques for demographic events prediction and emergent patterns for searching significant and potentially useful sequences.

Keywords: Demographic sequences · Sequence mining · Emergent patterns · Emergent sequences · Decision trees · Machine learning

1 Introduction and Related Work

The analysis of demographic sequences is a very popular and promising direction of study in demography[1,2]. The life courses of people consist of the chains of events in different spheres of life. Scientists are interested in the transition from the analysis of separate events and their interrelation to the analysis of the whole sequences of the events. However, this transition is slowing by the technical peculiarity of working with sequences. As of today, demographers and sociologists do not have an available and simple instrument of such analysis. Some demographers possessing programming skills are successfully making sequence analysis [3–5] and developing statistical methods [6–10], but the majority of the social scientists have the only option to cooperate with other scientists to extract knowledge from demographic data. Commonly, demographers rely on statistics, but sophisticated sequence analysis techniques only start to emerge in this field [11]. Since traditional statistical methods cannot face the emerging needs of demography, demographers start showing a great interest in techniques of computer science [12].

Human demographic behaviour can be very different varying over different generations, gender, education level, religious views etc., however, hidden similarities can be found and generalised by specially designed techniques. Even though there

P. Klinov and D. Mouromtsev (Eds.): KESW 2015, CCIS 518, pp. 1–15, 2015.
DOI: 10.1007/978-3-319-24543-0_17

are many methods developed so far, the field is far from convergence with traditional sequence mining techniques that studied in Data Mining. Machine Learning (ML) and Data Mining (DM) are rather young and rapidly developing fields that require professional knowledge of computer science, which is usually missing in social sciences.

Another positive tendency is the availability of easy to use tools from ML & DM community like Weka, Orange, SPMF etc. that do not presume expertise in programming. For those social scientists who are able to program, Python (and R as well) and many its packages as scikitlearn are ready to use.

So, one of the goals of this study is to find possible ties between areas and try to use Machine Learning and Pattern Mining to this end.

We essentially rely on two previous works [8,12] and strive to obtain similar results by means of decision tree learning implemented in Orange. However, those papers demonstrate only classification techniques as a tool of choice. The main goal of the authors was to find rules (patterns) that discern demographic behaviour of Italian and Austrian people. The classification itself was rather the mean but not a goal. Good classification results only assured us that the classifier is suitable, but the if-then rules from an obtained decision tree give us the patterns to interpret from demography viewpoint. From this point, black-box approaches like SVM and artificial neural networks do not match the task; they can be better in prediction but do not produce interpretable patterns.

Thus, the next natural avenue is pattern mining and sequence mining in particular, so we use SPMF [13] and its sequence mining techniques as a tool. To make these methods more suitable for finding significant patterns in demographic setting we adapt so called emergent patterns approach from [14].

The paper is organized as follows. In Section 2, we describe our demographic data. In section 3, we propose how to use decision trees and find interesting patterns in demographic sequences. Section 4 introduces sequence mining and emergent patterns that we combined. Experimental results are reported in two subsections of Section 5. Section 6 concludes the paper.

2 Data Description and Problem Statement

The dataset for the study is obtained from the Research and educational group for Fertility, Family formation and dissolution of HSE[1]. We use the panel of three waves of the Russian part of Generation and Gender Survey (GGS), which took place in 2004, 2007 and 2011[2]. The dataset contains records of 4857 respondents (1545 men and 3312 women). The gender imbalance of the dataset is caused by the panel nature of the data: the leaving of the survey by the respondents is an uncontrollable process. That is why the representative waves combined in a panel with the structure less close to the general sample.

[1] http://www.hse.ru/en/demo/family/

[2] This part of GGS "Parents and Children, Men and Women in Family and in Society" is an all Russia representative panel sample survey: http://www.ggp-i.org/

In the database, for each person the following information is indicated: date of birth, gender (male, female), generation, type of education (general, higher, professional), locality (city, town, village), religion (yes, no), frequency of church attendance (once a week, several times in a week, minimum once a month, several times in a year or never) and the date of significant events in their lives such as: first job experience, completion of education of the highest level, leaving the parental house, first partnership, first marriage, birth of the first child. There are eleven generations: first (those who was born in 1930–34), second (1935–39), third (1940–44), fourth (1945–49), fifth (1950–54), sixth (1955–59), seventh (1960–64), eighth (1965–69), ninth (1970–74), tenth (1975–79) and eleventh (1980–84).

