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 класстеризации и поиске паттернов в многомерных временных рядах

Цель работы

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

Проблема

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

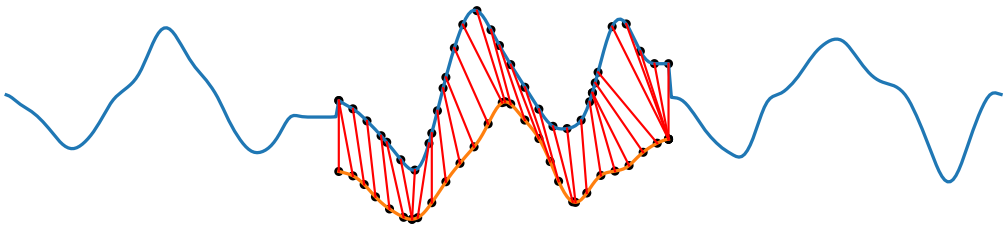
Метод решения

Рассмотрение задач класстеризации и поиска паттернов для нахождения функции расстония, позволяющей достичь на данных задачах наилучшего качества.

Многомерное выравнивание рядов

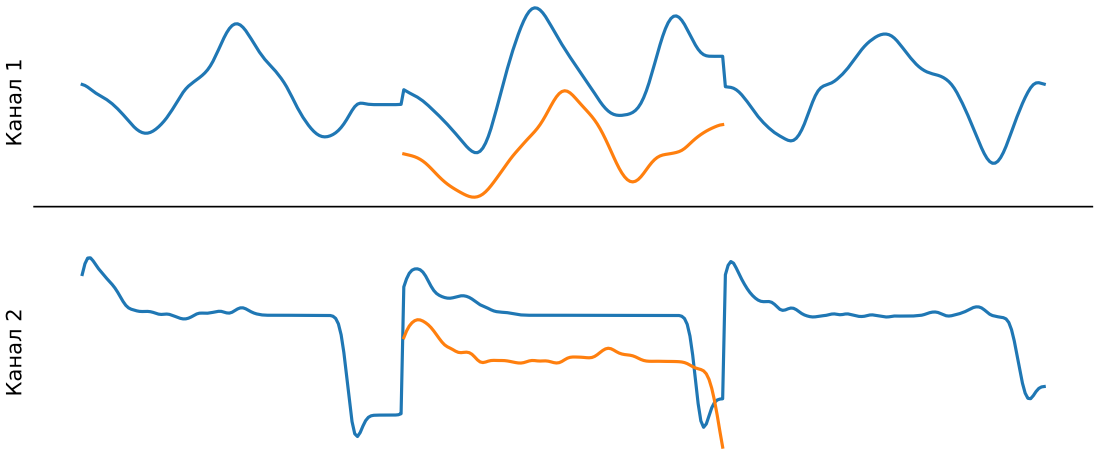
DTW: функция расстояния, учитывающая выравнивание рядов относительно сдвигов и сжатий. Использует расстояние между точками.

Одномерный случай



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

Многомерный случай



- S_1^i, S_2^i - вектора
- $\rho(S_1^i, S_2^i) = \|S_1^i - S_2^i\|_1$
- $\rho(S_1^i, S_2^i) = \|S_1^i - S_2^i\|_2$
- $\rho(S_1^i, S_2^i) = \cos\_\text{dist}(S_1^i, S_2^i)$

Литература

Одномерный случай:

1. Thanawin Rakthanmanon, Bilson Campana, Abdullah Mueen, Gustavo Batista, Brandon Westover, Qiang Zhu, Jesin Zakaria, and Eamonn Keogh. Searching and mining trillions of time series subsequences under dynamic time warping.
   2012

Многомерные обобщения:

1. Parinya Sanguansat. Multiple multidimensional sequence alignment using generalized dynamic time warping.
   2012
2. G.A. ten Holt, M.J.T. Reinders, and E.A. Hendriks. Multi-dimensional dynamic time warping for gesture recognition.
   2007

Постановка задачи

Дано множество $\boldsymbol{S}$ l-мерных временных рядов длины n.
$S_i = \{s_i^1 \dots s_i^n\},\ S_i \in \boldsymbol{S},\ s_i^j \in \mathbb{R}^l$.
Задано множество функций расстояния между векторами:

$$\mathrm{R} = \{\rho : \mathbb{R}^l \times \mathbb{R}^l \to \mathbb{R}_+\},$$

$$\mathrm{DTW}_\rho : \boldsymbol{S} \times \boldsymbol{S} \to \mathbb{R}_+.$$

Общая постановка задачи

Для задач кластеризации и поиска паттернов вводятся соответствующие функции качества Q_i*. Рассматривается поиск оптимального ρ.

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

Задача кластеризации

Для всех $S_i \in \boldsymbol{S}$ задано $y_i \in \mathbb{Y}$–множество меток классов.
Задана матрица попарных расстояний:

$$D(\mathrm{DTW}_\rho(\boldsymbol{S})) = ||D_{ij}||, \quad D_{ij} = \mathrm{DTW}_\rho(S_i, S_j), \quad S_i, S_j \in \boldsymbol{S}.$$

Модель кластеризации: $f : D \to Z^N$,
Z - множество меток кластеров.

Метод класстеризации

Иерархическая с функциями расстояния между кластерами:

1. *complete:* $d(A, B) = \max\limits_{a \in A, b \in B}(dist(a, b))$
2. *weighted:* $d(A, B) = \dfrac{(dist(S, B) + dist(T, B))}{2}$, где кластер $A = S \cup T$
3. *weighted:* $d(u, v) = \sum\limits_{a \in A, b \in B} \dfrac{d(a, b)}{(|A| * |B|)}$

Функции качества кластеризации

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

- 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 длинны n, содержащий сегменты класса $\boldsymbol{P}$.

$\boldsymbol{P}$ - временные ряда длины $m \ll n$.

Известны представители класса $\boldsymbol{P}$, необходимо найти участки S, соответствующие данному классу.

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

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

Функция качества поиска шаблонов

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

Эксперимент

Цель

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

Данные: класстеризация

- Размеченные данные ускорений акселерометра телефона: 6 состояний 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	t	$t_{\text{no optim}}$	Q	t	$t_{\text{no optim}}$
L_1	DBA	0.866	2.117	10.064	0.744	14.335	13.064
	mean	0.901	2.524	10.819	0.744	13.541	13.912
L_2	DBA	0.831	1.308	9.628	0.687	12.342	13.205
	mean	0.864	1.255	10.495	0.687	14.199	12.738
cos	DBA	0.805	3.221	13.650	0.687	12.342	13.205
	mean	0.819	3.142	14.285	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		
		compl.	*aver.*	*weight.*	*compl.*	*aver.*	*weight.*
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**

Результаты

Выводы

Во всех экспериментах связанных с поиском паттерном лучших результатов позволила достичь использование L_1 метрики. В эксперементах с кластеризацией, напротив, самой эффективной оказалось L_2 метрика.