Классификация демографических последовательностей на основе скрытых марковских моделей

Студент: Степанюк Ирина, БПМИ131
Научный руководитель: доц. Игнатов Дмитрий Игоревич

Факультет компьютерных наук
НИУ ВШЭ

Москва, 2017

Содержание

Демографические последовательности

Возможные события

- Завершение наивысшей ступени образования (education)
- Отъезд от родителей (separation)
- Первый опыт работы (work)

Примеры последовательностей

- $\langle\{separation\}, \{work\}, \{marriage\}, \{child\}\rangle$
- $\langle\{education\}, \{partner\}, \{separation\}\rangle$

Классификация касательно

- пола респондента (мужской/женский)
- поколения респондента (советское/современное)

Актуальность задачи

Актуальность

Анализ последовательностей (sequence analysis) - один из наиболее передовых методов анализа демографических данных. Методы машинного обучения и майнинга данных позволяют извлекать демографические знания, которые могут быть интересны демографам. Область нуждается в разработке новых методов и алгоритмов для улучшения качества анализа.

Актуальность задачи

Демографические исследования с использованием скрытых марковских моделей

- анализ механизмов перехода к взрослой жизни [1]
- анализ профессиональных "траекторий" населения Швейцарии [2]
- кластеризация для определения групп людей со схожим демографическим поведением среди жителей Германии [3]

Цель и задачи дипломной работы

Цель работы

Разработка метода классификации демографических последовательностей с использованием скрытых марковских моделей (Hidden Markov Models)

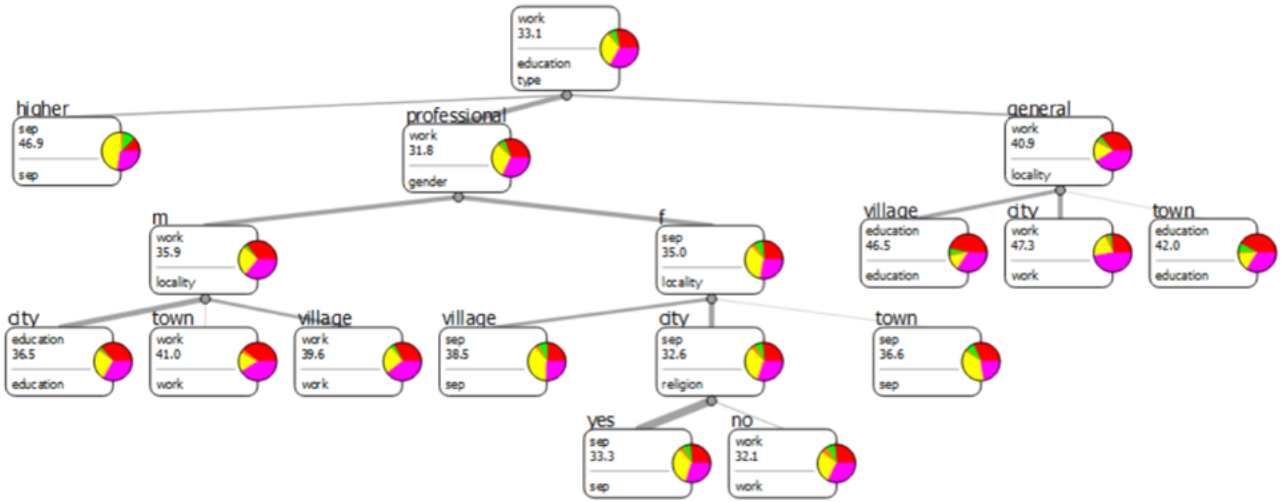
Задачи работы

- Подтвержение релевантности использования метода классификации, основанном на HMM, для анализа траекторий жизненных событий
- Сравнение предложенного подхода с другими методами классификации
- Получение результатов, представляющих интерес для демографов для их дальнейшей интерпретации

Обзор существующих методов

Методы классификации демографических последовательностей

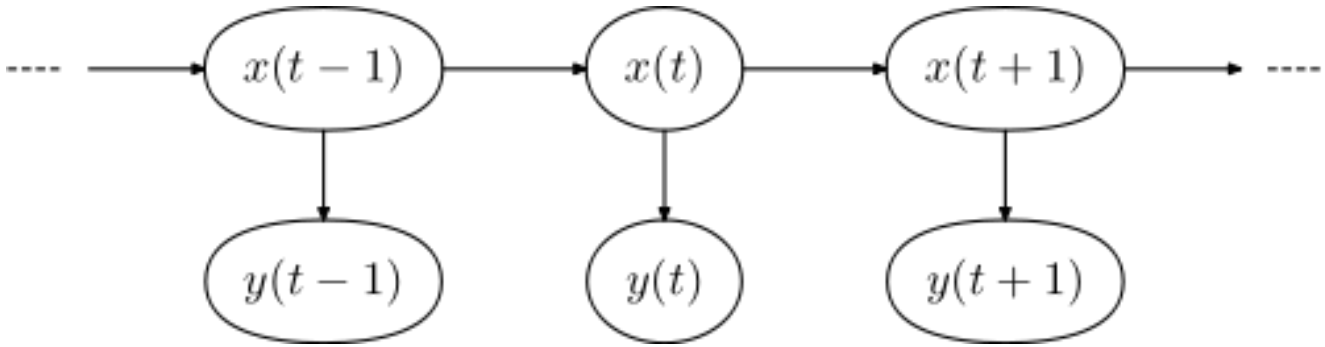
- Классификация на основе сходства (оценка расстояний между последовательностями) [4]
- Деревья решений [5]
- Марковские модели



Предложенный метод

Скрытая марковская модель

Скрытая марковская модель (Hidden Markov Model) - статистическая модель, где рассматриваются два процесса: видимый процесс наблюдаемых состояний и скрытый процесс. Скрытый процесс представляет собой марковскую цепь, а наблюдаемые состояния обусловлены этим скрытым процессом.



