

ЛОКАЛЬНАЯ КЛАСТЕРИЗАЦИЯ ВРЕМЕННЫХ РЯДОВ*

А. В. Грабовой¹, В. В. Стрижов²

Аннотация: Данная работа посвящена поиску периодических сигналов во временном ряду, для распознавания физических действий человека с помощью акселерометра. Предлагается метод кластеризации точек временного ряда для поиска характерных периодических сигналов внутри временного ряда. Для построения признакового описания используется локальное снижение размерности фазового пространства при помощи метода главных компонент. Для оценки близости двух периодических сигналов рассматривается расстояние между базисными векторами, которые получены методом главных компонент. Используя матрицу попарных расстояний между точками временного ряда выполняется кластеризация точек временного ряда. Для анализа качества представленного алгоритма проводятся эксперименты на синтетических данных и данных полученных при помощи мобильного акселерометра.

Ключевые слова: машинное обучение; метод главных компонент; временные ряды; кластеризация; распознавание физической активности.

DOI: 00.00000/000000000000000

*Работа выполнена при поддержке РФФИ и правительства РФ.

¹Московский физико-технический институт, grabovoy.av@phystech.edu

²Вычислительный центр им. А. А. Дородницына ФИЦ ИУ РАН, strijov@ccas.ru

1 Введение

Стремительное развитие портативных мобильных устройств, таких как мобильные телефоны или разумных часов открывает большие возможности для анализа повседневной человеческой активности. Большинство устройств включает в себя большое количество датчиков, например акселерометры. В данной работе стоит цель локального распознавания и локальной разметки человеческой активности в течении некоторого времени. Некоторые сходства с данной задачей можно увидеть в работах по классификации временных рядов [7], а также задача об обнаружении периодов активности [8, 1].

Временные ряды это объекты сложной структуры, при классификации которых большую роль играет построение признакового пространства. Для этой цели возможно использование экспертного задания базовых функций и метода построения признаков на основе гипотезы порождения данных. В работе [2] рассматривается комбинированное признаковое описание на основе этих двух методов. В статье [3] рассматривается построение признаков и предлагается критерий избыточности выбранных признаков.

В данной работе рассмотрено две проблемы. Во-первых предлагается алгоритм локальной аппроксимации временного ряда при помощи метода главных компонент [6] для получения признакового описания временного ряда. Во-вторых рассматривается метрика в новом пространстве признакового описания. После получения расстояния между точками временного ряда используется метод кластеризации KMeans [5] для объединения точек которые соответствуют одному и тому же характерному сигналу внутри временного ряда.

Для решения данной задачи вводится ряд предположений о временном ряде. Предполагается, что периоды всех различных сигналов различаются не значительно, причем известен максимальный период сигналов и количество различных сигналов внутри временного ряда. Также предполагается, что класс сигнала во времени меняется не часто, а также что фазовые траектории сигналов из разных классов являются различными.

Проверка и анализ метода проводится на синтетической выборке, построенной при помощи обрезанного ряда Фурье с произвольными коэффициентами. Также качество кластеризации проводится на рядах с разной физической активностью человека полученных при помощи акселерометра.

2 Постановка

Задан временной ряд:

$$\mathbf{X} \in \mathbb{R}^{N \times 1}. \quad (2.1)$$

Пусть временной ряд состоит из последовательности сигналов из множества \mathcal{V} :

$$\mathbf{X} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_M], \quad (2.2)$$

где \mathbf{v}_i некоторый сигнал из множества возможных сигналов \mathcal{V} . Причем $\forall i$ выполняется или $\mathbf{v}_i = \mathbf{v}_{i-1}$ или $\mathbf{v}_i = \mathbf{v}_{i+1}$. Пусть множество \mathcal{V} удовлетворяет следующим свойствам:

$$|\mathcal{V}| = K, \quad \forall \mathbf{v} \in \mathcal{V} \quad |\mathbf{v}| \leq T, \quad (2.3)$$

где $|\mathcal{V}|$ мощность множества сигналов, а $|\mathbf{v}|$ длина сигнала.

