Кластеризация временных рядов*

Гончаров А. В.,Юсупов И. Р
yusupov.ir@phystech.edu, alex.goncharov@phystech.edu
МФТИ

В работе исследуется задача метрической кластеризации временных рядов из носимых устройств для выделения общих и характерных паттернов сигнала для движений человека различного типа. Используется устойчивая к линейным и нелинейным деформациям временной шкалы функция расстояния, основанная на методе динамического выравнивания шкалы времени DTW. При помощи этой функции расстояния вычисляется матрица расстояний между объектами исходной выборки, которая и задает новое признаковое пространство для временного ряда. В новом признаковом пространстве решается задача кластеризации набора временных рядов различными методами. Основной целью статьи является анализ зависимости качества кластеризации от метода кластеризации для указанного признакового пространства. Исследуется возможность снижения временной сложности расчета матрицы расстояний при сохранении качества кластеризации с помощью оптимизированной функции расстояния DTW. Вычислительный эксперимент проводится на открытых данных акселерометра мобильного устройства.

Ключевые слова: *временные ряды, многомерные временные ряды, DTW.*

Введение

Одной из актуальных задач анализа данных является задача кластеризации временных рядов. Такого рода задачи возникают при построении моделей объектов в трудноформализуемых областях исследований, например в медицине, когда требуется дать описание типичных групп пациентов со сходной динамикой развития заболевания на основе данных об изменениях клинических показателей и диагностических признаков. Типизация пациентов позволяет, в частности, разрабатывать методики лечения, оптимальные для каждой группы.

В задаче кластерного анализа требуется разбить множество объектов, описываемых набором некоторых переменных или матрицей попарных расстояний на кластеры так, чтобы критерий качества принял наилучшее значение. Критерий качества – функционал, зависящий от разброса объектов внутри кластера и расстояний между кластерами. Определение расстояния или меры различия между временными рядами имеет дополнительные трудности: ряды могут быть разной длины, состоять из разнотипных компонентов, иметь большую размерность. Кроме того, предполагается наличие зависимостей между наблюдаемыми характеристиками в различные моменты времени.

Евклидово расстояние имеет чувствительность к искажению по временной оси, поэтому для измерения расстояния между временными рядами используется функция расстояния DTW. [1]. Гибкость этого метода позволяет оценить сходство двух временных рядов, несмотря на фазовые сдвиги. DTW имеет вычислительную сложность $O(n^2)$, но тем не менее является лучшим известным решением для определения сходства между временными рядами. На сегодня предложено большое количество подходов для понижения вычислительной сложности: отбрасывание заведомо непохожих подпоследовательностей на основе оценки нижней границы расстояния [3, 2], индексирование [5], раннее прекращение заве-

Задачу поставил: Гончаров А. В. Консультант: Гончаров А. В.

домо нерезультативных вычислений [8]. В данной работе используется отбрасывание подпоследовательностей на основе оценки нижней границы расстояния (lower-bounding, LB). Это ускорение позволяет добиться вычислительной сложности $O(n)$[8]. Также используется оптимизация DTW Barycenter Averaging (DBA) и ускоренный DTW (fastDTW). Затем сравним результаты и выявим самый подходящий алгоритм кластеризации и оптимизацию DTW.

Постановка задачи

Имеется выборка $\mathbb{D} = \{(s_i, y_i)\}_i^m$, где $s_i \in \mathbb{S}$ – множество временных рядов, а $y_i \in \mathbb{Y}$ – множество идентификаторов кластера.

Опредедение 1. Алгоритм кластеризации — функция a: $X \rightarrow Y$

Временные ряды не являются статическими данными, поэтому обычные методы кластеризации для них не всегда работают. Кластеризацию временных рядов можно осуществить двумя способами:

1) Для необработанных данных подобрать функцию расстояния так, чтобы известные методы кластеризации работали.

2) Преобразовать данные в статические и использовать известные методы кластеризации.

В данной работе мы используем первый способ и для этого вводим функцию расстояния DTW. Качество метода кластеризации должна оцениваться некоторыми критериями. Выделяют две категории критериев в зависимости от того, известно ли количество кластеров или нет.

Пусть количество кластеров известно и равно k и пусть G и C – множества индентификаторов кластера, известные изначально и определенные алгоритмом кластеризации соответственно. Тогда вводят следующий критерий

Опредедение 2. Мера сходства:

$$Sim(G, C) = \frac{1}{k} \sum_{i=1}^{k} \max_{1 \leqslant j \leqslant k} Sim(G_i, C_j)$$
$$\text{где } Sim(G_i, C_j) = \frac{2|G_i \cap C_j|}{|G_i| + |C_j|}$$

Опредедение 3 Функция расстояния f(x, y) – функция от двух параметров, которая неотрицательна и удовлетворяет аксиомам тождества, симметрии и треугольника.