There is a variety of questions that demographers would like to answer:

– What are the most typical first life-course events for different generations?
– What is the difference between men and women in terms of demographic behaviour?
– What are the non-trivial but robust patterns in life-course events which are not evident from the first glance?
– What are the prospective starting event and next (after the last surveyed) event for an individual of a certain type (e.g. youngest generations)?

There are many different variations of similar questions and all they in fact need proper means of pattern mining.

3 Why Decision Tree Learning?

First, to instantiate more focused questions from Section 2 we answer those requests that can be formulated in classification setting:

– What is the first life course event given the general person's descriptions?
– What is the next course event given the general person's descriptions and previous demographic behaviour?
– What are the typical patterns that discern men and women?

We decided to use decision trees (DT) [15,16] to fulfill these queries since DT provide us with not only predictions but with classification if-then rules. In fact, we cannot only predict the first life course event of a given individual but must to say which features in his/her profile is the reason. These if-then rules are also good discriminative patterns in analysing men and women behaviour.

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 $\{+, -\}$:

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

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

On the one hand $a_1, a_2, \dots,$ and a_n are the most discriminative attributes w.r.t. to attribute selection function for branching, but on the other hand r_1 and r_2 differ only in the value of the last node. This form of rules seems to be a restriction since objects of different classes share a majority of similar attribute values. Hence, treatment of other types of patterns is necessary as well.

4 Sequence Mining and Emergent Patterns

4.1 Frequent Sequence Mining

The frequent sequence mining problem was first introduced by Agrawal and Srikant [17] for analysis of customer purchase sequences: "Given a set of sequences, where each sequence consists of a list of events (or elements) and each event consists of a set of items, and given a user-specified minimum support threshold of *minsup*, sequential pattern mining finds all frequent subsequences, that is, the subsequences whose occurrence frequency in the set of sequences is no less than *minsup*."

We reproduce the more formal definitions from [18]. Let $I = \{i_1, i_2, \ldots, i_n\}$ be the set of all items (or atomic events). An *itemset* is a nonempty set of items. A *sequence* is an ordered list of events (itemsets). A sequence s is denoted $\langle e_1 e_2 e_3 \ldots e_l \rangle$, where event e_1 happens before e_2, which happens before e_3, etc. Event e_i is also called an element of s. The itemset (or event) is denoted $\{a_1, a_2, \ldots, a_q\}$, where a_i is an item. The brackets are omitted if an element has only one item, that is, element $\{a\}$ is written as a (an atomic event).

A particular item can occur at most once in an event of a sequence, but can occur multiple times in different events of a sequence. In case of first life-course events an item cannot occur more than once, since e.g. the event "first child birth" cannot happen again in the lifecourse of a particular person. The number of occurrences of items in a sequence is called the *length of the sequence*, i.e. an l-sequence has length l. A sequence $\alpha = \langle a_1 a_2 \ldots a_n \rangle$ is called a *subsequence* of another sequence $\beta = \langle b_1 b_2 \ldots b_m \rangle$, and β is a *supersequence* of α, denoted as $\alpha \sqsubseteq \beta$, if there exist integers $1 \leq j_1 < j_2 < \ldots < j_n \leq m$ such that $a_1 \subseteq b_{j1}$,$a_2 \subseteq a_{j2}, \ldots, a_n \subseteq b_{jn}$.

A *sequence database*, D, is a set of sequences. For the discussed domain, D contains sequences for all individuals in the demographic survey. A sequence s is said to contain a sequence α, if α is a subsequence of s. The *support of a sequence* α in a sequence database D is the number of sequences in the database containing α: $sup_D(\alpha) = \#\{s | s \in D \ \& \ \alpha \sqsubseteq s\}$. When the sequence database is clear from the context, it can be denoted as $sup(\alpha)$. Given a positive integer *minsup* (or relative *rminsup* $\in [0, 1]$) as the minimum support threshold, a sequence α is called *frequent* in sequence database D if $sup_D(\alpha) \geq minsup$. Sometimes a frequent sequence is called a *sequential pattern*.

