Муратова Анна Александровна
Национальный исследовательский университет
Высшая школа экономики
Российская Федерация, Москва
amuratova@hse.ru

Гиздатуллин Данил Кутдусович
Национальный исследовательский университет
Высшая школа экономики
Российская Федерация, Москва
dgizdatullin@hse.ru

Игнатов Дмитрий Игоревич
Национальный исследовательский университет
Высшая школа экономики
Российская Федерация, Москва
dignatov@hse.ru

Митрофанова Екатерина Сергеевна
Национальный исследовательский университет
Высшая школа экономики
Российская Федерация, Москва
emitrofanova@hse.ru

# ВЫЯВЛЕНИЕ ЗНАНИЙ В ДЕМОГРАФИЧЕСКИХ ПОСЛЕДОВАТЕЛЬНОСТЯХ[1]

**Аннотация.** В данной статье обобщены результаты недавних исследований приложений анализа паттернов и машинного обучения для анализа демографических последовательностей. Главной целью является решение задач демографов, включая предсказание следующего события и извлечение интересных паттернов из существующих наборов демографических данных, которые не могут быть обработаны с помощью обычных демографических методов. Мы используем деревья решений в качестве метода

[1] Статья подготовлена в результате проведения исследования № 16-05-0011 «Разработка и апробация методик анализа демографических последовательностей» в рамках Программы «Научный фонд Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ)» в 2016 г. и в рамках государственной поддержки ведущих университетов Российской Федерации "5-100".

для предсказания демографических событий, а также эмерджентные последовательности паттернов и структуры паттернов для поиска подходящей интерпретации последовательностей. Успешный опыт использования анализа паттернов в области демографии может быть использован и в других областях анализа данных.
**Ключевые слова:** демографические последовательности; анализ последовательностей; эмерджентные паттерны; структуры паттернов; деревья решений.

Muratova Anna Alexandrovna
National Research University Higher School of Economics
Russian Federation, Moscow
amuratova@hse.ru

Gizdatullin Danil Kutdusovich
National Research University Higher School of Economics
Russian Federation, Moscow
dgizdatullin@hse.ru

Ignatov Dmitry Igorevich
National Research University Higher School of Economics
Russian Federation, Moscow
dignatov@hse.ru

Mitrofanova Ekaterina Sergeevna
National Research University Higher School of Economics
Russian Federation, Moscow
emitrofanova@hse.ru

## KNOWLEDGE DISCOVERY IN DEMOGRAPHY SEQUENCES

**Abstract.** In this paper, we summarize the results of recent studies on the application of pattern mining and machine learning to the analysis of demographic sequences. The main goal is the demonstration of demographers' needs, including next-event prediction and the extraction of interesting patterns from substantial datasets of demographic data, which cannot be handled by conventional demographic techniques. We use decision trees as a technique for demographic event prediction, and emerging sequential patterns and pattern structures for discovering relevant

interpretable sequences. The emerging problem statements and positive prospects of the usage of pattern mining in the demography domain are worth dissemination in the data mining community.
**Keywords:** demographic sequences; sequence mining; emerging patterns; pattern structures; decision trees.

**Задачи демографов и методы анализа последовательностей**

Анализ демографических последовательностей является очень популярным и перспективным направлением исследования в демографии [Aisenbrey, 2010, p. 420-462; Billary, 2006, p. 37-65]. Жизненные пути людей состоят из набора событий в различных сферах. Ученые заинтересованы в переходе от анализа отдельных событий и их взаимосвязи к анализу последовательностей из сразу нескольких событий. Однако этот переход затруднен техническими особенностями работы с последовательностями. На данный момент демографы и социологи не имеют доступного и простого инструмента для такого анализа. Некоторые демографы, владеющие навыками программирования, успешно анализируют последовательности [Aassve, 2007, p. 369-388; Braboy, 2005, p. 55-90] и разрабатывают статистические методы [Billary, 2005, p. 81-106; Billary, 2006, p. 37-65; Gauthier, 2009, p. 197-231; Ritschard, 2005, p. 283-314], однако большинство социологов имеют только возможность сотрудничать с учеными из других сфер для извлечения знаний из демографических данных. Как правило, демографы полагаются на статистику, так как методы анализа последовательностей из области интеллектуального анализа данных только начинают появляться [Gabadinho, 2011, p. 1-37]. Поскольку традиционные статистические методы не удовлетворяют возрастающим потребностям демографии, демографы начинают проявлять большой интерес к методам обнаружения знаний из области интеллектуального анализа данных [Blockeel, 2001, p. 29-41; Billary, 2006, p. 37-65].

На данный момент в статистических пакетах существуют следующие возможности анализа последовательностей:

- в программе SPSS это функция «Sequence plot», позволяющая графически изобразить индивидуальные траектории людей. Недостатком является минимальное количество возможностей по настройке данного графика и отсутствие каких-либо иных инструментов анализа последовательностей;
- в программе Stata представлен больший спектр возможностей, но они также весьма ограничены заложенными в программу функциями [Brzinsky-Fay, 2006, p. 435];

- в программе R есть пакет TraMineR [Gabadinho, 2011, p. 1-37] и ряд вспомогательных пакетов, которые предоставляют достаточно широкий спектр возможностей по анализу последовательностей. Более того, обладая навыками программирования, можно создавать собственные функции, что создает дополнительное поле возможностей для работы с последовательностями.

Несмотря на достаточно долгое сотрудничество статистиков и программистов с социологами и демографами в создании данных приложений к статистическим пакетам, существующие решения пока еще не удовлетворяют всем нуждам практиков. Авторы освоили самый перспективный из представленных вариантов – пакет R, но даже в нем еще есть зоны роста.

