



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

Выявление знаний в демографических последовательностях

Муратова А.А.

Гиздатуллин Д.К.

Игнатов Д.И.

Митрофанова Е.С.

План

- Постановка задачи
- Демографические данные
- Методы и результаты
 - Деревья решений
 - Предсказание первого события
 - Предсказание последнего события
 - Правила для целевого признака пол
 - Частые и эмерджентные последовательности
 - Последовательности, характерные для мужчин и женщин
- Выводы



Постановка задачи

- Анализ демографических данных методами и программными средствами машинного обучения и майнинга данных.
- Вопросы демографов:
 - Какие существуют различия между демографическим поведением мужчин и женщин?
 - Какие частые последовательности событий встречаются в жизни человека?
 - Какие первые и последние стартовые события могут быть у человека?
- Используемые инструменты:
 - SPMF
 - Orange
 - Программы собственной разработки на Python



Демографические данные

Таблица содержит результаты опроса «Родители и дети, мужчины и женщины в семье и обществе» www.socpol.ru/gender/RIDMIZ.shtml

- 4857 человек, среди которых 1545 мужчин и 3312 женщин
- 11 поколений с интервалом в 5 лет с 1930 по 1984 год

пол	тип образования	место жительства	религиозность	частота посещения служб	поколения	дата рождения	партнер	брак	расставание	развод	ребенок	отделение от родителей	работа	образование
1 - мужчины 2 - женщины	1 - общее 2 - профессионал 3 - высшее	1 - город 2 - ПГТ 3 - село	1 - верит 0 - не верит/ не указал	1 - неск. раз в неделю 2 - раз в неделю 3 - мин. раз в месяц										
2	1	2	1	4	9	01.03.1970		01.03.1988			01.02.1989	01.03.1988	01.04.1998	01.06.1988
2	2	2	1	4	10	01.08.1976		01.11.1999			01.04.2000	01.11.1999	01.09.1993	01.04.2007
2	3	2	1	4	3	01.11.1941		01.01.1970			01.11.1970	01.09.1961	01.07.1961	01.06.1965
1	2	2	1	0	9	01.10.1973	01.06.1996		01.12.1999		01.09.1997	01.12.1991	01.08.1994	01.06.1988
2	2	2	1	3	3	01.04.1941		01.12.1964			01.06.1966	01.01.1965	01.06.1964	01.06.1964
1	3	2	1	0	7	01.06.1961		01.06.1989			01.08.1990	01.09.1978	01.11.1987	01.06.1984
1	1	2	1	0	2	01.03.1939		01.08.1968			01.03.1969	01.10.1957	01.08.1957	01.08.1957
1	3	2	1	0	8	01.01.1967	01.02.1991	01.08.1998	01.01.1999	01.07.1999	01.12.1993	01.09.1984	01.02.1990	01.02.1997
2	1	2	1	3	1	01.02.1932		01.12.1955			01.11.1956	01.12.1955	01.04.1951	01.06.1945
1	2	2	1	0	8	01.09.1966		01.10.1989			01.03.1990	01.11.1989	01.10.1986	01.05.1985
2	2	2	1	0	7	01.11.1962		01.01.1984			01.02.1985	01.05.1985	01.07.1982	01.06.1982
2	2	2	1	4	6	01.12.1955		01.07.1981			01.06.1982	01.03.1993	01.01.1977	01.06.1976
1	2	2	1	0	8	01.11.1965		01.02.1987			01.07.1988	01.12.1985	01.03.1983	01.02.1983
1	2	2	1	0	4	01.04.1949		01.01.1969			01.04.1973	01.09.1973	01.06.1971	01.08.1968
2	1	2	1	2	3	01.03.1940					01.10.1965	01.06.1964	01.04.1968	01.06.1960
2	1	2	1	3	4	01.03.1949		01.12.1966			01.04.1967	01.12.1966	01.03.1967	01.06.1966
2	3	2	1	4	3	01.05.1943					01.10.1965	01.05.1965	01.09.1963	01.06.1968
2	1	2	1	3	4	01.01.1947					01.10.1968	01.09.1962	01.09.1962	01.06.1965
2	2	2	1	4	2	01.10.1939	01.12.1961	01.05.1962			01.04.1962	01.08.1957	01.06.1961	01.06.1961
1	1	2	1	0	8	01.04.1965		01.05.1986			01.06.1987	01.06.1986	01.07.1985	01.03.1983
2	2	2	1	3	3	01.04.1943		01.10.1964			01.09.1965	01.08.1967	01.08.1963	01.11.1963

