НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Методы классификации для поиска закономерностей в демографических последовательностях

Муратова Анна Александровна, группа ИССА

Научный руководитель:

к.т.н. доцент, Игнатов Дмитрий Игоревич

План

- Постановка задачи
- Демографические данные
- Результаты
  - Формулы мер сходства последовательностей (без разрыва)
    - Префиксы и подпоследовательности
    - Добавление новых префиксов
    - Добавление новых подпоследовательностей
  - Использование специального ядра в SVM
  - Нейросетевые алгоритмы
  - Чистка данных
- Выводы

Постановка задачи

- Из накопленных демографических данных можно извлечь больше полезной информации, применив современные методы майнинга данных
- Сравнение методов классификации демографических данных
- Новизной данной работы является использование специальных вариантов ядер в методе SVM
- Используемые инструменты:
  - Программы собственной разработки на Python

Демографические данные

Институтом демографии НИУ ВШЭ была предоставлена база данных, содержащая 6626 человек (3314 мужчин и 3312 женщин)

Признаки:

- Пол (мужской, женский);
- Поколение (советские 1930-1969 и современные 1970-1986);
- Уровень образования (общее, профессиональное, высшее);
- Тип населенного пункта (город, поселок городского типа, село);
- Религиозность (да, нет);
- Частота посещения религиозных служб (несколько раз в неделю, раз в неделю, минимум раз в месяц, несколько раз в год и реже);

Также указаны даты рождения и даты значимых первых событий в их жизни, такие как: партнер, брак, расставание, развод, образование, работа, отделение от родителей, рождение ребенка

Основные определения

Пусть *I* – конечное множество объектов, а *X* – непустой набор объектов из *I*. **Последовательностью** *S* называется упорядоченное множество $< X_1 \dots X_n >$ где X_i – непустой набор объектов.

Размер последовательности обозначим $|S|=n$.

Длиной последовательности называется общее число объектов, встречающихся в последовательности $l(S) = \sum_{i=1}^{n} |X_i|$.

***l*-префиксом** S^l последовательности *S* называется последовательность из первых *l* элементов $< X_1 \dots X_l >$. Аналогично для **суффикса** – последние *l* элементов.

j-й **набор объектов** X_j последовательности *S* обозначим *S*[*j*].

Основные определения

Последовательность $T =< Y_1 \dots Y_m >$ называется **подпоследовательностью** последовательности $S =< X_1 \dots X_n >$, если существуют такие $1 \le i_1 < i_2 < \dots < i_m \le n$, что $Y_j \subseteq X_{i_j}$ для всех $j = 1 \dots m,\ m < n$. Тогда последовательность *S* называется **надпоследовательностью** последовательности *T*.

Множество всех подпоследовательностей последовательности *S* обозначим за $\varphi(S)$, а $\phi(S) = |\varphi(S)|$.

Обозначим за $\varphi(S,T) = \varphi(S) \cap \varphi(T)$ множество всех **общих подпоследовательностей** двух последовательностей *S* и *T* и аналогично $\phi(S,T) = |\varphi(S,T)|$.

Пусть $S =< X_1 \dots X_n >$ – последовательность, а *Y* – набор объектов. **Конкатенацией** набора объектов *Y* с последовательностью *S* $S \circ Y$ называется последовательность $< X_1 \dots X_n Y >$.

Меры сходства последовательностей без разрыва: префиксы

Пусть даны последовательности *S* и *T*.

Общие префиксы:

$$sim_{CP}(S,T) = \frac{|LCSP(S,T)|}{\max\{|S|,|T|\}}$$

где *LCSP* – наибольший общий префикс последовательностей.

Меры сходства последовательностей без разрыва: подпоследовательности

Самая длинная общая подпоследовательность:

$$sim_{LCS}(S,T) = \frac{|LCS(S,T)|}{\max\{|S|,|T|\}}$$

Все общие подпоследовательности:

$$sim_{ACS}(S,T) = \frac{2 \cdot \sum_{k \le l} \phi(S,T,k)}{l(l+1)}$$

$$l = \max\{|S|,|T|\}$$

k – длина общей подпоследовательности, $\phi(S,T,k)$ – количество общих подпоследовательностей *S* и *T* без разрывов длины *k*.

Добавление новых префиксов

Предположим, что мы хотим расширить данную последовательность $S =< X_1, \dots, X_n >$ набором объектов *Y*.