A frequent sequence s is called *closed* for given $minsup = \theta$ and sequence set D iff there is no its supersequence with the same support. Mining closed sequential patterns results in a significantly less number of sequences than in case of the full set of sequential patterns. Moreover, the full set of frequent subsequences (with their supports) can be recovered from the closed subsequences.

4.2 Emergent Sequences

Emergent patterns [14] in data mining is simply another instantiation of John Stuart Mill's ideas on formalisation of inductive reasoning (for example, cf. the difference method in [19]). To find a hypothesis for classification of an object to be positive or negative (one can consider several classes as well), one can compare all positive examples and use their common descriptions to this end. If such a common positive description does not occur in the descriptions of negative examples, it can be called a hypothesis [20,21].

We define *emergent sequences* as frequent subsequences of sequences of a particular class, which less frequent in the sequences of the rest classes. Thus, one can find such sequences for classes men and women to reveal discriminative patterns.

Let D_1 and D_2 be two datasets and s be a sequence. For $i \in \{1,2\}$ denote support of s in dataset D_i as $sup_i(s)$. The *growth rate* of s is defined as follows:

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

A sequence s is called *emergent* (for class 2) if its GrowthRate exceeds a predefined threshold. The *sequence contribution* to class C_i is defined as follows:

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

where $e \sqsubseteq s$. In case of unbalanced classes the contribution values need to be normalised e.g. by arithmetic mean or median.

Note that in DM community there are alternative definitions of statistically significant and unexpected patterns [22].

5 Experiments and Results

5.1 Machine Learning Experiments

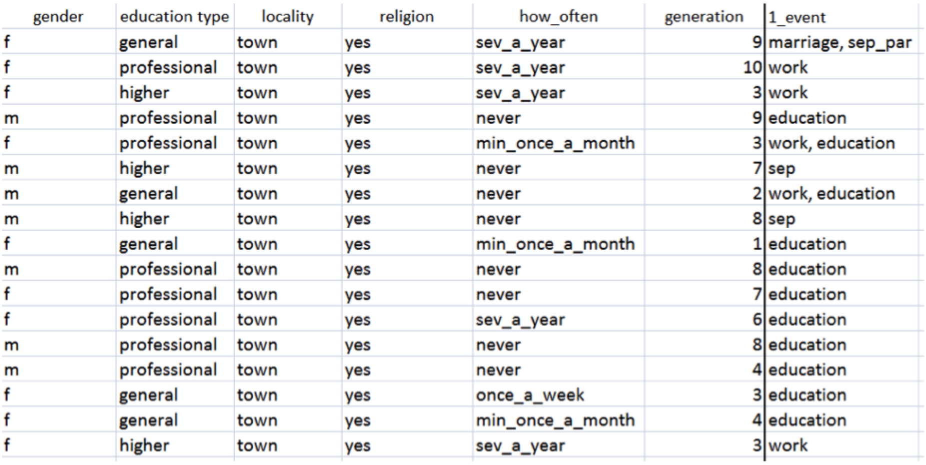
To perform classification experiments[3] we mainly use Orange [23] and WEKA [24] environments and adhoc scripts in Python.

Prediction of First Life-Course Events. QUESTION: *What is the first life course event given the general person's descriptions?*

In the first task the class attribute is "First event". In fact it can be a conjunction of several atomic events that happened together w.r.t to the time granularity, e.g. within the same month (see Table 1).

However, such events are rather rare and will impose low prediction quality. So, we untangle these events to be atomic keeping their feature description the same. That is, for an individual with class value $\{work, separation\}$ we produce two new rows with the class values *work* and *separation* respectively 2.

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



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

Fig. 1. Excerpt of data for the first event prediction task

The classification accuracy on this dataset for complex events and atomic is 0.37 and 0.41 respectively. Usually, such low value is unacceptable for machine learning task. One of the reasons is the data unbalance, i.e. $\frac{\#women}{\#men} \approx 2$. To overcome the difficulty we use SMOTE oversampling technique from WEKA package. After oversampling the number of men has risen from 1680 to 3360. However, the classification accuracy has not gained dramatically: $CA = 0.43$.

