

Построение метода динамического выравнивания многомерных
временных рядов, устойчивого к локальным колебаниям
сигнала.

Кулагин Петр Андреевич

Московский физико-технический институт

Консультанты: Г. И. Глеб Моргачев, А. В. Гончаров

Эксперт: Стрижов В. В.

8 мая 2020 г.

Цель работы — построение многомерного алгоритма выравнивания

Задача

Нахождение оптимального алгоритма, устойчивого к колебаниям сигнала при близком расположении датчиков.

Методы решения

Использование L2 расстояния между сигналами.

Использование расстояния DTW между сигналами.

- ① *Parinya Sanguansat* Multiple Multidimensional Sequence Alignment Using General Dynamic Time Warping, 2020.
- ② *Skutkova Helena, Vitek Martin, Babula Petr, Kizek Rene, Provaznik Ivo* Classification of genomic signals using dynamic time warping, BMC Bioinformatics, 2007.
- ③ *ten Holt, Gineke and Reinders, Marcel and Hendriks, Emile* Multi-dimensional dynamic time warping for gesture recognition, Annual Conference of the Advanced School for Computing and Imaging, 2007
- ④ *Jörg P. Bachmann and Johann-Christoph Freytag* High Dimensional Time Series Generators. CoRR, 2018.

Постановка задачи

Классификация временных рядов

Рассматриваем множество временных рядов $\{X_i\}_{i=1}^n \in \mathbb{R}^k$ и метки классов $Y_i \in \{0, 1\}$

Требуется для $X \in \mathbb{R}^k$ предсказать класс.

Метод kNN. Используем функцию расстояния между 2 временными рядами $\rho(x, y)$ Ищем k ближайших к объекту x по метрике ρ .

Предсказание - самый частый класс.

$$\text{accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Требуется найти $\rho = \text{argmax}(\text{accuracy})$ в методе kNN.

Гипотеза порождения данных

Выборка порождена K датчиками-измерителями.

Каждый датчик имеет определённое положение в пространстве и поэтому значения сигналов у каждого датчика имеют жёсткую привязку не только ко времени, но и к расположению датчика.

Временной ряд

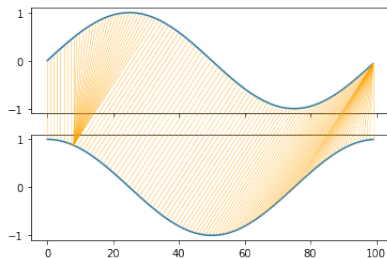
Временной ряд - последовательность измерений, произведенных в определённые промежутки времени.

$\mathbf{X} = \{X_1, \dots, X_n\}$ - временной ряд.

$X_i \in \mathbb{R}$ - одномерный временной ряд.

$X_i \in \mathbb{R}^K$ - многомерный временной ряд.

DTW - алгоритм, позволяющий посчитать расстояние между 2 временными рядами, устойчивый к сдвигам, растяжениям/сжатиям.



Определение

Рассмотрим два временных ряда Q и C разной длины: $Q = q_1, q_2, \dots, q_i, \dots, q_n$; $C = c_1, c_2, \dots, c_j, \dots, c_m$

Рассматриваем матрицу расстояний

1) $d_{ij} = \rho(q_i, c_j)$

2) Матрица трансформаций $D_{ij} = d_{ij} + \min(D_{i-1j}, D_{i-1j-1}, D_{ij-1})$;

3) Выравнивающий путь: строим путь трансформации W – минимизирует общее расстояние между Q и C $W = w_1, w_2, \dots, w_k, \dots, w_K$, где $w_k = (i, j)_k$, $d(w_k) = (q_i c_j)$

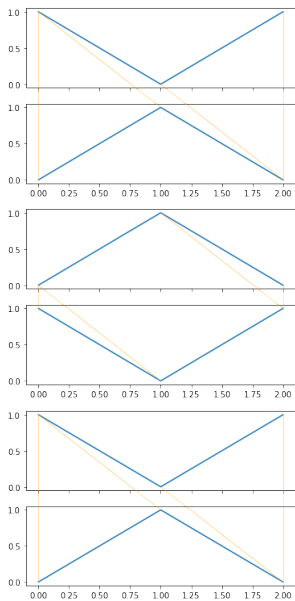
4) $\rho_{DTW}(Q, C) = \min \frac{\sum_{k=1}^K d(w_k)}{K}$

В случае многомерных временных рядов большое значение имеет расстояние между измерениями.