В частности, сложность вызывает стратификация функций. Разные функции подразумевают разные команды для того, чтобы сделать группировку по какому-то признаку (полу, поколению и др.). Чтобы сделать группировку по двум и более переменным, приходится прибегать к более трудозатратным методам (создавать дополнительные переменные, объединяющие нужные признаки). Необходимо создание более простого механизма стартификации наподобие настраиваемых таблиц в SPSS.

Следующий сюжет, требующий доработки – работа с подпоследовательностями. В пакете подпоследовательностями считаются все возможные варианты: и то, что состоит из одного элемента, и то, что состоит из нескольких; и уникальные подпоследовательности, и подпоследовательности, дублирующие уже найденные. Необходима возможность настройки пользователем того, что он в данном анализе будет считать подпоследовательностью.

Еще достаточно много частных замечаний к работе пакета, на которых мы сейчас останавливаться не будем. Те же недостатки, что были перечислены, мы постарались решить средствами майнинга данных в данной работе.

Таким образом, одной из целей данного исследования является попытка решить актуальные вопросы демографов, которые пока не способны решить существующие статистические пакеты. В своей работе мы, в основном, опираемся на две новаторские публикации [Blockeel, 2001, p. 29-41; Billary, 2006, p. 37-65]. Однако эти статьи используют только ассоциативные правила и деревья решений в качестве инструмента принятия решений. Основной целью авторов являлось нахождение правил (паттернов), которые рассматривают демографическое поведение итальянцев и австралийцев. Здесь такие подходы, как SVM и искусственные нейронные сети не соответствуют задаче; они могут давать лучшее предсказание, однако они не приводят к интерпретируемым паттернам. Таким образом, следующими естественными направлениями работы являются

анализ паттернов и анализ последовательностей. Для того, чтобы эти методы больше подходили для обнаружения значимых закономерностей в демографических данных, мы будем применять эмерджентные паттерны из [Dong, 1999, p. 43-52] и структуры последовательных паттернов [Kaytoue, 2015, p. 227-231; Buzmakov, 2016, p. 135-159].

**Проблемы и результаты**

Набор данных для исследования был получен от научно-учебной группы "Модели и методы анализа демографических последовательностей". Мы использовали панель из трех волн обследования «Родители и дети, мужчины и женщины в семье и обществе», которые проходили в 2004, 2007 и 2011 годах. База данных содержит ответы 4857 респондентов (1545 мужчин и 3312 женщин). Гендерный дисбаланс набора данных вызван панельной природой данных.

В наборе данных для каждого отдельного человека была представлена следующая информация: дата рождения, пол, поколение, уровень образования, тип местности проживания (город, поселок городского типа, сельская местность), верующий ли человек, частота посещения церкви. Также указаны даты значимых событий в их жизни таких как: первый опыт работы, дата завершения обучения (самого высокого из пройденных на данный момент уровней), отделение от родителей, первое партнерство, первое замужество/женитьба, дата рождения первого ребенка. В данных были представлены 11 разных поколений в период между 1930 – 1984.

Существует целый ряд вопросов, на которые демографы хотели бы получить ответы:

- Каковы наиболее типичные первые события для разных поколений?
- В чем отличие между мужчинами и женщинами с точки зрения демографического поведения?
- Какие существуют нетривиальные, но устойчивые паттерны событий, которые неочевидны с первого взгляда?
- Каковы предполагаемые первые события и следующие события (после последнего из опрошенных событий) для человека определенного типа (например, среди младших поколений)?

**Деревья решений**

Сначала ответим на вопросы, которые могут быть сформулированы как задачи классификации:

1. Каково первое событие в жизни человека, исходя из общего описания человека?

2. Каково следующее событие человека, исходя из общего описания человека и его предыдущего демографического поведения?

3. Каковы типичные паттерны, которые отличают мужчин от женщин?

Для нахождения ответов на эти вопросы были использованы деревья решений (ДР), так как они дают нам не только предположения, но и классификацию с правилами вида «если-то». Мы не можем просто предсказать первое событие в жизни конкретного человека, мы должны указать какие признаки в его/ее профиле являются причиной. Эти «если-то» правила также являются хорошими паттернами при анализе различий в поведении между мужчинами и женщинами.

Согласно экспериментам, ответ на первый вопрос, предсказание первого события, используя только общее описания человека, дает низкую точность классификации (СА), равную 0.43. Однако, на остальные два вопроса можно ответить довольно хорошо: значения СА составляют 0,88 и 0,69 соответственно. Подробные результаты экспериментов могут быть найдены в [Ignatov, 2015, p. 225-239].

Существует особенность в методе: проблема высокого сходства в правилах решений путей с разными листьями, но с одинаковым начальным подпутем. Это может быть ограничением при интерпретации, а значит использование других типов паттернов также необходимо.

**Анализ частых последовательностей и эмерджентных последовательностей**

Проблема анализа частых последовательностей была впервые введена в работе [Agrawal, 1995, p. 3-14] для анализа последовательностей покупок. Эмерджентные паттерны [Dong, 1999, p. 43-52] в анализе данных можно рассматривать как пример конкретизации идей Джона Стюарта Милля по формализации индуктивных рассуждений. Чтобы найти гипотезы для классификации объекта в качестве положительного или отрицательного (можно также рассмотреть несколько классов), можно сравнить все положительные примеры и использовать их общие описания для этой цели. Если такого общего положительного описания нет в описаниях отрицательных примеров, то его можно назвать гипотезой [Kuznetsov, 1999, p. 384-391]. Определим эмерджентные последовательности как частые подпоследовательности последовательностей определенного класса, которые реже встречаются в последовательностях других классов. Таким образом, можно найти такие же последовательности для классов мужчин и женщин для выявления различий в паттернах.

