

Динамическое выравнивание многомерных временных рядов

Моргачев Г., Смирнов В., Липницкая Т.

Московский физико-технический институт

*Курс: Автоматизация научных исследований в машинном обучении
(практика, В.В. Стрижов)/2019*

Консультант: Гончаров А.

Выбор функции расстояния

при кластеризации и поиске паттернов во временных рядах

Цель работы

Исследовать влияние выбора функции расстояния между векторами на качество алгоритма DTW.

Проблема

При обобщении метода выравнивания временных рядов на многомерный случай остается открытым вопрос определения расстояния между парами векторов.

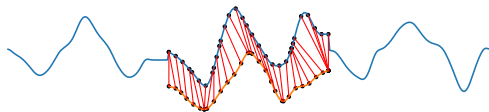
Метод решения

Получение оптимальной функции расстояния путем проведения эксперимента на задачах поиска паттернов и кластеризации.

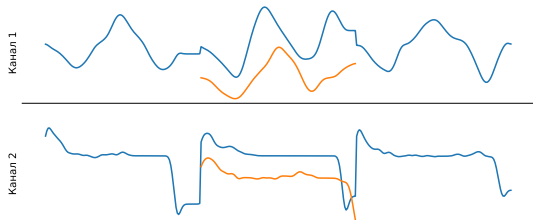
Многомерное выравнивание рядов

DTW: функция расстояния, учитывающая выравнивание рядов, относительно сдвигов и сжатий. Использует расстояние между точками.

Одномерный случай



Многомерный случай



- S_1^i, S_2^i - числа

- $\rho(S_1^i, S_2^i) = |x - y|$

- S_1^i, S_2^i - вектора

- $\rho(S_1^i, S_2^i) = \|S_1^i - S_2^i\|_1$

- $\rho(S_1^i, S_2^i) = \|S_1^i - S_2^i\|_2$

- $\rho(S_1^i, S_2^i) = \cos_dist(S_1^i, S_2^i)$

Постановка задачи

Дано множество \mathbf{S} l -мерных временных рядов длины n .

$S_i = \{s_i^1 \dots s_i^n\}$, $S_i \in \mathbf{S}$, $s_i^j \in \mathbb{R}^l$.

Задано множество функций расстояния между векторами:

$$\mathbf{R} = \{\rho : \mathbb{R}^l \times \mathbb{R}^l \rightarrow \mathbb{R}_+\},$$

$$\text{DTW}_\rho : \mathbf{S} \times \mathbf{S} \rightarrow \mathbb{R}_+.$$

Общая постановка задачи

Для задач кластеризации и поиска паттерна вводятся соответствующие функции качества Q_i^* . Рассматривается поиска оптимального ρ .

$$\rho_i = \underset{\rho}{\operatorname{argmax}} Q_i(\rho)$$

Задача кластеризации

Для всех $s_i \in \mathbf{S}$ задано $y_i \in \mathbb{Y}$ —множество меток классов.

Задаана матрица попарных расстояний:

$$D(\text{DTW}_\rho(\mathbf{S})) = ||D_{ij}||, \quad D_{ij} = \text{DTW}_\rho(s_i, s_j), \quad s_i, s_j \in S.$$

Модель кластеризации: $f : D \rightarrow Z^N$,

Z - множество меток кластеров.

Метод класстеризации

Иерархическая с функциями расстояния между кластерами:

- ❶ *complete*: $d(A, B) = \max_{a \in A, b \in B} (\text{dist}(a, b))$
- ❷ *weighted*: $d(A, B) = \frac{(\text{dist}(S, B) + \text{dist}(T, B))}{2}$, где кластер $A = S \cup T$
- ❸ *weighted*: $d(u, v) = \sum_{a \in A, b \in B} \frac{d(a, b)}{(|A| * |B|)}$

Функции качества кластеризации

$$Q_1(\rho) = \frac{1}{|Z|} \sum_{z \in Z} \max_y \frac{N_z^y}{N_z}, \quad Q_2(\rho) = \frac{1}{|Z|} \sum_{z \in Z} \max_y \frac{(N_z^y)^2}{N_z N^y}.$$

- N_z - количество элементов в кластере с меткой z .
- N^y - количество элементов в классе y .
- N_z^y - количество элементов класса y в классе z .

Задача поиска паттернов

Задан временной ряд S длины n , содержащий сегменты класса P .

P - временные ряда длины $m \ll n$.

Известны представители класса P , необходимо найти участки A , соответствующие данному классу.

$T = \{t_1, \dots, t_j\}$ - множество начал таких событий.

Участок найден, если пересечение с предполагаемым более 80% от m .

Функция качества поиска шаблонов

$$Q_3(DTW_{\rho}, A, P_k, T) = \frac{\sum_{i=1}^j [t_i - \text{найден}]}{j}.$$

Цель

Изучить зависимость качества кластеризации и поиска паттернов в зависимости от выбора функции расстояния между векторами, при различных методах определения расстояния между кластерами и получения среднего ряда.

Данные: класстеризация

- Размеченные данные ускорений акселерометра телефона: из 6 состояния человека, 3 канала, разбиты по 50 точек.

Данные: поиск паттернов

- Данные ECG: 4 состояния человека, 3 канала, разбиты на ряды по 206 точек.
- Написание букв: 20 символов, 3 канала, разбиты по 182 точки.

Результаты: поиск паттернов

ρ	average	characters			epi		
		Q_3	t	$t_{\text{no optim}}$	Q_3	t	$t_{\text{no optim}}$
L_1	DBA	0.857	2.123	11.767	0.744	14.335	13.064
	mean	0.894	2.361	11.614	0.744	13.541	13.912
L_2	DBA	0.818	1.551	11.499	0.687	12.342	13.205
	mean	0.854	1.527	10.164	0.687	14.199	12.738
ED	DBA	0.08	17.511	17.511	0.172	1.620	1.620
	mean	0.09	17.645	17.645	0.172	1.540	1.540

t , $t_{\text{no optim}}$ - время работы алгоритма с оптимизациями и без них.

Результаты: кластеризация

ρ	N_{clust}	Q_1			Q_2		
		<i>compl.</i>	<i>aver.</i>	<i>weight.</i>	<i>compl.</i>	<i>aver.</i>	<i>weight.</i>
L_1	24	0.506	0.585	0.638	0.273	0.376	0.449
	36	0.533	0.620	0.616	0.299	0.425	0.414
	48	0.556	0.639	0.631	0.330	0.443	0.431
L_2	24	0.488	0.622	0.626	0.270	0.417	0.425
	36	0.498	0.646	0.643	0.270	0.455	0.449
	48	0.534	0.648	0.653	0.270	0.455	0.462

Выводы