Рассмотрим отношение между общими префиксами до конкатенации $\gamma(S, T)$ и после $\gamma(S \circ Y, T)$. Существуют два варианта:

1) Если $Y[1] \neq T[p+1]$, где *p* – длина наибольшего префикса последовательностей *S* и *T*, тогда конкатенация набора объектов *Y* с последовательностью *S* не повлияет на набор $\gamma(S, T)$.

2) Если $Y[1] = T[p+1]$, где *p* – длина наибольшего префикса последовательностей *S* и *T*, тогда после конкатенации объектов *Y* с *S*, возникнут новые префиксы.

Добавление новых префиксов

Количество всех префиксов последовательностей после добавления *Y* будет равно количеству старых префиксов последовательностей и новых, полученных в результате конкатенации объектов *Y* с *S*.

$$|\gamma(S \circ Y, T)| = |\gamma(S, T)| + A(S, T, Y),$$

$$A(S, T, Y) = |S \otimes \gamma(D, Y)|,$$

$\gamma(S, T)$ – все общие префиксы последовательностей *S* и *T*,
$A(S, T, Y)$ – количество общих префиксов последовательностей, которые появляются после конкатенации,
$D = T - S$ – суффикс последовательности *T*, то есть, элементы последовательности *T* без элементов последовательности *S* где $T = S \circ D$,
$\gamma(D, Y)$ – все общие префиксы двух последовательностей *D* и *Y*,
$S \otimes \gamma(D, Y)$ – конкатенация *S* с каждым из объектов $\gamma(D, Y)$.

Добавление новых подпоследовательностей

Предположим, что мы хотим расширить данную последовательность $S =< X_1, \dots, X_n >$ набором объектов *Y*.

Рассмотрим отношение между общими подпоследовательностями до конкатенации $\beta(S, T)$ и после $\beta(S \circ Y, T)$. Существуют два варианта:

1) Если *Y* не имеет общих элементов с *T*, тогда конкатенация набора объектов *Y* с последовательностью *S* не повлияет на набор $\beta(S, T)$.

2) Если *Y* имеет общие элементы с *T*, тогда после конкатенации объектов *Y* с *S*, возникнут новые общие подпоследовательности без разрывов. Если *S* и *Y* имеют общие элементы, то некоторые подпоследовательности могут в итоге встретиться несколько раз, для этого необходимо ввести корректирующий член.

Добавление новых подпоследовательностей

Количество всех подпоследовательностей без разрыва после конкатенации с *Y* будет равно количеству старых подпоследовательностей без разрыва, новых, полученных в результате конкатенации объектов *Y* с *S*, и без учета повторяющихся подпоследовательностей без разрыва.

$$|\beta(S \circ Y, T)| = |\beta'(S,T)| + A(S,T,Y) - R(S,T,Y)$$

$$A(S,T,Y) = |\{P \circ Q \in \beta(S,T) \otimes \beta'(T,Y) \mid {}_{P \in \beta(S,T), Q \in \beta'(T,Y), P[m] \circ Q[1] \sqsubseteq S \circ Y}\}|$$

$$R(S,T,Y) = |\beta'(S,Y,T)|,$$

$\beta(S,T)$ – все общие подпоследовательности *S* и *T*,
$\beta'(S,T)$ – все непустые общие подпоследовательности *S* и *T*,
$\beta'(S,Y,T)$ – все непустые общие подпоследовательности *S*, *Y* и *T*,
$\beta(S,T) \otimes \beta'(T,Y)$ – конкатенация каждой подпоследовательности из $\beta(S,T)$ с каждой из подпоследовательностей $\beta'(S,T)$,
[*m*] – указывает на последний символ последовательности,
[1] – указывает на первый символ последовательности.

Использование специального ядра в SVM

- Для оценки качества классификации разными методами использовалось предсказание класса «пол» для тестовой выборки. Исходные данные были разделены на тренировочные и тестовые в соотношении 80/20.

- Классификация с использованием специальных функций ядра (CP, ACS, LCS) метода SVM на основе мер сходства последовательностей без разрыва.

Параметр	CP	ACS	LCS
Время обучения, с	400.97	1580.86	1544.21
Время предсказания, с	98.66	394.20	388.06
Общее время, с	499.62	1975.06	1932.27
Точность	0.648	**0.659**	0.490

Использование специального ядра в SVM

Классификация по признакам методом SVM.

Используем метод SVM с параметрами по умолчанию (функция ядра - RBF).

Параметр	Значение
Время обучения, с	3.62
Время предсказания, с	0.52
Общее время, с	4.14
Точность	**0.615**

Использование специального ядра в SVM

Классификация по последовательностям, по признакам и по взвешенной сумме вероятностей последовательностей и признаков.