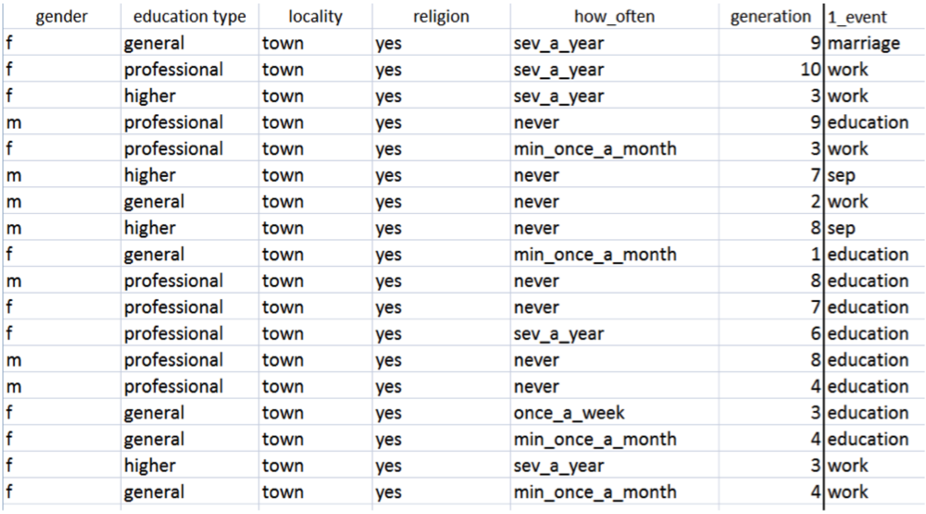
From the obtained decision tree 3 one can see that with probability 46.9 % if a person obtained higher education, then his/her first event is separation from parents.

If a person has general education and lives in a countryside or a town, then the first event is finishing of education in 46.5 % and 42.0 % cases respectively. However, if a person lives in a city, this is “first job” in 47.3 % cases.

If a man has professional education and lives in a town or in a countryside, then his first event will be “first job” in 41.0 % and 39.6 % respectively; for a man living in a city, he will complete education in 36.5 % cases. However, if a woman has professional education, her first events are different. Thus, if a woman lives in a countryside or a town, then the first event is separation from parents in 38.5 % and 36.6 % cases respectively. If she is living in a city being not a religious person, then in 32.1 % cases her first even is “first job” but, in case she is religious, she will separate from parents first in 33.3 % cases.

All the first events have not that high frequencies which means that there is no strong dependencies between individuals’ descriptions and their first event outcome. However, in case we suppose that all the events are uniformly distributed, we obtain that each event has probability $100/6 \approx 16.6\%$. So, even 30 % of cases observed in the tree leaf is a rather high result.

Note that we have chosen kNN and SVM just to see whether decision trees (DT) are not much worse than the other popular ML methods based on



gender	education type	locality	religion	how_often	generation	1_event
f	general	town	yes	sev_a_year	9	marriage
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
m	higher	town	yes	never	7	sep
m	general	town	yes	never	2	work
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
f	general	town	yes	min_once_a_month	4	work

Fig. 2. The dataset for the first event prediction task with untangled events in the class attribute

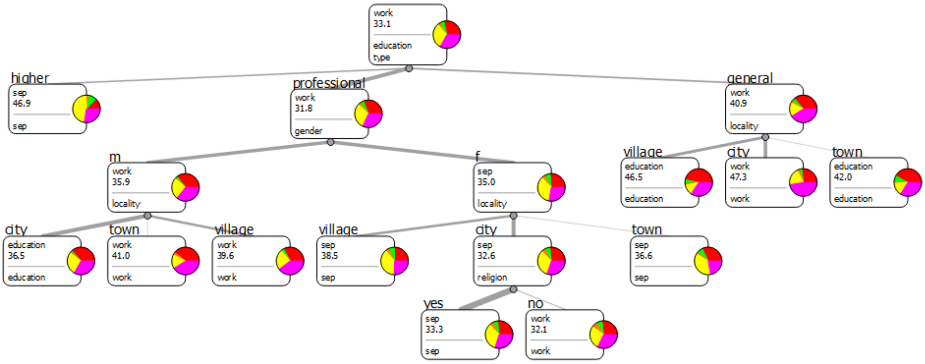


Fig. 3. The decision tree diagram for the first event prediction task

different approaches. Thus, in all our ML experiments we use 15-fold cross-validation with parameter tuning, C-SVM with RBF kernel and starting cost parameter 1.0, DT with Information Gain splitting criterion and no less 70 objects in leaves without binarisation. From the result summary one can conclude that decision trees (DT) classifier demonstrates comparable quality (Brier score 0.68) as kNN and SVM (Brier score 0.686). Moreover, this set of methods has the same events of low prediction quality: first child, first marriage, and first partner are rather unpredictable as first events at all. So, separation from parents, first job, and completion of first education are the typical first events.

