Динамическое выравнивание многомерных временных рядов*

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

В данной работе исследуется кластеризация многомерных временных рядов с использованием алгоритма DTW. При использовании DTW в многомерном случае возникает проблема определения функций расстояния между элементами временных рядов. Основной целью статьи является нахождение зависимости качества кластеризации от выбора этой функции расстояния. В связи с повышением размерности возникает вопрос эффективности и применимости DTW на многомерных рядах.

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 [5], раннее прекращение заведомо нерезультативных вычислений [7]. В данной работе используется отбрасывание подпоследовательностей на основе оценки нижней границы расстояния (lower-bounding, LB). Это ускорение позволяет добиться вычислительной сложности $O(n)$[7]. 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

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

элементов, в то время как это мешает точности ввиду того, что временные ряды не являются статическими и неоторые элементы из набора данных вовсе не должны быть собраны в кластер [1], но есть оптимизация этого метода для DTW, которая называется RSTMF [6]. 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.

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:

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

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:

$$P(C^*) = \min_{C_j \in C \in P_k} \sum_{j=1}^{k} p(C_j)$$
$$\text{где } p(C) = \frac{1}{2w(C)} \sum_{X,Y \in C} w(X) w(Y) D(X, Y)$$
$$w(X) \text{ - вес элемента } X,\ w(C) = \sum_{X \in C} w(X) \text{ - вес всех элементов.}$$
$$D(X, Y) \text{ - функция расстояния между элементами.}$$

*****

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] 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] 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.

[7] 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.