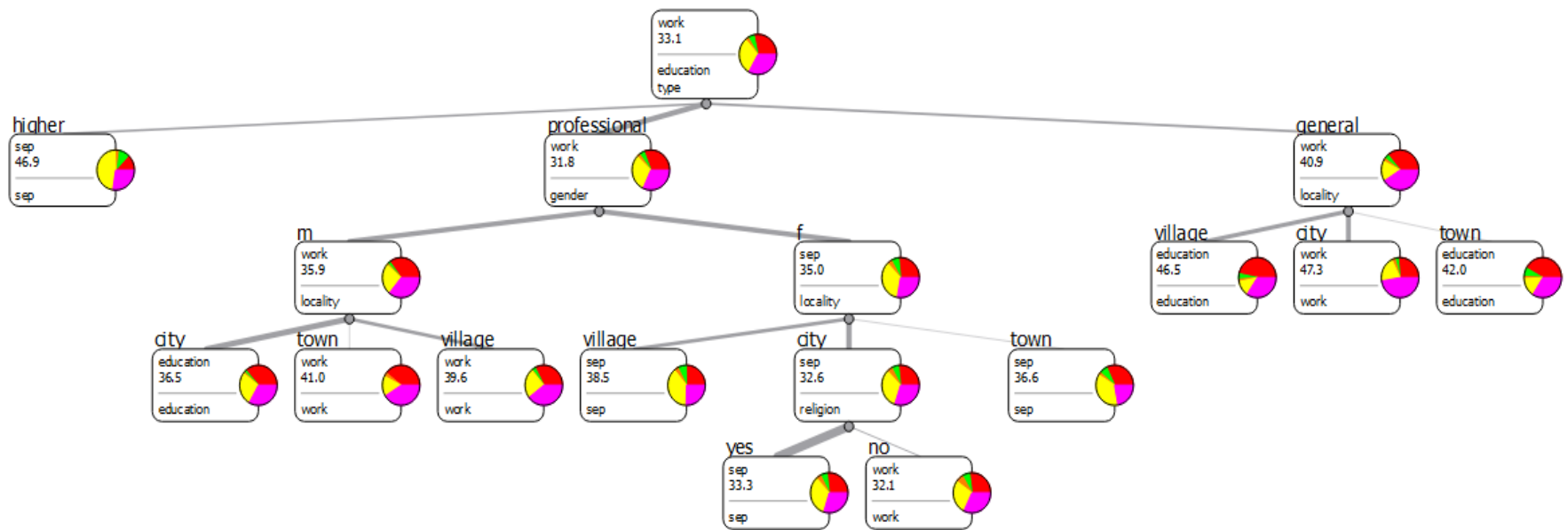


Сравнение методов на примере первого события

Методы	Правильность	F-мера	Точность	Чувствительность
Образование				
Деревья решений	0.43	0.42	0.44	0.39
kNN	0.39	0.40	0.40	0.40
SVM	0.42	0.42	0.45	0.39
Отделение от родителей				
Деревья решений	0.43	0.47	0.41	0.53
kNN	0.39	0.42	0.41	0.44
SVM	0.42	0.50	0.40	0.64
Первая работа				
Деревья решений	0.43	0.45	0.44	0.47
kNN	0.39	0.42	0.41	0.43
SVM	0.42	0.40	0.45	0.36
Первый ребенок				
Деревья решений	0.43	n/a	n/a	0
kNN	0.39	n/a	0	0
SVM	0.42	n/a	n/a	0
Первый брак				
Деревья решений	0.43	n/a	0	0
kNN	0.39	0.08	0.12	0.06
SVM	0.42	n/a	n/a	0
Первый партнер				
Деревья решений	0.43	n/a	0	0
kNN	0.39	0.1	0.16	0.07
SVM	0.42	n/a	n/a	0