Рассмотрим отображение:

$$a : x \rightarrow \{1, \dots, K\}, \quad (2.4)$$

где $x \in \mathbf{X}$ некоторая точка временного ряда.

Требуется, чтобы отображение удовлетворяло следующим свойствам:

$$\begin{cases} a(x_1) = a(x_2), & \text{если } \exists \mathbf{v} \in \mathcal{V} : x_1, x_2 \in v \\ a(x_1) \neq a(x_2), & \text{если } \nexists \mathbf{v} \in \mathcal{V} : x_1, x_2 \in v \end{cases}$$

3 Кластеризация точек

Рассмотрим фазовую траекторию ряда \mathbf{X} :

$$\mathcal{H} = \{\mathbf{h}_t | \mathbf{h}_t = [x_{t-T}, x_{t-T+1}, \dots, x_t], \quad T \leq t \leq N\}. \quad (3.1)$$

Фазовая траектория разбивается на фазовые подпространства из $2T$ векторов:

$$\mathcal{S} = \{\mathbf{s}_t | \mathbf{s}_t = [h_{t-2T}, h_{t-2T+1}, \dots, h_t], \quad 2T \leq t \leq N\}. \quad (3.1)$$

Каждое T -мерное подпространство s_t спроектируем на плоскость при помощи метода главных компонент. Получим представление базисных векторов плоскости, а также собственные числа, которые соответствуют данным базисным векторам каждого подпространства s_t в T -мерном

пространстве:

$$\mathcal{W} = \{\mathbf{W}_t | \mathbf{W}_t = [\mathbf{w}_t^1, \mathbf{w}_t^2]\}, \quad \mathcal{L} = \{\boldsymbol{\lambda}_t | \boldsymbol{\lambda}_t = [\lambda_t^1, \lambda_t^2]\}, \quad (3.3)$$

где $[\mathbf{w}_t^1, \mathbf{w}_t^2]$ и $[\lambda_t^1, \lambda_t^2]$ это базисные векторы и соответствующие им собственные числа плоскости построенной при помощи метода главных компонент для подпространстве s_t .

Рассмотрим расстояние между элементами \mathcal{W} :

$$\rho(\mathbf{W}_1, \mathbf{W}_2) = \max_{\{\mathbf{a}, \mathbf{b}, \mathbf{c}\} \subset \mathbf{W}_1 \cup \mathbf{W}_2} V(\mathbf{a}, \mathbf{b}, \mathbf{c}), \quad (3.4)$$

где $V(\mathbf{a}, \mathbf{b}, \mathbf{c})$ — объем параллелепипеда построенного на векторах $\mathbf{a}, \mathbf{b}, \mathbf{c}$.

$\rho(\mathbf{W}_1, \mathbf{W}_2)$ является метрикой, если дополнительно указать, что базисы соответствующие параллельным плоскостям не различимы.

Рассмотрим расстояние между элементами \mathcal{L} :

$$\rho(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) = \sqrt{(\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2)^\top (\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2)}. \quad (3.5)$$

$\rho(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2)$ является метрикой в пространстве \mathcal{L} .

Матрица попарных расстояний между базисными векторами для временного ряда \mathbf{X} :

$$\mathbf{M}_c = [0, 1]^{N \times N}. \quad (3.6)$$

Матрица попарных расстояний между собственными значениями для временного ряда \mathbf{X} :

$$\mathbf{M}_l = [0, 1]^{N \times N}. \quad (3.7)$$

Используя выражения (3.4-7) определим расстояние между двумя точками t_1, t_2 временного ряда:

$$\rho(t_1, t_2) = \rho(\mathbf{W}_1, \mathbf{W}_2) + \rho(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2), \quad \mathbf{M} = \mathbf{M}_l + \mathbf{M}_c, \quad (3.8)$$

где $\rho(t_1, t_2)$ является метрикой, как сумма двух метрик. Матрица \mathbf{M} является матрицей попарных расстояний между двумя точками временного ряда.