Для экспериментов с эмерджентными последовательностями мы использовали PrefixSpan [Pei, 2004, p. 1424-1440]. Так как у нас есть два класса, мужчины и женщины, мы получили два множества эмерджентных последовательностей для относительной минимальной поддержки, равной 0.005; для каждого класса мы использовали 3312 последовательностей после переобучения. Лучшая точность классификации (0.936) была получена с 80:20 перекрестной проверкой при минимальной скорости роста 1.0, с 577 правилами для мужчин и 1164 для женщин и 3 неиспользованными объектами.

**Замкнутые подпоследовательности без пропусков на основе префиксных подпоследовательностей**

Использование только эмерджентных последовательных паттернов недостаточно и может привести к неверным выводам. Например, в частых подпоследовательностях, между двумя последующими событиями из соответствующей последовательности в наборе данных D, может произойти несколько других событий. Таким образом, демографы хотят использовать "последовательности начальных событий без пропусков", которые мы назвали префиксными подпоследовательностями. Чтобы устранить такие пропуски мы реализовали анализ префиксных подпоследовательностей строк, основанный на структурах паттернов [Kaytoue, 2015, p. 227-231; Buzmakov, 2016, p. 135-159]. Несмотря на то, что у нас есть снижение точности классификации результирующих паттернов и ненулевая скорость без классификации, обнаруженные начальные события дают больше полезной информации о первых значимых событиях. Использование только замкнутых паттернов улучшило точность классификации на 5-10%.

**Основные определения**

**Неразрывной префиксной подпоследовательностью** последовательности $s = \langle s_1, \dots, s_k \rangle$ назовем последовательность $s' = \langle s_1', \dots, s_k' \rangle$, обозначается это как $s = s' *$, если $k' \leq k, \ \forall i \ s_i = s_i'$.

**Поддержка в неразрывных префиксных подпоследовательностях** последовательности s на наборе последовательностей T Support($s'$, T) – это количество последовательностей в наборе последовательностей T, которые имеют неразрывную префиксную подпоследовательность $s'$.

**Частота** $s'$ теперь обозначается как

$$Freq(s', T) = \frac{Support(s', T)}{n}$$

где n – количество последовательностей в наборе T.

При данной минимальной частоте $0 < minFreq \leq 1$ задача **поиска префиксных неразрывных паттернов** – это задача поиска всех неразрывных префиксных последовательностей s таких, что $Freq(s,T) \geq minFreq$. Каждая последовательность $s'$такая, что $Freq(s',T) \geq minFreq$ называется **префиксным неразрывным паттерном**.

Неразрывный префиксный паттерн p называется **замкнутым**, если не существует неразрывного префиксного паттерна d большей длины с такой же поддержкой и d = p*.

Также важно отметить такое полезное свойство, как **антимонотонность**:
для любых последовательностей s и s′ таких, что s=s′* ⇒ Support(s,D) ≤ Support(s′, D).

Теперь приведем пример, на котором проиллюстрируем определения данные выше. Пусть дан следующий набор последовательностей D.

**Таблица 1**

**Набор последовательностей**

| | |
|---|---|
| s1 | $\langle\{a\},\{b\},\{d\}\rangle$ |
| s2 | $\langle\{a\},\{b\},\{c\}\rangle$ |
| s3 | $\langle\{a,b\},\{b,c\}\rangle$ |

Наш алфавит I тогда будет равен I = {a, b, c}. Количество последовательностей n=3.
Неразрывная префиксная последовательность $\langle\{a\},\{b\}\rangle$ содержится одновременно в последовательности s1 и последовательности s2, но не в последовательности s3. Поэтому Support($\langle\{a\},\{b\}\rangle, D$) = 2. При minFreq = $\frac{2}{3}$, $\langle\{a\},\{b\}\rangle$ можно назвать неразрывным префиксным паттерном, удовлетворяющим минимальной поддержке. Множество всех частых замкнутых последовательностей будет таким: $\{\langle\{a\},\{b\}\rangle\}$.
Видим, что $\langle\{a\}\rangle$ не будет замкнутым префиксным паттерном, так как $Support(\langle\{a\}\rangle, D) = Support(\langle\{a\},\{b\}\rangle, D)$ и при этом $\langle\{a\}\rangle$ есть неразрывная префиксная подпоследовательность для $\langle\{a\},\{b\}\rangle$.

**Эмерджентные паттерны**

**Эмерджентный неразрывный префиксный паттерн –** это эмерждентный паттерн, характерный для одного класса последовательностей, но при этом не характерный для другого. Понятие было введено в статье [Dong, 1999, p. 43-52].