Параметр	CP	ACS	LCS
Точность по последовательностям (SVM, custom kernel)	0.648	0.659	0.490
Точность по признакам (SVM default - RBF)	0.615	0.615	0.615
Точность по взвешенной сумме вероятностей: $P = \frac{As \cdot Ps + Af \cdot Pf}{As + Af}$	**0.678**	0.670	0.612

As — точность метода классификации по последовательностям,
Af — точность метода классификации по признакам,
Ps — вероятность, рассчитанная методом классификации по последовательностям,
Pf — вероятность, рассчитанная методом классификации по признакам.

Использование специального ядра в SVM

Классификация по признакам и последовательностям в качестве дополнительного признака.

Параметр \ Классификация	По последовательностям	По признакам	По последовательностям и признакам
Количество последовательностей	6626	0	6626
Количество уникальных последовательностей максимальной длины (количество значений признака)	1228	0	1228
Количество исходных признаков	0	5	5
Количество сгенерированных признаков (из последовательностей)	1	0	1
Время обучения, с	4.80	3.62	5.79
Время предсказания, с	0.79	0.52	0.91
Общее время, с	5.59	4.14	6.70
Точность	**0.675**	0.615	**0.716**

Нейросетевые алгоритмы

Для классификации использовался классификатор на основе модели перцептрона MLPClassifier из пакета sklearn.

Параметр \ Классификация	По последова-тельностям (как признакам)	По признакам	По последова-тельностям и признакам
Количество последовательностей	6626	0	6626
Количество уникальных последовательностей максимальной длины (количество значений признака)	1228	0	1228
Количество исходных признаков	0	5	5
Время обучения, с	5.52	8.11	8.73
Время предсказания, с	0.02	0.05	0.05
Общее время, с	5.54	8.16	8.78
Точность	0.524	0.624	**0.640**

Нейросетевые алгоритмы

Проведена классификация по последовательностям с использованием рекуррентных нейронных сетей с помощью ПО Keras и Tensorflow.

Метод классификации	По последовательностям			По признакам	По последова-тельностям и признакам
	Слои нейросети (количество)				
Параметр	SimpleRNN(1) Dense(1)	GRU(1) Dense(1)	LSTM(1) Dense(1)	Dropout(1) Dense(3)	SimpleRNN(1) Dense(5) Dropout(3)
Количество событий в последовательностях (максимум)	8	8	8	0	8
Количество признаков	0	0	0	5	5
Время обучения, с	168.80	452.27	585.73	348.49	418.05
Время предсказания, с	2.28	3.38	3.93	0.68	1.36
Общее время, с	171.07	455.66	585.73	349.17	419.40
Точность	0.676	0.672	0.675	0.626	**0.754**

Сравнение методов классификации

Методы	По последовательностям	По признакам	По последовательностям и признакам
SVM			
Специальная функция ядра CP	0.648	0.615	**0.678**
Специальная функция ядра ACS	0.659	0.615	0.670
Специальная функция ядра LCS	0.490	0.615	0.612
Специальное преобразование последовательностей в признак	0.675	0.615	**0.716**
Нейросетевые алгоритмы			
Перцептрон sklearn.neural_network.MLPClassifier	0.524	0.624	0.640
Рекуррентные нейронные сети (Keras, Tensorflow)			
SimpleRNN, Dense	0.676	0.626	**0.754**
GRU, Dense	0.672	0.626	
LSTM, Dense	0.675	0.626	
Деревья решений, временное кодирование			**0.661**
Марковские модели			
Hidden Markov Models, HMM			0.582
Предобработка последовательностей с помощью HMM и затем SVM (RBF kernel)			**0.620**

Чистка данных

Институтом демографии НИУ ВШЭ была предоставлена база данных, содержащая результаты опроса 16804 человек. Для каждого человека указаны даты наступления значимых событий в их жизни, такие как:

- Партнерство
- Брак
- Расставание
- Развод

В данных содержится большое количество разнообразных ошибок и уникальных случаев опечаток. Например: неправильная очередность порядка дат, длительность событий, завершенность событий и так далее.

Разработан и осуществлен универсальный алгоритм по чистке данных. Программы на языке Microsoft VBA (19 макросов общей длиной 3036 строк).

Выводы

- Лучшие результаты классификации получены с помощью специального преобразования последовательностей в признак в методе SVM и с использованием рекуррентной нейронной сети.
- Эти методы лучше других выявляют закономерности в последовательностях событий.
- Данная работа, в том числе разработанные программы, могут быть использованы при анализе различных последовательностей.

Спасибо за внимание!