Классификация с помощью HMM

ML-подход

$$Class(O) = \arg\max_i P(O|M_i),$$

где M_i - модель, относящаяся к i-ому классу.

Байесовский подход

$$Class(O) = \arg\max_i P^i(O, M) = \arg\max_i \frac{P^i P(O|M_i)}{\sum_{j=1}^{C} P^j P(O|M_j)}$$

где M_i - модель, относящаяся к i-ому классу.

Построение пространства признаков

Пространство признаков

$$D(O) = \frac{1}{L}\begin{pmatrix} logP(O|M_1) \\ ... \\ logP(O|M_C) \end{pmatrix}$$

где M_i - HMM, обученная на всех последовательностях i-го класса, C - количество классов, L - длина последовательности O.

Признаки

Пусть число всех последовательностей равно N, тогда вектора признаков будут образовывать признаковое пространство $P_{HMM} = \{D_j(O)\}_{j=1}^{N}$ размерности $N \times C$.

Экспериментальная оценка

Данные

- Пол (мужской, женский)
- Поколение
- Дата рождения
- Тип населенного пункта (город, поселок городского типа, село)
- Уровень образования (высшее, профессиональное, общее)
- Религиозность (да, нет)
- Даты важных событий в жизни с точностью до месяца (окончание образования, отъезд от родителей, работа, партнер, брак, ребенок)

Результаты классификации

Результаты классификации касательно пола респондента

Метод	Точность
Деревья решений [5]	**0.693**
RNN [6]	**0.754**
HMM_{ML}	0.582
HMM_{Bayes}	0.633
SVM (linear) на P_{HMM}	0.617
SVM (rbf) на P_{HMM}	**0.687**

Результаты классификации

Классификация касательно пола для советского поколения

Метод	Точность
HMM_{ML}	0.68
HMM_{Bayes}	**0.72**
SVM (rbf kernel) на P_{HMM}	**0.72**
Random Forest на P_{HMM}	**0.7**

Классификация касательно пола для современного поколения

Метод	Точность
HMM_{ML}	0.613
HMM_{Bayes}	0.616
SVM (rbf kernel) на P_{HMM}	0.645
Random Forest на P_{HMM}	0.664

Результаты классификации

Результаты классификации касательно поколения респондента

Метод	Точность
Деревья решений	**0.661**
HMM_{ML}	0.6
HMM_{Bayes}	0.597
SVM (linear kernel) на P_{HMM}	0.634
SVM (rbf kernel) на P_{HMM}	**0.674**

Результаты классификации

Классификация с помощью SVM на P_{HMM} касательно поколения респондентов

	[1930-39]	[1940-49]	[1950-59]	[1960-69]	[1970-79]	[1980-86]
1	-	0.535	0.663	**0.703**	**0.685**	**0.804**
2	-	-	0.642	0.658	0.654	**0.802**
3	-	-	-	0.558	0.613	**0.801**
4	-	-	-	-	0.58	0.582
5	-	-	-	-	-	**0.73**

Выводы

Выводы

- Подтверждена релевантность использования метода классификации, основанном на HMM, для анализа демографических последовательностей
- Предложенный подход сравнен с работой других методов классификации последовательностей
- На основе полученых результатов сделаны интересные выводы о демографическом поведении мужчин и женщин и о поколенческих различиях

Программная реализация

Программная реализация

Скрытые марковские модели были реализованы на языке R с использованием библиотек seqHMM, HMM. Предобработка данных и построение классификаторов были реализованы на языке Python.

Список литературы I

S. Han, A. Liefbroer, C. Elzinga, G. Ritschard, and M. Studer.
Mechanisms of the transition to adulthood: an application of hidden markov models.
Sequence Analysis and Related Methods (LaCOSA II), page 155, 2016.

P. Adamopoulou, G. Ritschard, and A. Berchtold.
Using dynamic microsimulation to understand professional trajectories of the active swiss population.
In *LaCOSA II: Proceedings of the International Conference on Sequence Analysis and Related Methods*, 2016.

S. Helske, J. Helske, and M. Eerola.
Analysing complex life sequence data with hidden markov modelling.
In *LaCOSA II: Proceedings of the International Conference on Sequence Analysis and Related Methods*, 2016.

Список литературы II


Z. Xing, J. Pei, and E. Keogh.
A brief survey on sequence classification.
ACM SIGKDD Explorations Newsletter, 12:40–48, 2010.


D. Ignatov, E. Mitrofanova, A. Muratova, and D. Gizdatullin.
Pattern mining and machine learning for demographic sequences.
Communications in Computer and Information Science, 518:225–239, 2015.


А. Муратова.
Методы классификации для поиска закономерностей в демографических последовательностях.
Master's thesis, Высшая школа экономики, 2017.