Эту характеристику мы можем посмотреть по отношению поддержек паттерна для разных классов. Это отношение назовем **отношение прироста** (growth rate). Тогда отношение прироста для паттерна p на двух классах ‘+’ и ‘-‘ будет выглядеть как:

$$\text{GrowthRate(X)} = \begin{cases} 0, \text{если } \text{supp}_1(\text{X}) = 0 \text{ и } \text{supp}_2(\text{X}) = 0 \\ \infty, \text{если } \text{supp}_1(\text{X}) = 0 \text{ и } \text{supp}_2(\text{X}) \neq 0 \\ \dfrac{\text{supp}_2(\text{X})}{\text{supp}_1(\text{X})}, \text{ иначе} \end{cases}$$

$$GR(p,+,-) = \begin{cases} 0, & \text{если } Support(p, D_-) = 0 \text{ и } Support(p, D_+) = 0 \\ \infty, & \text{если } Support(p, D_-) = 0 \text{ и } Support(p, D_+) \neq 0 \\ \dfrac{Support(p, D_+)}{Support(p, D_-)}, & \text{иначе} \end{cases}$$

Паттерны отбираются путем задания минимальной границы отношения прироста $\theta$. Как и в статье [Dong, 1999, p. 43-52]. То есть мы задаем минимальное отношение прироста, для которого мы хотим отбирать паттерны:

$$GR(g,+,-) > \theta$$

**Использование эмерджентных паттернов для классификации**

Для нового объекта мы должны рассчитать рейтинг его принадлежности каждому из доступных классов. Затем сравниваем полученные рейтинги и принимаем решение о классификации нового объекта.

Пусть есть новый объект, обозначим его s. Тогда рейтинг для положительного класса будет записан следующим образом:

$$score_+(s) = \frac{\sum_{p \in P_+} GR(p,+,-)}{median\big(GR(P_+)\big)},$$

где $p$ – неразрывная префиксная подпоследовательность для $s$.

То есть мы из множества всех паттернов P, отобранных по минимальному отношению прироста, выбираем те, которые являются неразрывными префиксными подпоследовательностями для нашей новой последовательности, а затем суммируем все отношения приростов. Далее мы нормируем оценку путем деления ее на медиану значений отношения прироста для данного класса. Все это соотносится с методом из статьи [Dong, 1999, p. 43-52], только примененном к другой структуре паттернов.

**Использование префиксного дерева**

Для поиска узорных понятий мы будем использовать специальную структуру, похожую на префиксное дерево с некоторыми дополнительными атрибутами на вершинах. В обычном префиксном дереве одной вершине соответствует какой-то набор символов. В нашем дереве одной вершине будет соответствовать только один символ.

Например, для последовательности $\langle\{event\_1\},\{event\_2\},\{event\_3\}\rangle$; $\langle\{event\_2\},\{event\_3\},\{event\_1\}\rangle$; $\langle\{event\_2\},\{event\_3\},\{event\_4\}\rangle$ получим следующее дерево.

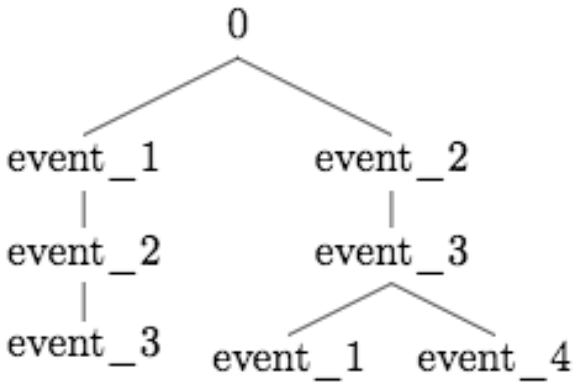


Рисунок 1

Рассмотрим теперь пример приближенный к нашим данным о жизни людей. Пусть у нас есть два набора последовательностей, один – соответствующий женщинам, второй соответствующий мужчинам.

Мужчины:

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

Женщины:

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

Теперь построим префиксно-подобное дерево для этих последовательностей. В каждой вершине будут храниться события и количество последовательностей каждого класса, в которых данная цепочка событий произошла. class0 – женщины, class1 – мужчины.

Теперь поиск гипотез и эмерджентных паттернов сведется к тому, что мы будем идти по дереву, смотреть на вершину и, если поддержка вершины удовлетворяет значению минимальной поддержки, а также поддержка вершины для конкретного класса больше, чем поддержка его ребенка для этого же класса, то можно считать путь от корня до этой вершины замкнутым паттерном.

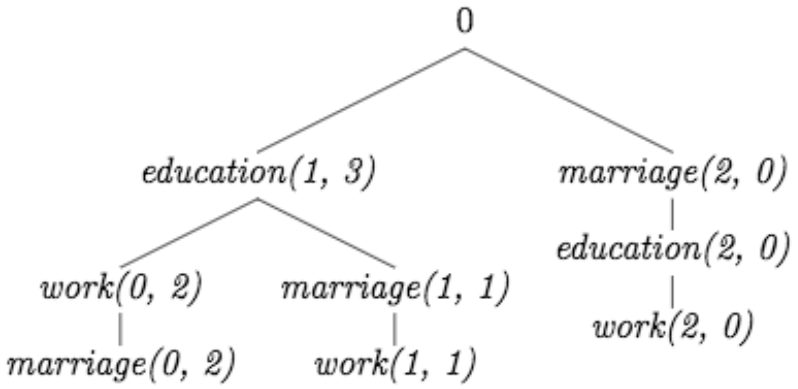


Рисунок 2

Например, мы можем увидеть, что согласно префиксно-подобному дереву у одной женщины и трех мужчин есть неразрывная префиксная подпоследовательность $\langle\{education\}\rangle$. И это правда. У всех трех мужчин последовательности начинаются с {education}, также только у одной женщины последовательность начинается с {education}. Также согласно префиксному дереву один мужчина и одна женщина имеют следующую неразрывную префиксную подпоследовательность $\langle\{education\},\{marriage\},\{work\}\rangle$, что совпадает с нашими данными.

