Динамическое выравнивание многомерных временных рядов*

Гончаров А. В., Моргачев Г. И., Смирнов В., Липницкая Т.
morgachev.gi@phystech.edu, smirnov.vs@phystech.edu, tanya.lipnizky@yandex.ru
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. В качестве прикладной задачи исследуется кластеризация размеченных данных о деятельности человека полученных с акселерометра. Оценка качества кластеризации производится при сравнении с результатами кластеризации на основе авторегрессионной модели и анализу распределения классов данных в полученных кластерах.

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

1 Введение

Для описания различных данных широко используются временные ряды. Чтобы найти их сходство вводится функция расстояния, однако стандартный поточечный подход не является информативным вследствие того, что ряды могут содержать общие паттерны, деформированные относительно временной оси: претерпевшие сдвиги либо сжатия [1]. Одним из способов решения этой проблемы является выравнивание временных рядов (DTW) [2] и его модификаций [3]. Этот подход в большом спектре задач позволяет достичь максимального качества среди его аналогов.

В работе рассматривается применения DTW для кластеризации в случае многомерных временных рядов. Использование DTW на подобных данных описано в [4], [5]. В работе [4] предлагается способ выравнивания многомерных рядов, основанный на нормализации исходных данных и нахождениии векторной нормы. В [5] рассматривается алгоритм, позволяющий выполнить выравнивание временных рядов между координатами. Многомерное DTW предполагает различные варианты выравнивания, такие как выравнивание относительно общей временной шкалы и между соответсвующими каналами.

В процессе работы алгоритма DTW происходит вычисление расстояний между точками сравниваемых рядов. Поскольку в многомерном случае координаты точек описываются векторами, на результат будет влиять выбор функций расстояния между ними. Исследование влияние выбора этих функций на качество кластеризации является главной особенностью этой работы. В работе используются функции расстояния порождённые L_1 и L_2 нормами.

Ещё одним стандартным подходом к нахождению сходства между рядами является сравнение представления рядов коэффициентами их регрессионных моделей. Полученная в ходе работы DTW кластеризация сравнивается кластеризацией на основе авторегрессионной модели.

* Работа выполнена при финансовой поддержке РФФИ, проект № 00-00-00000. Научный руководитель: Гончаров А. В. Задачу поставил: Гончаров А. В. Консультант: Гончаров А. В.

В статьях [6] [7] рассматриваются различные виды алгоритмов кластеризации временных рядов, среди которых неплохие результаты показывают варианты иерархической кластеризации. Данный вид кластеризации был выбран в качестве базового.

Данные [8] представляют собой измерения акселерометра некоторого носимого устройства, например мобильного телефона, находящегося в кармане человека, и используется для индентификации действия человека в конкретный момент времени. Данные разделены на 6 классов: ходьба, бег, подъём по лестнице, спуск по лестнице, сидение, лежание.

2 Постановка задачи

Пусть задано множество временных рядов $\mathbb{S} \subset \mathbb{R}^{l \times n}$, где l - количество каналов, n - длина временного ряд.

$\forall s_i \in \mathbb{S}$ задано $y_i \in \mathbb{Y}$ - множество меток классов.

Пусть есть множество функция расстояния между векторами R:

$$\mathrm{R} = \{\rho : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^+\}$$

Для каждой функции расстояния между векторами существует соответствующая функция расстояния между временными рядами:

$$g_\rho : \mathbb{S} \times \mathbb{S} \to \mathbb{R}^+$$

Возьмем выборку $S \subset \mathbb{S}, \quad |S| = N$

Определим матрицу попарных расстояния:

$$D(g_\rho(S)) = ||D_{ij}||, \quad D_{ij} = g_\rho(s_i, s_j), \quad s_i, s_j \in S$$

Определим кластеризатор:

$$f : D \to Z^N$$

Где Z - множество меток кластеров.

