Анализ кластеризации временных рядов

Юсупов Игорь

Московский физико-технический институт

Курс: Численные методы обучения по прецедентам (практика, В. В. Стрижов)/Группа 674, весна 2019

Цель исследования

Проблема

Кластеризация временных рядов с помощью функции расстояния DTW имеет высокую временную сложность

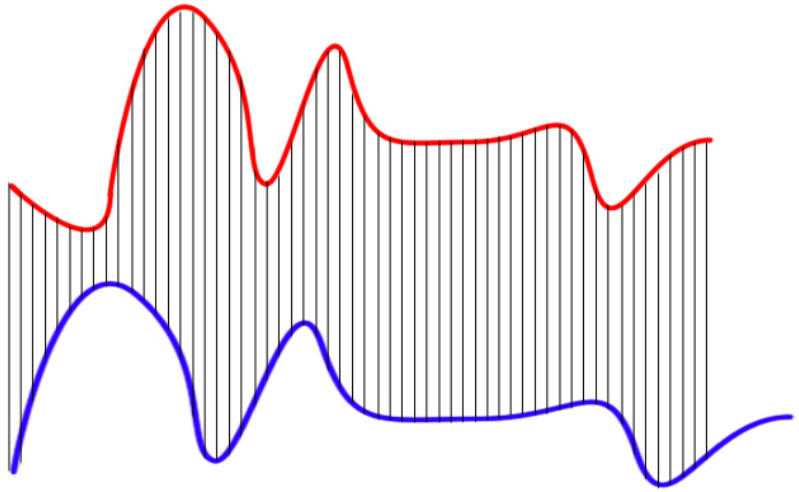
Цель работы

Сделать анализ зависимости качества кластеризации от метода кластеризации и исследовать возможность снижения временной сложности расчета матрицы расстояний при сохранении качества кластеризации с помощью оптимизированной функции расстояния DTW

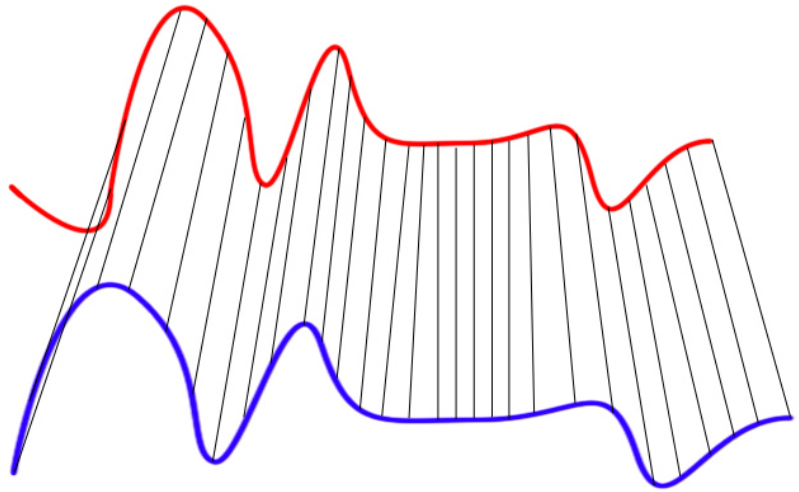
Обзор литературы

- Petitjean, François, Alain Ketterlin, and Pierre Gançarski. "A global averaging method for dynamic time warping, with applications to clustering."Pattern Recognition 44.3 (2011): 678-693.
- Vidal, Enrique, et al. "On the use of a metric-space search algorithm (AESA) for fast DTW-based recognition of isolated words."IEEE Transactions on Acoustics, Speech, and Signal Processing 36.5 (1988): 651-660.
- Assent, Ira, et al. "Anticipatory DTW for efficient similarity search in time series databases."Proceedings of the VLDB Endowment 2.1 (2009): 826-837

Выравнивание временных рядов



Euclidean Matching



Dynamic Time Warping Matching

Постановка задачи

Дано

$\mathbf{D} = \{(\mathbf{s_i}, y_i)\}_i^m$, где $\mathbf{s_i} \in \mathbb{S}$ – множество временных рядов, а $y_i \in \mathbb{Y}$ – множество идентификаторов кластера.
Модель кластеризации – функция a: $\mathbf{D} \to Y$

Мера сходства

В задаче число кластеров известно и равно k. Обозначим Y_1 и Y_2 как множества идентификаторов кластера, известные изначально и определенные алгоритмом кластеризации соответственно. Тогда вводят следующий критерий:

$$\mathrm{Sim}(Y_1, Y_2) = \frac{1}{k} \sum_{i=1}^{k} \max_{1 \le j \le k} \mathrm{Sim}(Y_{1i}, Y_{2j})$$

$$\text{где } \mathrm{Sim}(Y_{1i}, Y_{2j}) = \frac{2|Y_{1i} \cap Y_{2j}|}{|Y_{1i}| + |Y_{2j}|}$$