Основываясь на этой структуре данных, мы можем вычислительно эффективно посчитать поддержку и отношение прироста для каждой последовательности.

**Эксперименты и результаты**

Для реализации экспериментов классификации и выявления паттернов был использован язык программирования Python, а также библиотека Gapless Sequences Analysis, написанная нами.

**Паттерны присущие конкретным классам**

Мы поставили ограничение на частоту в 9%. То есть полученный при этом ограничении замкнутый неразрывный префиксный паттерн должен встретиться как минимум у 9% всех опрошенных. Получили следующие паттерны:

**Таблица 2**

**Женщины**

| | |
|---|---|
| $\langle\{work\}\rangle$ | 0.287 |
| $\langle\{work\},\{education\}\rangle$ | 0.120 |
| $\langle\{separation\}\rangle$ | 0.283 |
| $\langle\{education\}\rangle$ | 0.239 |
| $\langle\{education\},\{work\}\rangle$ | 0.168 |
| $\langle\{separation\},\{education\}\rangle$ | 0.110 |
| $\langle\{separation\},\{education\},\{work\}\rangle$ | 0.097 |

**Таблица 3**

**Мужчины**

| | |
|---|---|
| $\langle\{work\}\rangle$ | 0.329 |
| $\langle\{work\},\{education\}\rangle$ | 0.155 |
| $\langle\{separation\}\rangle$ | 0.266 |
| $\langle\{education\}\rangle$ | 0.276 |
| $\langle\{education\},\{work\}\rangle$ | 0.103 |
| $\langle\{separation\},\{education\}\rangle$ | 0.199 |
| $\langle\{separation\},\{education\},\{work\}\rangle$ | 0.099 |

Можем сделать вывод, что начала траекторий жизни людей не зависит сильно от пола и самые популярные начала траекторий совпадают для обоих полов.

**Классификация при помощи эмерждентных паттернов**

Все наши данные мы разделили на 2 группы: обучающее множество и тестовое множество в процентном соотношении 80% - 20%.

Мы выбрали одинаковые значения минимальной поддержки для обоих классов – это 0.004. Это значит, что паттерн должен быть как минимум у 5 мужчин и 9 женщин. Затем мы провели несколько классификаций с различными минимальными значениями отношения прироста в рамках массива [1.2, 1.5, 2, 3, 5, 7, infinity].

Полные таблицы с результатами приведены в приложении. Красным выделены наилучшие результаты по каждой метрике.

Так как нам важно было выявить интересные ярко отличительные паттерны, мы не пытались решить проблему того, что много объектов из тестового множества остается совсем без предположения о принадлежности какому-то классу. Например, в эксперименте 41 классифицировано всего чуть больше 1% людей из тестовой выборки. Но при этом у нас средние значения precision и recall достигают 0.79%. То есть из результатов можно сделать вывод, что паттерны, которые являются ярко отличительными для какого-то класса относительно другого, имеют небольшое покрытие, а значит средние поведения мужчин и женщин не имеет какого-то сильного отличия, если смотреть в общем, но присутствуют локальные группы обоих классов, которые ведут себя с очень большими отличиями.

Наилучшее качество классификации показал классификатор с минимальным значением отношения прироста (7, infinity). Ему соответствуют следующие эмерджентные паттерны:

**Таблица 4**

**Женщины**

| Pattern | Growth-rate | Support |
|---|---|---|
| ⟨{work, separation}, {marriage}, {children}, {education}⟩ | inf | 0.006 |
| ⟨{ separation, partner}, {marriage}⟩ | inf | 0.005 |
| ⟨{work, separation}, {marriage}, {children}⟩ | inf | 0.008 |
| ⟨{work, separation}, {marriage}⟩ | inf | 0.009 |

**Таблица 5**

**Мужчины**

| Pattern | GR | Sup |
|---|---|---|
| ⟨{education}, {marriage}, {work}, {children}, {separation}⟩ | 10.6 | 0.006 |
| ⟨{education}, {marriage}, {work}, {children}⟩ | 12.7 | 0.007 |
| ⟨{education}, {work}, {partner}, {marriage}, {separation}, {children}⟩ | 10.6 | 0.006 |

Мы получили 7 последовательностей, наиболее четко разделяющих мужчин и женщин: 3 последовательности, определяющих мужчин, и 4, определяющих женщин. Отношение прироста показывает, что все «женские» последовательности типичны только для женщин (growth_rate = infty), отношение прироста «мужских» последовательностей находится в диапазоне от 10,6 до 12,7, что означает, что данные последовательности показательны для мужчин в 10-12 раз больше, чем для женщин.

Внутри «женских» последовательностей первое событие «отделение» происходит одновременно с другими событиями: «работа» (3 случая из 4) и «партнер» (1 случай из 4). Второе событие для женщин «замужество», третье, если есть, это «дети» и четвертое «образование».

Полученные выше результаты показывают, что женщины, наиболее непохожие на мужчин, склонны начинать взрослую жизнь с отделения от семьи. Только в одном случае отделение связано с рождением ребенка, в остальных случаях мы видим образ независимой женщины, у которой есть работа и которая отделилась от родителей. Второй шаг во всех случаях – замужество. Мы видим, что финансово независимая женщина создает свою семью и рожает ребенка. Самая длинная последовательность содержит событие «получение высшего образования». Таким образом, только после 4 важных социально-экономических и социально-демографических событий наша типичная женщина завершает образование.