Пусть функция качества:

$$Q(f(D), S) = \frac{\sum\limits_{s_i, s_j \in S} \mathbb{I}(z_i = z_j \wedge y_i = y_j)}{N^2}$$

Тогда, решаемая задача:

$$Q(D(g_\rho(S), S) \to \max_\rho$$

3 Описание основных методов

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

3.1 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 n и Y длины m.

$$X = x_1, x_2, ..., x_i, ..., x_n$$
$$Y = y_1, y_2, ..., y_j, ..., y_m$$
$$x_i, y_j \in \mathbb{R}^n$$

Требуется построить матрицу размера $n \times m$ 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, $w_k = (i, j)_k$.

$W = w_1, w_2, ..., w_k, ..., w_K$

$max(n, m) \leqslant K \leqslant m + n + 1$, где K-длина выравнивающего пути

Выравнивающий путь должен удовлетворять следующим условиям:

1. $w_1 = (1, 1)$, $w_K = (n, m)$
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.

3.2 Описание алгоритма кластеризации

В качестве алгоритма кластеризации используется иерархическая кластеризация, который базируется на последовательном слияние ближайших кластеров. Рассматриваются различные функции 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:* $d(u, v) = \sum\limits_{a \in A, b \in B} \dfrac{d(a, b)}{(|A| * |B|)}$

3.3 Описание авторегрессионного подхода

Авторегрессия - представляет собой подход, в котором элемент временного ряда представляется линейной комбинацией некоторого числа прошлых элементов.

Пусть есть временной ряд $X = x_1, ..., x_n,\ x \in \mathbb{R}^n$. Размер окна авторегрессионной модели l.

В модели авторегрессии

$$x_k = \sum_{i=1}^{l} k_i \cdot x_{k-i},\ k_i \in \mathbb{R}$$

$\boldsymbol{k} = \{k_i\}$ - коэффициенты, обучаемые для наилучшего описания выборки.

В дальнейшей работе с временным рядом, набор коэффициентов используется как вектор описания ряда.

В нашей работе, на множестве векторов коэффициентов производится иерархическая классификация с L_2 расстоянием между векторами.

4 Эксперимент

В ходе эксперимента были использованы данные акселерометра мобильного телефона. Они представляли собой временные ряды длинной в 600 точек ускорений по осям X, Y, Z. Из них была сгенерирована выборка из 2048 рядов по 50 точек. Каждых из рядов принадлежал к одному из шести возможных классов. Данные были равномерно распределены по всем классам.

Проводилась кластеризация этих данных описаными методами. В качестве расстояния между векторами использовались L_1 и L_2 нормы. Так как в процессе получения данных были возможны различные положения телефона в кармане, кластеризация проводилась на 24, 36 и 48 кластеров.

4.1 Результаты

		имя метрики			имя метрики2		
ρ	n clust	*weighted*	*average*	*complete*	*weighted*	*average*	*complete*
L_1	24						
	36						
	48						
L_2	24						
	36						
	48						

5 Заключение

Литература

[1] Hui Ding, Goce Trajcevski, Peter Scheuermann, Xiaoyue Wang, and Eamonn Keogh. *Querying and mining of time series data: Experimental comparison of representations and distance measures*, volume 1, pages 1542–1552. 2 edition, 8 2008.

[2] Eamonn J. Keogh and Michael J. Pazzani. Derivative dynamic time warping. In *In SIAM International Conference on Data Mining*, 2001.

[3] Stan Salvador and Philip Chan. Toward accurate dynamic time warping in linear time and space. *Intell. Data Anal*, 11(5):561–580, 2007.

[4] G.A. ten Holt, M.J.T. Reinders, and E.A. Hendriks. Multi-dimensional dynamic time warping for gesture recognition. In *Thirteenth annual conference of the Advanced School for Computing and Imaging*, 2007.

[5] Parinya Sanguansat. Multiple multidimensional sequence alignment using generalized dynamic time warping. 2012.

[6] T. Warren Liao. Clustering of time series data—a survey. *Pattern Recognition*, 38(11):1857 – 1874, 2005.

[7] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah. Time-series clustering – a decade review. *Information Systems*, 53:16 – 38, 2015.

[8] Jennifer R. Kwapisz, Gary M. Weiss, and Samuel A. Moore. Activity recognition using cell phone accelerometers. *SIGKDD Explor. Newsl.*, 12(2):74–82, March 2011.