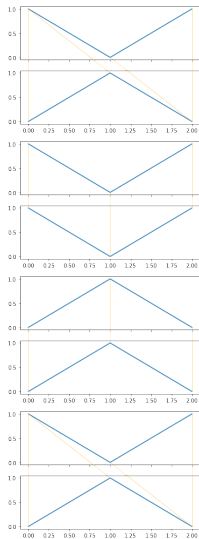
$$L2(q, c) = \sum_{i=1}^n (q_i - c_i)^2$$

$$DTW(q, c) = DTW(q, c) \text{ с } \rho = (x - y)^2$$

Пример выравнивания многомерных временных рядов с метрикой L2



Пример выравнивания многомерных временных рядов с метрикой DTW



Базовый эксперимент

Цель базового эксперимента

Показать, что существуют случаи, когда DTW расстояние между сигналами показывает лучший результат, но зачастую работает медленнее, чем L2 расстояние.

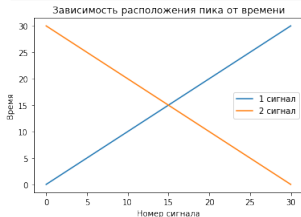
Генерация временных рядов

2 временных ряда, в которых сигнал "гуляет" в противоположных направлениях по датчикам.

Длина временного ряда - 30

DTW - 4

L2 - 60



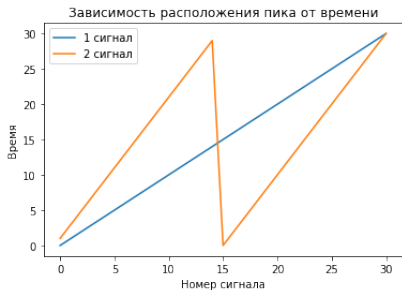
Генерация временных рядов

2 временных ряда, в котором одни датчики улавливают сигнал в 2 раза быстрее, например, стоят ближе к источнику.

Длина временного ряда - 30

DTW - 2

L2 - 60



Временной эксперимент

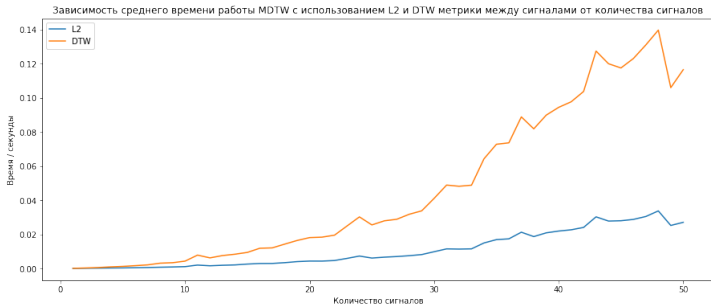
Подсчитаем среднее время работы DTW алгоритма с использованием L2 и DTW метрик.

Размерность сигнала: 3

Количество измерений: 50

Количество временных рядов: 20

Среднее время было посчитано 100 запусками на произвольной паре временных рядов.



Реальные данные

Данные - EEG Eye State

Результат - DTW метрика лучше в 9% случаев.

Полученные результаты

Было продемонстрировано значение метрики DTW для поиска расстояний между сигналами в многомерном случае. Однако затрачиваемое время на порядок медленнее L2 алгоритма. Поэтому стоит использовать L2 или DTW в зависимости от задачи и временных ограничений.

Дальнейшие исследования

- К сожалению, загруженный датасет оказался не с очень хорошим качеством данных (по крайней мере несколько первых строк), поэтому оба классификатора выдавали ассигасу = 1, что не позволяет корректно сравнивать 2 подхода. Нужно подробнее исследовать набор данных или найти новые.
- Найти конструкции построения датасета для классификации алгоритмически, когда классы, задаваемые DTW и L2 будут сильно различаться.