Рассмотрим «мужские» последовательности. В них первое событие для мужчин – образование. В отличие от женщин, мужчины раньше получают образование. Это показывает не только жизненные приоритеты мужчин и женщин, но и разницу в наивысшей ступени получения образования. Второе событие для мужчин – это женитьба

(2 случая из 3) и работа (1 случай из 3). Как и женщины, мужчины склонны создавать семью достаточно рано, но в отличие от женщин, которые к моменту создания семьи уже имеют работу и независимы, мужчины к этому моменту обладают только образованием. Мужчины, для которых женитьбы является вторым шагом затем получают работу и становятся отцами. В качестве последнего шага они покидают родительский дом, что является окончательным шагом к взрослению. Мужчины, для которых работа – второй шаг, имеют другой набор следующих шагов: у них появляется первый партнер, затем они женятся, покидают родительский дом и последним шагом становятся родителями.

**Заключение**

Основным результатом работы является применение различных методов анализа данных к анализу демографических последовательностей. Следующие выводы можно сделать по результатам работы:

- В данной работе было изучено применение методов анализа последовательностей к задачам демографического направления. В частности, задаче поиска паттернов, характеризующих отдельные классы данных.
- Был разработан и реализован новый метод анализа паттернов специального типа (неразрывных и префиксных).
- Получены и проинтерпретированы паттерны поведения для разных классов.
- Разработан и протестирован классификатор на языке Python на основе эмерджентных частых последовательностей и узорных структур.

**Библиографический список**

Aassve A., Billari F.C., Piccarreta R. Strings of adulthood: A sequence analysis of young british womens work-family trajectories // European Journal of Population. 2007. 23(3/4). P. 369–388.

Agrawal R., Srikant R. Mining sequential patterns // In: Proceedings of the Eleventh International Conference on Data Engineering. 1995. P. 3–14.

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. 2010. №38(3). P. 420–462.

Billari F., Frnkranz J., Prskawetz A. Timing, Sequencing, and Quantum of Life Course Events: A Machine Learning Approach // European Journal of Population. 2006. №22(1). P. 37–65.

Billari F., Piccarreta R. Analyzing demographic life courses through sequence analysis // Mathematical Population Studies. 2005. №12(2). P. 81–106.

Blockeel H., Furnkranz J., Prskawetz A., Billari F.C. Detecting temporal change in event sequences: An application to demographic data // In: Principles of Data Mining and Knowledge Discovery, 5th Eur. Conf., PKDD. 2001. P. 29–41.

Braboy P., 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. 2005. №9. P. 55–90.

Brzinsky-Fay C., Kohler U., Luniak M. Sequence analysis with Stata // Stata Journal 6, №4. 2006. P. 435.

Buzmakov A., Egho E., Jay N., Kuznetsov S.O., Napoli A., Raissi C. On mining complex sequential data by means of FCA and pattern structures // Int. J. General Systems. 2016. №45(2). P. 135–159.

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 ’99, ACM. 1999. P. 43–52.

Gabadinho A., Ritschard G., Miller N.S., Studer M. Analyzing and Visualizing State Sequences in R with TraMineR // J. of Stat. Software. 2011. №40(4). P. 1–37.

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. 2009. №38(1). P. 197–231.

Ignatov D.I., Mitrofanova E., Muratova A., Gizdatullin D. Pattern mining and machine learning for demographic sequences // In: Knowledge Engineering and Semantic Web, KESW. 2015. P. 225–239.

Kaytoue M., Codocedo V., Buzmakov A., Baixeries J., Kuznetsov S.O., Napoli A. Pattern structures and concept lattices for data mining and knowledge processing // In: ECML PKDD, Proceedings, Part III. 2015. P. 227–231.

Kuznetsov S.O. Learning of simple conceptual graphs from positive and negative examples // In: PKDD ’99, Proceedings. 1999. P. 384–391.

Pei J., Han J., Mortazavi-Asl B., Wang J., Pinto H., Chen Q., Dayal U., Hsu M. Mining sequential patterns by pattern-growth: The prefixspan approach // IEEE Trans. Knowl. Data Eng. 2004. №16(11). P. 1424–1440.

Ritschard G., Oris M. Life course data in demography and social sciences: Statistical and data-mining approaches // Adv. in Life Course Research. 2005. №10. P. 283–314.