Table 1. Classification performance for the first life-course 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

Prediction of Next Life-Course Events. QUESTION: *What is the next course event given the general person's descriptions and previous demographic behaviour?*

As an input for the next life-course event prediction we take the same individuals with all their general descriptors like gender, education, living place, religion, visiting frequency of religious events, generation and important demographic events with the indication of time they happened in months starting from the person's birth. The current last event that happened in the life-course of a particular person is a target attribute (class).

Here we deal with the same problem of multiple atomic events in the target variable. Therefore, we use the same untangling transformation.

We encode the events as features in three different ways:

1. BE or binary encoding (value 1 means that event has happened for a given person and 0 otherwise),
2. TE or time encoding (the age when the event happened in months),
3. PE or pairwise encoding.

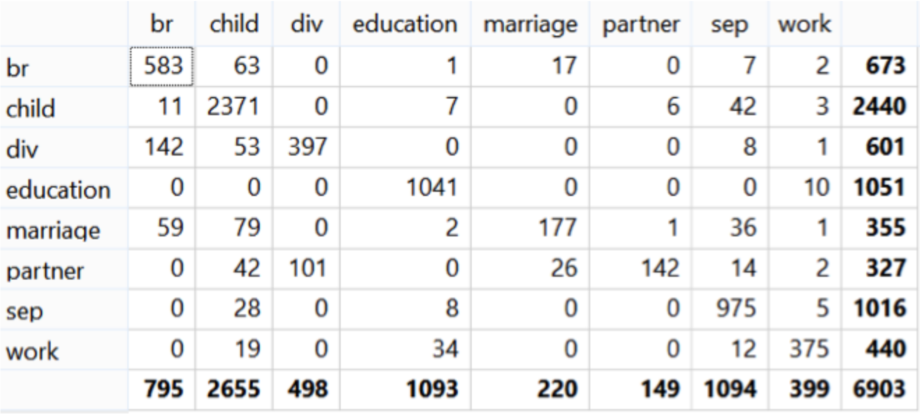
There are 7 possible combinations of these encoding schemes.

For pairwise encoding of two events a and b as values: "<" means that either a precedes b or b has not happened yet, ">" means that either a follows b or a has not occurred yet, "=" designates that a and b has happened simultaneously w.r.t. time granularity, "n/a" denotes that neither a or b has happened.

Table 2. Comparison of classification accuracy of different encoding schemes for next life-course event prediction. Bold font means the best result for given encoding, sign (∼) means very close results, and (*) means the best result in the column.

Encoding scheme	Unbalanced data	Balanced data
	Classification Accuracy	Classification Accuracy
Binary	0.8498(*)	**0.8780(*)**
Time-based	0.3516	**0.3591**
Pairwise	**0.7161**	0.7013
Binary and time-based	0.7293(∼)	**0.7459**
Binary and pairwise	0.8407	**0.8438**
Time-based and pairwise	**0.5465**	0.4959
BE, TE, and PE	0.7295(∼)	**0.7503**

From the resulting Table 2 it is easy to see that the best classification accuracy is attained via binary encoding scheme on the balanced data, 0.88, but for the unbalanced dataset the best value does not differ dramatically, 0.85. Pure time-based encoding scheme is the worst in both cases. Pairwise encoding usually helps to improve the results of other encoding schemes, but in combination with binary one it may slightly worsen the accuracy.