Предсказание первого события с помощью дерева решений



Дерево построено с помощью программы **Orange**

Первое событие больше всего зависит от типа образования

Кодирование

Общие признаки:

- Пол
- Поколение
- Тип образования
- Населенный пункт
- Религиозность
- Частота посещения религиозных служб

События:

- Бинарное кодирование (0 - нет, 1 - да)
- Временное кодирование (возраст в месяцах)
- Пары событий, очередность ($<$, $>$, $=$, n)



Сравнение типов кодирования для предсказания последнего события

Тип кодирования	СА (точность)	
	Несбалансированные данные	Сбалансированные данные
Бинарное	0.8498	0.8780 (*)
Временное	0.3516	0.3591
Пары событий	0.7076	0.7013
Бинарное и временное	0.7293 (~)	0.7459
Бинарное и пары событий	0.8407	0.8438
Временное и пары событий	0.5465	0.4959
Бинарное, временное и пары событий	0.7295 (~)	0.7503

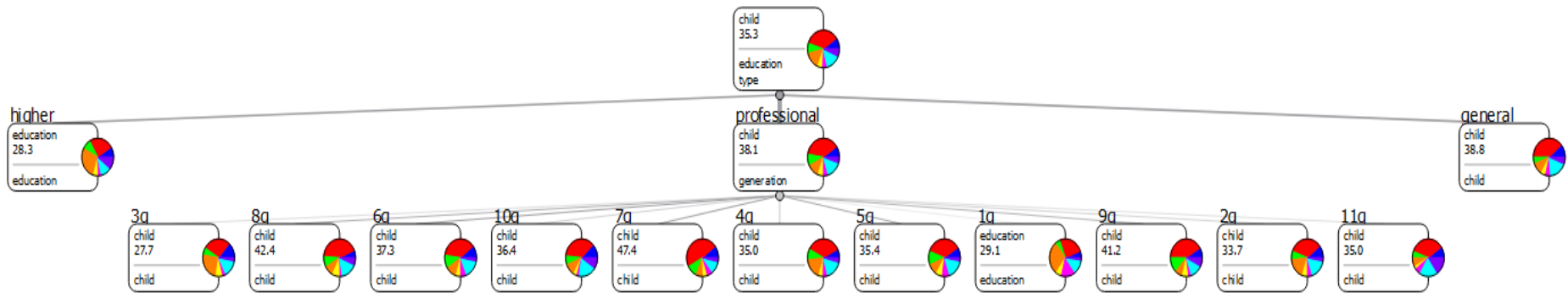
(*) – лучший результат, (~) – эквивалентные результаты

Примеры правил для предсказания последнего события

Посылка (путь в дереве)	Заключение (лист)	Вероятность
Образование и рождение ребенка	Отделение от родителей	93.9%
Образование и первая работа	Отделение от родителей	76.7%
Образование, отделение от родителей и рождение ребенка	Первая работа	98.9%
Образование	Первая работа	90.3%
Мужчина, образование, отделение от родителей, первая работа, партнер и рождение ребенка	Брак	83.2%
Образование, отделение от родителей и брак	Рождение ребенка	95.7%
Женщина, образование, отделение от родителей и первая работа	Рождение ребенка	78.1%
Образование (общее или профессиональное), отделение от родителей, первая работа и партнер	Рождение ребенка	60.5% и 54.5% соответственно
Образование, отделение от родителей, первая работа, партнер, брак и рождение ребенка	Расставание	78.1%
Первая работа, отделение от родителей, образование, брак и рождение ребенка	Развод	72.9%



Предсказание последнего события на основе признаков



Дерево решений, построенное только на основе признаков
Последнее событие также больше всего зависит от типа образования

Сравнение типов кодирования для предсказания целевого признака «пол»

Тип кодирования	СА (точность)	
	Несбалансированные данные	Сбалансированные данные
Бинарное	0.6838	0.5824
Временное	0.6827	0.6758
Пары событий	0.6817	0.5896
Бинарное и временное	0.6842 (~)	0.6647
Бинарное и пары событий	0.6815	0.5923
Временное и пары событий	0.6827	0.6743
Бинарное, временное и пары событий	0.6842 (~)	0.6915 (*)