Постановка задачи

Функция расстояния

В данной задаче функциями расстояния являются элементы из множества оптимизаций функции расстояния DTW. Обозначим это множество **F**

Итоговая задача оптимизации

В работе мы должны определить оптимальный алгоритм кластеризации и функцию расстояния, то есть решаемая задача:

$$\arg\max_{a,F} \mathrm{Sim}(Y_1,\ Y_2)$$

Вычислительный эксперимент

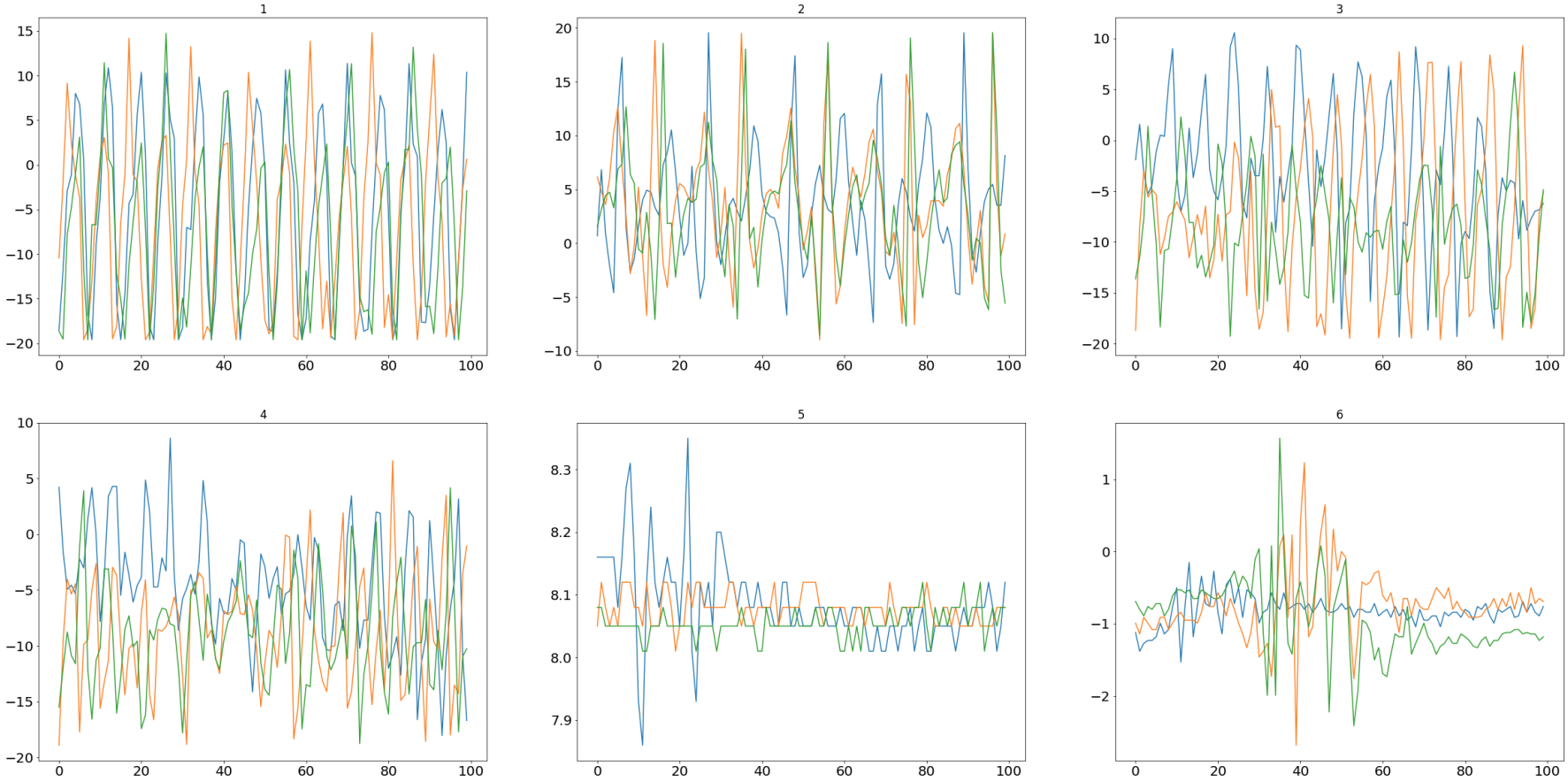
Цель

Оценить качество кластеризации, используя различные методы и оптимизации функции расстояния

Данные

Данные представляют собой измерения акселерометра, использующегося для идентификация действия человека в конкретный момент времени. Каждое движение имеет метку.