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

$$\underset{a,f}{\operatorname{argmax}} Sim(G, C)$$

Описание алгоритма

Для построения функции выравнивания и проверки её качества используются модель DTW (и её отптимазации).

Описание функции расстояния между объектами В данной работе в качестве метрического расстояния между объектами предлагается использовать строимость *пути наименьшей стоимости* между объектами.

Dynamic time warping - измерение расстояния между двумя временными рядами.

Задано два временных ряда, X длины m_1 и Y длины m_1.

$$X = x_1, x_2, ..., x_i, ..., x_{m_1}$$
$$Y = y_1, y_2, ..., y_j, ..., y_{m_2}$$
$$x_i, y_j \in \mathbb{R}^n$$

Требуется построить матрицу размера $m_1 \times m_2$ с элементами $D_{ij} = d(x_i, y_j)$, где d - выбранная метрика. Чтобы найти наибольшее соответсвие между рядами нужно найти выравнивающий путь W, который минимизирует расстояние между ними. W - набор смежных элементов матрицы D, $w_k = (i, j)_k$. $W = w_1, w_2, ..., w_k, ..., w_K$

$max(m_1, m_2) \leqslant K \leqslant m_1 + m_2 + 1$, где K-длина выравнивающего пути Выравнивающий путь должен удовлетворять следующим условиям:

1. $w_1 = (1, 1)$, $w_K = (m_1, m_2)$
2. $w_k = (a, b)$, $w_{k-1} = (a', b') : a - a' \leqslant 1$, $b - b' \leqslant 1$
3. $w_k = (a, b)$, $w_{k-1} = (a', b') : a - a' \geqslant 0$, $b - b' \geqslant 0$

Оптимальный выравнивающий путь должен минимизировать выравнивающую стоимость пути:

$$DTW(X, Y) = \sum_{k=1}^{K} w_k$$

Путь находится рекуррентно:
$\gamma(i, j) = d(q_i, c_j) + min(\gamma(i-1, j-1), \gamma(i-1, j), \gamma(i, j-1))$, где $\gamma(i, j)$ суммарное расстояние, $d(q_i, c_j)$ расстояние в текущей клетке.

DBA

fastDTW

LB

Базовый эксперимент

Алгоритмы кластеризации В данной работе используются следующие алгоритмы кластеризации: DBSCAN [9], K-Means [7] [4], Hierarchical Clustering [6], Agglomerative Clustering [6]
Цель эксперимента – оценить качество работы алгоритмов, используя функцию расстояния DTW, на небольшой выборке данных.
В ходе эксперимента были использованы данные акселерометра. Они представляли собой временные ряды длинною в 600 точек. Было выделено 120 рядов (по 20 из каждого класса) и сгенерировано 5 выборок, в каждой по n классов.

Результаты

fastDTW	n	DBSCAN	HIERARCHY	SPECTRAL CLUSTERING
	2	1.00	0.80	
	3	1.00	0.8	
	4	0.725	0.8	
	5	0.62	0.72	
	6	0.52	0.63	

DTW	n	DBSCAN	HIERARCHY	SPECTRAL CLUSTERING
	2	1.00	0.80	
	3	1.00	0.8	
	4	0.75	0.8	
	5	0.64	0.72	
	6	0.80	0.63	

*

Список литературы

[1] Donald J Berndt и James Clifford. «Using dynamic time warping to find patterns in time series.» в: *KDD workshop*. т. 10. 16. Seattle, WA. 1994, с. 359—370.

[2] Alessandro Camerra и др. «iSAX 2.0: Indexing and mining one billion time series». в: *2010 IEEE International Conference on Data Mining*. IEEE. 2010, с. 58—67.

[3] Hui Ding и др. «Querying and mining of time series data: experimental comparison of representations and distance measures». в: *Proceedings of the VLDB Endowment* 1.2 (2008), с. 1542—1552.

[4] John A Hartigan и Manchek A Wong. «Algorithm AS 136: A k-means clustering algorithm». в: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28.1 (1979), с. 100—108.

[5] Seung-Hwan Lim, Hee-Jin Park и Sang-Wook Kim. «Using multiple indexes for efficient subsequence matching in time-series databases». в: *International Conference on Database Systems for Advanced Applications*. Springer. 2006, с. 65—79.

[6] Daniel Müllner и др. «fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python». в: *Journal of Statistical Software* 53.9 (2013), с. 1—18.

[7] François Petitjean, Alain Ketterlin и Pierre Gançarski. «A global averaging method for dynamic time warping, with applications to clustering». в: *Pattern Recognition* 44.3 (2011), с. 678—693.

[8] Thanawin Rakthanmanon и др. «Searching and mining trillions of time series subsequences under dynamic time warping». в: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2012, с. 262—270.

[9] Thanh N Tran, Klaudia Drab и Michal Daszykowski. «Revised DBSCAN algorithm to cluster data with dense adjacent clusters». в: *Chemometrics and Intelligent Laboratory Systems* 120 (2013), с. 92—96.