

Кластеризация временных рядов*

Гончаров А. В., Юсупов И. Р

yusupov.ir@phystech.edu, alex.goncharov@phystech.edu

МФТИ

В данной работе исследуется кластеризация временных рядов с использованием алгоритма DTW. При использовании DTW кластеризация временных рядов имеет большую временную сложность. Основной целью статьи является снижение временной сложности и кластеризация временных рядов с помощью оптимизированной функции расстояния DTW.

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

Введение

Одной из актуальных задач анализа данных является задача кластеризации многомерных 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. [2]. 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 [4, 3], индексирование [6], раннее прекращение заведомо нерезультативных вычислений [10]. В данной работе используется отбрасывание подпоследовательностей на основе оценки нижней границы расстояния (lower-bounding, LB). Это ускорение позволяет добиться вычислительной сложности $O(n)$ [10]. Затем сравним алгоритмы кластеризации и выявим самый подходящий для кластеризации временных рядов с помощью DTW. Известно, что не все алгоритмы кластеризации подходят. Например, k-means не подходит, так как этот алгоритм настаивает на кластеризации всех элементов, в то время как это мешает точности ввиду того, что временные ряды не являются статическими и некоторые элементы из набора данных вовсе не должны быть собраны в кластер [1], но есть оптимизация этого метода для DTW, которая называется RSTMF [7].

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

Также DBSCAN не подходит для DTW, так как DTW не является метрикой, то возникает сложность в индексации, особенно для многомерных рядов [1].

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.

Определение 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 идентификаторов кластера, известные изначально и определенные алгоритмом кластеризации соответственно. Тогда вводят следующий критерий

Определение 2. Мера сходства:

$$Sim(G, C) = \frac{1}{k} \sum_{i=1}^k \max_{1 \leq j \leq k} Sim(G_i, C_j)$$

где $Sim(G_i, C_j) = \frac{2|G_i \cap C_j|}{|G_i| + |C_j|}$

Рассмотрим теперь случай, когда количество кластеров неизвестно. Вводится множество P_k , которое обозначает множество всех кластеров, разбивающих множество временных рядов на k кластеров. Критерий определяющий лучшую среди возможных группировок:

$$P(C^*) = \min_{C_j \in C \in P_k} \sum_{j=1}^k p(C_j)$$

где $p(C) = \frac{1}{2w(C)} \sum_{X, Y \in C} w(X)w(Y)D(X, Y)$

$w(X)$ - вес элемента X , $w(C) = \sum_{X \in C} w(X)$ - вес всех элементов.

$D(X, Y)$ - функция расстояния между элементами.

Описание алгоритма

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

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_2 .

$$X = x_1, x_2, \dots, x_i, \dots, x_{m_1}$$

$$Y = y_1, y_2, \dots, y_j, \dots, y_{m_2}$$

$$x_i, y_j \in \mathbb{R}^n$$

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 соответствие между рядами нужно найти выравнивающий путь W , который минимизирует расстояние между ними. W - набор смежных элементов матрицы D , $w_k = (i, j)_k$. $W = w_1, w_2, \dots, w_k, \dots, w_K$

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

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

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.

Кроме того, выравнивающий путь ограничивают тем, насколько он может отклоняться от диагонали. Типичным ограничением является полоса Сако-Чиба, в которой говорится, что путь искривления не может отклоняться от диагонали больше, чем на определённый процент клеток.

Алгоритмы кластеризации В данной работе используются следующие алгоритмы кластеризации: DBSCAN [11], K-Means [9] [5], Hierarchical Clustering [8], Agglomerative Clustering [8]

Базовый эксперимент

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 длиной в 600 точек. Из них была сгенерирована выборка из 90 рядов. Каждые 30 рядов принадлежат определённому классу.

Результаты

	DBSCAN	K-Means	Hierarchical Clustering	Agglomerative Clustering
Sim	1.00	1.00	1.00	0.47

*

Список литературы

- [1] Nurjahan Begum и др. «Accelerating dynamic time warping clustering with a novel admissible pruning strategy». в: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2015, с. 49—58.
- [2] 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.
- [3] 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.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] Warissara Meesrikamolkul, Vit Niennattrakul и Chotirat Ann Ratanamahatana. «Shape-based clustering for time series data». в: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer. 2012, с. 530—541.
- [8] 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.
- [9] 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.
- [10] 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.
- [11] 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.