Визуализация данных



Примеры временных рядов акселерометра для разных видов физической активности

Вычислительный эксперимент

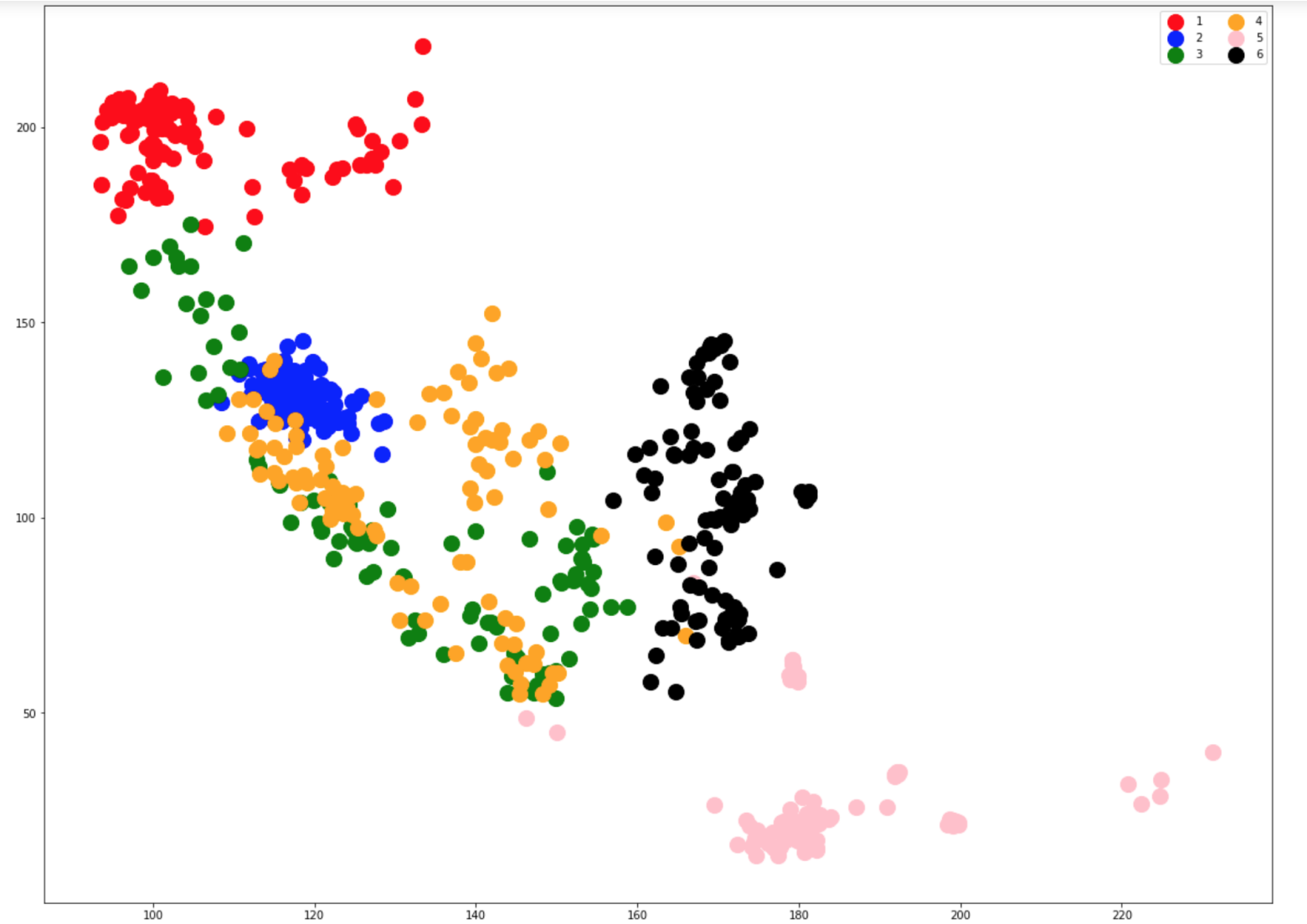


Диаграмма рассеяния для признаков 1 и 5

Результаты

В таблицах приведены результаты вычислительного эксперимента. Проведена кластеризация отдельно для пяти выборок, каждая из которых содержит по n классов

DBA	n	DBSCAN	HIERARCHY	SPECTRAL CL.
	2	1.000	0.800	1.000
	3	1.000	0.800	1.000
	4	0.750	0.800	0.800
	5	0.640	0.720	0.660
	6	0.800	0.630	0.766

Результаты

fastDTW	n	DBSCAN	HIERARCHY	SPECTRAL CL.
	2	1.000	0.800	1.000
	3	1.000	0.800	1.000
	4	0.725	0.800	0.775
	5	0.620	0.720	0.740
	6	0.520	0.630	0.733

Результаты

LB	n	DBSCAN	HIERARCHY	SPECTRAL CL.
	2	1.000	0.650	1.000
	3	0.930	0.766	1.000
	4	0.850	0.725	0.850
	5	0.680	0.600	0.700
	6	0.733	0.666	0.816

Выводы

- Проведена кластеризация с помощью функции расстояния DTW.
- Результаты совпали с ожиданием.
- Несмотря на меньшую точность, оптимизация LB существенно понижает временную сложность.
- Требуется исследование оптимальных параметров алгоритмов кластеризации для увеличения точности.