	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**

Fig. 4. The confusion matrix for the next event prediction ("br" means "break up" and "div" means "divorce" events) in case of binary encoding and the balanced dataset

In Fig. 4 the confusion matrix (true class vs predicted) is shown. We can see that there are more zeros outside of the main diagonal in comparison to the first event prediction task. The most confusing classes are “first divorce” (“break-up” predicted in 142 cases and “child birth” in 53 cases), first partner (“child birth” and “divorce” in 42 and 101 cases respectively) and first marriage (“break-up” and “child birth” predicted in 59 and 79 cases respectively). Since the tree does not have a domain knowledge, divorce may be very similar to the break-up outcome ceteris paribus without taking into account marriage event. The hypothesis that in some cases a father will rather prefer divorce than to care about his child being married need to be separately tested.

Since the tree is rather humongous to be put as a figure we provide several examples of the obtained rules below (Table 3).

Table 3. Several rules from the obtained decision tree for the next events' prediction task

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%

So, the obtained rules is food for thoughts to demographers, however even an ordinary data analyst may be curious why 3rd and 4th rules in Table 3 differ only by gender but result in marriage for men and break-up for women.

The obtained rules may include attribute “generation” in the premise as well. For example, if generation is 10th (1975-1979) a person has religious beliefs and professional education, then 40.9% the last event is child birth, but in case the person is atheistic then it will most likely result in break up.

Gender Prediction Rules. QUESTION: *What are the typical patterns that discern men and women?*

To find discriminative patterns from men and women, we perform similar data transformations, namely balancing and combinations of three encoding schemes. We use both general descriptions and demographic events as object attributes.

Table 4. Comparison of 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(*)**

For this task balancing does not improve predictive accuracy but rather makes it lower; however, the slightly better result than in case of binary encoding (0.69 vs 0.68) is obtained via using balanced data and all three encoding schemes (see Table 4).

However, for the best unbalanced and balanced data we have different values of precision and recall. Thus for the unbalanced data and binary enconding scheme we have $F_1 = 0.17$, Precision=0.25, and Recall=0.16 for men class and $F_1 = 0.8$, Precision=0.7, and Recall=0.93 for women. The situation is quite better for the balanced data and the full combination of encoding schemes: $F_1 = 0.7$, Precision=0.68, and Recall=0.73 for male and $F_1 = 0.68$, Precision=0.71, and Recall=0.65 for female.

Let us present several found patterns.

The predicted target class value is Men:

- 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 (confidence is 65.9%)
- First job after 19.9, marriage in 20.6-22.4, break-up before 27.6 (conf. 61.1%)
- First job before 17.2, marriage in 20.6-22.4, break-up before 27.6 (confidence 61.3%)
- First job after 21, marriage after 29.5 (confidence 70.2%)
- Child after 22.9, marriage in 23.9-29.5 (confidence 69.3%)
- Marriage in 22.4-23.3 (confidence 65.4%)
- Marriage in 23.3-23.9, child after 24.7 (confidence 67%)

The predicted target class value is Women:

- First job in 18.2-19.9, marriage in 20.6-22.4, break-up after 27.6, divorce after 30.5 (confidence 71.9%)

- First job in 18.2-19.9, marriage in 20.6-22.4, break-up after 27.6, divorce before 30.5 (confidence 70.9%)
- First job in 17.2-19.9, marriage in 20.6-22.4, break-up before 27.6 (conf. 62.8%)
- First job in 17.7-21, marriage after 29.5 (confidence 62.8%)
- Child before 22.9, marriage in 23.8-29.5 (confidence 61.2%)

Note that the patterns are rather specific; it causes low predictive ability.

5.2 Pattern Mining Experiments

Mining Frequent Closed Sequences. To find frequent closed sequences we can use any efficiently implemented algorithm; thus, we used BIDE [25] from SPMF and set the minimal support threshold to $minsup = 0.1$.

The most frequent 1-event sequence is education with $sup = 4857$, the most frequent 2-event sequence is first job, and then child birth with $sup = 3828$, the most frequent 3-event sequence is first job, then marriage, then child birth with $sup = 2762$, and the most frequent 4-event sequence is education, then first job, marriage, and child birth with support equals to 1091.

Event "Child birth" happened for 4399 out of 4857 respondents: The longest closed sequential pattern starting with "Child birth" is $\langle child\ birth, education\rangle$ with support 1154.

The longest closed sequential pattern starting with "Education" is

$\langle education, first\ job, marriage, child\ birth\rangle$ with $sup = 1091$. "Marriage" event took place for 4201 out of 4857. The longest closed sequential pattern starting with "Marriage" is $\langle marriage, child\ birth, education\rangle$, $sup = 941$.