**Приложение**

| idx | $\text{minGR}_M$ | $\text{minGR}_W$ | $\text{n\_rules}_M$ | $\text{n\_rules}_W$ | accuracy | MAVG_Precision |
|---|---|---|---|---|---|---|
| 0.0 | 1.2 | 1.2 | 82.0 | 94.0 | 0.55 | 0.56 |
| 1.0 | 1.2 | 1.5 | 82.0 | 69.0 | 0.51 | 0.56 |
| 2.0 | 1.2 | 2.0 | 82.0 | 50.0 | 0.46 | 0.57 |
| 3.0 | 1.2 | 3.0 | 82.0 | 35.0 | 0.44 | 0.58 |
| 4.0 | 1.2 | 5.0 | 82.0 | 18.0 | 0.39 | 0.57 |
| 5.0 | 1.2 | 7.0 | 82.0 | 13.0 | 0.38 | 0.56 |
| 6.0 | 1.2 | inf | 82.0 | 4.0 | 0.36 | 0.61 |
| 7.0 | 1.5 | 1.2 | 57.0 | 94.0 | 0.61 | 0.57 |
| 8.0 | 1.5 | 1.5 | 57.0 | 69.0 | 0.58 | 0.58 |
| 9.0 | 1.5 | 2.0 | 57.0 | 50.0 | 0.54 | 0.59 |
| 10.0 | 1.5 | 3.0 | 57.0 | 35.0 | 0.51 | 0.6 |
| 11.0 | 1.5 | 5.0 | 57.0 | 18.0 | 0.46 | 0.59 |
| 12.0 | 1.5 | 7.0 | 57.0 | 13.0 | 0.44 | 0.58 |
| 13.0 | 1.5 | inf | 57.0 | 4.0 | 0.41 | 0.63 |
| 14.0 | 2.0 | 1.2 | 39.0 | 94.0 | 0.65 | 0.57 |
| 15.0 | 2.0 | 1.5 | 39.0 | 69.0 | 0.62 | 0.58 |
| 16.0 | 2.0 | 2.0 | 39.0 | 50.0 | 0.59 | 0.59 |
| 17.0 | 2.0 | 3.0 | 39.0 | 35.0 | 0.55 | 0.6 |
| 18.0 | 2.0 | 5.0 | 39.0 | 18.0 | 0.49 | 0.59 |
| 19.0 | 2.0 | 7.0 | 39.0 | 13.0 | 0.47 | 0.59 |
| 20.0 | 2.0 | inf | 39.0 | 4.0 | 0.43 | 0.63 |
| 21.0 | 3.0 | 1.2 | 19.0 | 94.0 | 0.7 | 0.56 |
| 22.0 | 3.0 | 1.5 | 19.0 | 69.0 | 0.7 | 0.57 |
| 23.0 | 3.0 | 2.0 | 19.0 | 50.0 | 0.68 | 0.58 |
| 24.0 | 3.0 | 3.0 | 19.0 | 35.0 | 0.65 | 0.59 |

| idx | MAVG_Recall | $\text{Precision}_M$ | $\text{Precisoon}_W$ | $\text{Recall}_M$ | $\text{Recall}_W$ | $\text{n\_cl}_M$ | $\text{n\_cl}_W$ |
|---|---|---|---|---|---|---|---|
| 0.0 | 0.57 | 0.38 | 0.74 | 0.61 | 0.53 | 0.03 | 0.07 |
| 1.0 | 0.57 | 0.37 | 0.76 | 0.72 | 0.41 | 0.16 | 0.2 |
| 2.0 | 0.56 | 0.36 | 0.78 | 0.85 | 0.26 | 0.21 | 0.27 |
| 3.0 | 0.55 | 0.36 | 0.79 | 0.9 | 0.2 | 0.23 | 0.32 |
| 4.0 | 0.52 | 0.35 | 0.78 | 0.95 | 0.09 | 0.27 | 0.36 |
| 5.0 | 0.51 | 0.35 | 0.77 | 0.96 | 0.07 | 0.27 | 0.38 |
| 6.0 | 0.51 | 0.35 | 0.86 | 0.99 | 0.02 | 0.3 | 0.41 |
| 7.0 | 0.57 | 0.39 | 0.74 | 0.47 | 0.67 | 0.26 | 0.25 |
| 8.0 | 0.59 | 0.39 | 0.76 | 0.61 | 0.57 | 0.4 | 0.41 |
| 9.0 | 0.59 | 0.4 | 0.78 | 0.77 | 0.42 | 0.48 | 0.53 |
| 10.0 | 0.58 | 0.4 | 0.79 | 0.84 | 0.33 | 0.52 | 0.59 |
| 11.0 | 0.55 | 0.4 | 0.78 | 0.92 | 0.18 | 0.56 | 0.67 |
| 12.0 | 0.53 | 0.4 | 0.77 | 0.93 | 0.14 | 0.57 | 0.68 |
| 13.0 | 0.51 | 0.4 | 0.86 | 0.99 | 0.04 | 0.59 | 0.72 |
| 14.0 | 0.57 | 0.41 | 0.74 | 0.38 | 0.76 | 0.37 | 0.33 |
| 15.0 | 0.59 | 0.4 | 0.76 | 0.51 | 0.67 | 0.52 | 0.5 |
| 16.0 | 0.61 | 0.4 | 0.78 | 0.67 | 0.55 | 0.63 | 0.64 |
| 17.0 | 0.61 | 0.4 | 0.79 | 0.76 | 0.45 | 0.68 | 0.7 |
| 18.0 | 0.57 | 0.4 | 0.78 | 0.87 | 0.26 | 0.72 | 0.78 |
| 19.0 | 0.55 | 0.4 | 0.77 | 0.89 | 0.21 | 0.73 | 0.79 |
| 20.0 | 0.52 | 0.4 | 0.86 | 0.98 | 0.06 | 0.75 | 0.83 |
| 21.0 | 0.53 | 0.38 | 0.74 | 0.16 | 0.9 | 0.53 | 0.43 |
| 22.0 | 0.55 | 0.38 | 0.76 | 0.25 | 0.85 | 0.69 | 0.61 |
| 23.0 | 0.58 | 0.38 | 0.78 | 0.39 | 0.77 | 0.8 | 0.75 |
| 24.0 | 0.6 | 0.38 | 0.79 | 0.5 | 0.7 | 0.85 | 0.81 |