(*) – лучший результат, (~) – эквивалентные результаты

Примеры правил для предсказания целевого признака «пол»

Мужчины:

Посылка	Вероятность
Работа после 21 лет, брак после 29.5 лет	70.2%
Ребенок после 22.9 лет, брак в 23.9-29.5 лет	69.3%
Брак в 23.3-23.9 лет, ребенок после 24.7 лет	67%
Работа после 19.9 лет, брак в 20.6-22.4 лет, образование до 20.7 лет, расставание после 27.6 лет, развод до 30.5 лет	65.9%

Женщины:

Посылка	Вероятность
Брак до 20.6 лет	77.1%
Работа в 18.2-19.9 лет, брак в 20.6-22.4 лет, расставание после 27.6 лет, развод после 30.5 лет	71.9%
Работа в 18.2-19.9 лет, брак в 20.6-22.4 лет, расставание после 27.6 лет, развод до 30.5 лет	70.9%
Работа в 17.2-19.9 лет, брак в 20.6-22.4 лет, расставание до 27.6 лет	62.8%



Префиксные подпоследовательности

- **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' обозначается как
$$Freq(s', T) = \frac{Support(s', T)}{n}$$
 где n – количество последовательностей в наборе T .

- Каждая последовательность s' такая, что $Freq(s', T) \geq minFreq$ называется **префиксным неразрывным паттерном**.

- Неразрывный префиксный паттерн называется **замкнутым**, если не существует паттерна d большей длины с такой же поддержкой.



Эмерджентные паттерны

- **Эмерджентный неразрывный префиксный паттерн** – это эмерджентный паттерн, характерный для одного класса последовательностей, но при этом нехарактерный для другого.
- **Отношение прироста** для паттерна p на двух классах '+' и '-' равно

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

- Паттерны отбираются путем задания **минимальной границы отношения прироста (порога)** $GR(g, +, -) > \theta$.
- Пусть есть новый объект, обозначим его s . Тогда **рейтинг** для положительного класса будет записан следующим образом:

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

где p – неразрывная префиксная подпоследовательность для s .



Характерные паттерны

- Последовательности, характерные для **мужчин**

$\langle\{work\}\rangle$	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

- Последовательности, характерные для **женщин**

$\langle\{work\}\rangle$	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

Классификация при помощи эмерджентных паттернов

- Обучающее множество 80%, тестовое множество 20%
- Значение минимальной поддержки для обоих классов 0.004.



Классификация при помощи эмерджентных паттернов

- Последовательности, характерные для **мужчин**

<u>Pattern</u>	<u>GR</u>	<u>Sup</u>
$\langle \{\text{education}\}, \{\text{marriage}\}, \{\text{work}\}, \{\text{children}\}, \{\text{separation}\} \rangle$	10.6	0.006
$\langle \{\text{education}\}, \{\text{marriage}\}, \{\text{work}\}, \{\text{children}\} \rangle$	12.7	0.007
$\langle \{\text{education}\}, \{\text{work}\}, \{\text{partner}\}, \{\text{marriage}\}, \{\text{separation}\}, \{\text{children}\} \rangle$	10.6	0.006

- Последовательности, характерные для **женщин**

<u>Pattern</u>	<u>Growth-rate</u>	<u>Support</u>
$\langle \{\text{work, separation}\}, \{\text{marriage}\}, \{\text{children}\}, \{\text{education}\} \rangle$	<u>inf</u>	0.006
$\langle \{\text{separation, partner}\}, \{\text{marriage}\} \rangle$	<u>inf</u>	0.005
$\langle \{\text{work, separation}\}, \{\text{marriage}\}, \{\text{children}\} \rangle$	<u>inf</u>	0.008
$\langle \{\text{work, separation}\}, \{\text{marriage}\} \rangle$	<u>inf</u>	0.009

Выводы

- Методы классификации на основе правил (деревьев решений) и частые последовательности являются оптимальным средством для поставленных задач
- Полученные паттерны просты в интерпретации
- Следующим этапом развития может служить использование других методов анализа данных



Спасибо за внимание!