Event "First partner" happened for 1839 out of 4857 respondents. The longest closed sequential pattern starting with "First partner" is

$\langle\ partner, marriage, child\ birth\rangle$, $sup = 676$. The longest closed sequential pattern starting with "First job" is $\langle first\ job, education, marriage, child\ birth\rangle$ with support 687.

The longest closed sequential pattern starting with "Separation from parents" ($sup = 4723$) is $\langle\ separation, first\ job, marriage, child\rangle$, $sup = 822$.

It is interesting that events "separation from parents" and "marriage" happened for 833 out of 4857 respondents. Here, the longest closed sequential pattern is $\langle\{separation\ from\ parents, marriage\}, child\ birth\rangle$, $sup = 777$.

Emergent Sequences. For experiments with emergent sequences we use PrefixSpan [26] from SPMF. Since we have two classes, men and women, we has obtained two sets of emergent sequences for relative $minsup = 0.005$; for each class we use 3312 sequences after oversampling. The best classification accuracy (0.936) has been reached via 80:20 cross-validation at minimal growth rate 1.0, with 577 rules for men and 1164 for women and 3 non-covered objects.
List of emergent sequences for women (with their class contribution):

$\langle\{partner, education\}, \{children\}, \{break\text{-}up\}\rangle$, 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

List of emergent sequences for men:

$\langle\{education\}, \{separation\}, \{work\}, \{marriage\}\rangle$, 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

6 Conclusion and Future Work

We have shown that decision trees and sequence mining could become the tools of choice for demographers. To this end we have found various patterns hoping that some of them are of interest to demographers and this is really the case for our second co-author, a professional demographer, and her colleagues.

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. For example, in frequent subsequences between their two subsequent events in a corresponding sequence from dataset D may happen several other events. Thus, our domain experts require us to use "sequences of starting events without gaps" which are just prefix strings in computer science terms. In the near future we are planning to implement emergent prefix-string mining to bridge the gap. As for decision trees, we mentioned the problem of high similarity of decision rules as paths with different terminal nodes and preceding leaves but sharing the same starting subpath. It may be a restriction for interpretation matters and we need to try different rule-based techniques as well. Thus, we need rule-based techniques that are able to cope with unbalanced multi-class data [27]. Another interesting venue is usage of Pattern Structures to sequence mining [28].

Acknowledgments. We would like to thank Jaume Baixeries, Mehdi Kaytoue, and anonymous reviewers. This study (grant no. 14-05-0054 "Studying of the dynamics of formation and development of families and fertility using data of selective surveys") was supported by the Academic Fund Program of National Research University Higher School of Economics in 2014. The first author has been supported by the Russian Foundation for Basic Research grants no. 13-07-00504 and 14-01-93960 and made a contribution within the project "Data mining based on applied ontologies and lattices of closed descriptions" supported by the Basic Research Program of the same university.