| idx | $\text{minGR}_M$ | $\text{minGR}_W$ | $\text{n\_rules}_M$ | $\text{n\_rules}_W$ | accuracy | MAVG_Precision |
|---|---|---|---|---|---|---|
| 25.0 | 3.0 | 5.0 | 19.0 | 18.0 | 0.56 | 0.58 |
| 26.0 | 3.0 | 7.0 | 19.0 | 13.0 | 0.53 | 0.57 |
| 27.0 | 3.0 | inf | 19.0 | 4.0 | 0.44 | 0.62 |
| 28.0 | 5.0 | 1.2 | 8.0 | 94.0 | 0.74 | 0.63 |
| 29.0 | 5.0 | 1.5 | 8.0 | 69.0 | 0.75 | 0.64 |
| 30.0 | 5.0 | 2.0 | 8.0 | 50.0 | 0.77 | 0.65 |
| 31.0 | 5.0 | 3.0 | 8.0 | 35.0 | 0.77 | 0.66 |
| 32.0 | 5.0 | 5.0 | 8.0 | 18.0 | 0.73 | 0.65 |
| 33.0 | 5.0 | 7.0 | 8.0 | 13.0 | 0.71 | 0.64 |
| 34.0 | 5.0 | inf | 8.0 | 4.0 | 0.68 | 0.69 |
| 35.0 | 7.0 | 1.2 | 3.0 | 94.0 | 0.74 | 0.73 |
| 36.0 | 7.0 | 1.5 | 3.0 | 69.0 | 0.76 | 0.74 |
| 37.0 | 7.0 | 2.0 | 3.0 | 50.0 | 0.79 | 0.75 |
| 38.0 | 7.0 | 3.0 | 3.0 | 35.0 | 0.8 | 0.75 |
| 39.0 | 7.0 | 5.0 | 3.0 | 18.0 | 0.79 | 0.75 |
| 40.0 | 7.0 | 7.0 | 3.0 | 13.0 | 0.78 | 0.74 |
| 41.0 | 7.0 | inf | 3.0 | 4.0 | 0.89 | 0.79 |
| 42.0 | inf | 1.2 | 0.0 | 94.0 | 0.74 | 0.37 |
| 43.0 | inf | 1.5 | 0.0 | 69.0 | 0.76 | 0.38 |
| 44.0 | inf | 2.0 | 0.0 | 50.0 | 0.79 | 0.39 |
| 45.0 | inf | 3.0 | 0.0 | 35.0 | 0.8 | 0.4 |
| 46.0 | inf | 5.0 | 0.0 | 18.0 | 0.79 | 0.39 |
| 47.0 | inf | 7.0 | 0.0 | 13.0 | 0.78 | 0.38 |
| 48.0 | inf | inf | 0.0 | 4.0 | 0.92 | 0.43 |

| idx | MAVG_Recall | $\text{Precision}_M$ | $\text{Precisoon}_W$ | $\text{Recall}_M$ | $\text{Recall}_W$ | $\text{n\_cl}_M$ | $\text{n\_cl}_W$ |
|---|---|---|---|---|---|---|---|
| 25.0 | 0.59 | 0.38 | 0.78 | 0.68 | 0.5 | 0.89 | 0.88 |
| 26.0 | 0.58 | 0.38 | 0.77 | 0.73 | 0.43 | 0.9 | 0.9 |
| 27.0 | 0.55 | 0.38 | 0.86 | 0.95 | 0.15 | 0.92 | 0.93 |
| 28.0 | 0.52 | 0.52 | 0.74 | 0.06 | 0.98 | 0.58 | 0.48 |
| 29.0 | 0.53 | 0.52 | 0.76 | 0.1 | 0.97 | 0.74 | 0.65 |
| 30.0 | 0.56 | 0.52 | 0.78 | 0.17 | 0.95 | 0.86 | 0.79 |
| 31.0 | 0.59 | 0.52 | 0.79 | 0.25 | 0.93 | 0.9 | 0.86 |
| 32.0 | 0.63 | 0.52 | 0.78 | 0.41 | 0.85 | 0.94 | 0.93 |
| 33.0 | 0.63 | 0.52 | 0.77 | 0.46 | 0.8 | 0.95 | 0.95 |
| 34.0 | 0.68 | 0.52 | 0.86 | 0.87 | 0.5 | 0.97 | 0.98 |
| 35.0 | 0.51 | 0.71 | 0.74 | 0.02 | 1.0 | 0.6 | 0.48 |
| 36.0 | 0.52 | 0.71 | 0.76 | 0.04 | 0.99 | 0.76 | 0.66 |
| 37.0 | 0.53 | 0.71 | 0.78 | 0.07 | 0.99 | 0.87 | 0.8 |
| 38.0 | 0.55 | 0.71 | 0.79 | 0.11 | 0.99 | 0.91 | 0.86 |
| 39.0 | 0.59 | 0.71 | 0.78 | 0.21 | 0.97 | 0.96 | 0.94 |
| 40.0 | 0.61 | 0.71 | 0.77 | 0.25 | 0.96 | 0.96 | 0.96 |
| 41.0 | 0.79 | 0.71 | 0.86 | 0.71 | 0.86 | 0.99 | 0.99 |
| 42.0 | 0.5 | 0.0 | 0.74 | 0.0 | 1.0 | 0.61 | 0.49 |
| 43.0 | 0.5 | 0.0 | 0.76 | 0.0 | 1.0 | 0.77 | 0.66 |
| 44.0 | 0.5 | 0.0 | 0.78 | 0.0 | 1.0 | 0.88 | 0.8 |
| 45.0 | 0.5 | 0.0 | 0.79 | 0.0 | 0.99 | 0.92 | 0.87 |
| 46.0 | 0.49 | 0.0 | 0.78 | 0.0 | 0.99 | 0.97 | 0.94 |
| 47.0 | 0.49 | 0.0 | 0.77 | 0.0 | 0.98 | 0.97 | 0.96 |
| 48.0 | 0.46 | 0.0 | 0.86 | 0.0 | 0.92 | 1.0 | 0.99 |