References

1. Aisenbrey, S., Fasang, A.E.: New life for old ideas: The second wave of sequence analysis bringing the course back into the life course. Sociological Methods & Research **38**(3), 420–462 (2010)
2. Billari, F.C.: Sequence analysis in demographic research. Canadian Studies in Population **28**(2), 439–458 (2001)
3. Aassve, A., Billari, F.C., Piccarreta, R.: Strings of adulthood: A sequence analysis of young british womens work-family trajectories. European Journal of Population **23**(3/4), 369–388 (2007)
4. Jackson, P.B., Berkowitz, A.: The structure of the life course: Gender and racioethnic variation in the occurrence and sequencing of role transitions. Advances in Life Course Research **9**, 55–90 (2005)
5. Worts, D., Sacker, A., McMunn, A., McDonough, P.: Individualization, opportunity and jeopardy in american womens work and family lives: A multi-state sequence analysis. Advances in Life Course Research **18**(4), 296–318 (2013)
6. Abbott, A., Tsay, A.: Sequence analysis and optimal matching methods in sociology: Review and prospect. Sociological Methods & Research (2000)
7. Billari, F., Piccarreta, R.: Analyzing demographic life courses through sequence analysis. Mathematical Population Studies **12**(2), 81–106 (2005)
8. Billari, F.C., Frnkranz, J., Prskawetz, A.: Timing, Sequencing, and Quantum of Life Course Events: A Machine Learning Approach. European Journal of Population **22**(1), 37–65 (2006)
9. Gauthier, J.A., Widmer, E.D., Bucher, P., Notredame, C.: How Much Does It Cost? Optimization of Costs in Sequence Analysis of Social Science Data. Sociological Methods & Research **38**(1), 197–231 (2009)
10. Ritschard, G., Oris, M.: Life course data in demography and social sciences: Statistical and data-mining approaches. Advances in Life Course Research **10**, 283–314 (2005)
11. Gabadinho, A., Ritschard, G., Mller, N.S., Studer, M.: Analyzing and Visualizing State Sequences in R with TraMineR. J. of Statistical Software **40**(4), 1–37 (2011)
12. Blockeel, H., Fürnkranz, J., Prskawetz, A., Billari, F.C.: Detecting temporal change in event sequences: an application to demographic data. In: Siebes, A., De Raedt, L. (eds.) PKDD 2001. LNCS (LNAI), vol. 2168, pp. 29–41. Springer, Heidelberg (2001)
13. Fournier-Viger, P., Gomariz, A., Gueniche, T., Soltani, A., Wu, C.W., Tseng, V.S.: SPMF: A Java Open-Source Pattern Mining Library. Journal of Machine Learning Research **15**, 3389–3393 (2014)
14. Dong, G., Li, J.: Efficient mining of emerging patterns: discovering trends and differences. In: Proc. of the Fifth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, KDD 1999, pp. 43–52. ACM (1999)

15. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Wadsworth (1984)
16. Quinlan, J.R.: Induction of decision trees. Machine Learning **1**(1), 81–106 (1986)
17. Agrawal, R., Srikant, R.: Mining sequential patterns. In: Proceedings of the Eleventh International Conference on Data Engineering, pp. 3–14 (1995)
18. Han, J., Kamber, M.: Data Mining: Concepts and Techniques, 2nd edn. Morgan Kaufmann (2006)
19. Mill, J.S.: A system of logic, ratonative and inductive, vol. 1. J. W. Parker, London (1843)
20. Finn, V.K.: On Machine-Oriented Formalization of Plausible Reasoning in the Style of F. BackonJ. S. Mill. Semiotika i Informatika **20**, 35–101 (1983)
21. Kuznetsov, S.O.: Learning of simple conceptual graphs from positive and negative examples. In: Żytkow, J.M., Rauch, J. (eds.) PKDD 1999. LNCS (LNAI), vol. 1704, pp. 384–391. Springer, Heidelberg (1999)
22. Low-Kam, C., Raissi, C., Kaytoue, M., Pei, J.: Mining statistically significant sequential patterns. In: IEEE 13th Int. Conf. on Data Mining, pp. 488–496 (2013)
23. Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Štajdohar, M., Umek, L., Žagar, L., Žbontar, J., Žitnik, M., Zupan, B.: Orange: Data Mining Toolbox in Python. Journal of Machine Learning Research **14**, 2349–2353 (2013)
24. Bouckaert, R.R., Frank, E., Hall, M.A., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: WEKA - Experiences with a Java Open-Source Project. Journal of Machine Learning Research **11**, 2533–2541 (2010)
25. Wang, J., Han, J.: BIDE: efficient mining of frequent closed sequences. In: Özsoyoglu, Z.M., Zdonik, S.B. (eds.) Proceedings of the 20th International Conference on Data Engineering, ICDE 2004, pp. 79–90. IEEE Computer Society (2004)
26. Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U., Hsu, M.: PrefixSpan: mining sequential patterns by prefix-projected growth. In: Proceedings of the 17th International Conference on Data Engineering, pp. 215–224 (2001)
27. Cerf, L., Gay, D., Selmaoui-Folcher, N., Crmilleux, B., Boulicaut, J.F.: Parameter-free classification in multi-class imbalanced data sets. Data & Knowledge Engineering **87**, 109–129 (2013)
28. Buzmakov, A., Egho, E., Jay, N., Kuznetsov, S.O., Napoli, A., Raïssi, C.: On projections of sequential pattern structures (with an application on care trajectories). In: 10th Int. Conf. on Concept Lattices and Their Applications, pp. 199–208 (2013)