Используя матрицу попарных расстояний \mathbf{M} выполним кластеризацию моментов времени временного ряда, получим следующее отображение:

$$a : x \rightarrow \{1, \dots, K\}, \quad (3.9)$$

где x некоторая точка временного ряда \mathbf{X} .

4 Эксперимент

Для анализа свойств предложенного алгоритма был проведен вычислительный эксперимент в котором кластеризация точек временного ряда проводилась используя матрицы попарных расстояний (3.6 – 8).

В качестве данных использовались две выборки временных рядов. Выборка "найти хорошую реальную выборку" это реальные временные ряды.

Синтетические временные ряды были построены при помощи обрзанного ряда Фурье с произвольными коэффициентами. Генерация данных состояла из двух этапов. На первом этапе генерировались короткие сигналы \mathbf{v} для построения множества \mathcal{V} . Вторым этапом генерации выборки \mathbf{X} является следующим случайным процессом:

$$\mathbf{X} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_M], \quad \begin{cases} \mathbf{v}_1 \sim \mathcal{U}(\mathcal{V}), \\ \mathbf{v}_i = \mathbf{v}_{i-1}, & \text{с вероятностью } \frac{3}{4}, \\ \mathbf{v}_i \sim \mathcal{U}(\mathcal{V}), & \text{с вероятностью } \frac{1}{4} \end{cases} \quad (4.1)$$

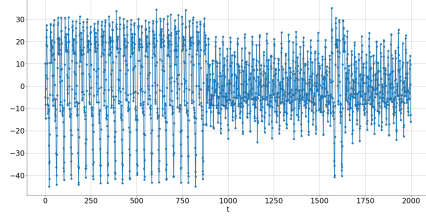
где $\mathcal{U}(\mathcal{V})$ — равномерное распределение на объектах из \mathcal{V} .

Таблица 1: Описание выборок

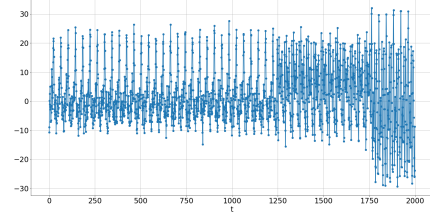
Выборка	N	K	T
Real			
Synthetic 1	2000	2	20
Synthetic 2	2000	3	20

Синтетические данные. На рис. 1 приведен пример синтетически построенных временных рядов. На рис. 1a показан пример ряда в котором количество сигналов $K = 2$, а длина каждого сигнала $T = 20$. На рис. 1b показан пример ряда в котором количество сигналов $K = 3$, а длина каждого сигнала $T = 20$.

На рис. 2 показан график зависимости значения сингулярных чисел локальной аппроксимации с течением времени. Значение сингулярных чисел, которые соответствуют первым двум главным компонентам значительно меняются с течением времени t .

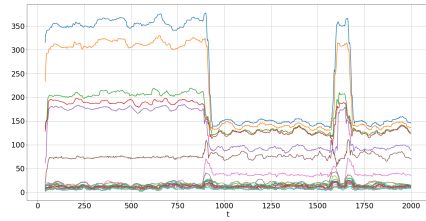


(a) Synthetic 1

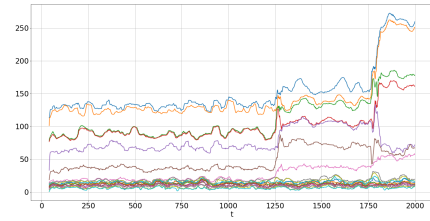


(b) Synthetic 2

Рис. 1: Пример синтетически построенных временных рядов

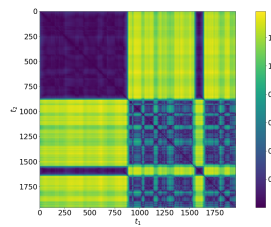


(a) Synthetic 1

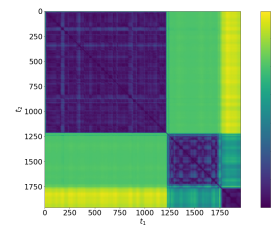


(b) Synthetic 2

Рис. 2: График зависимости значения сингулярных чисел метода главных компонент



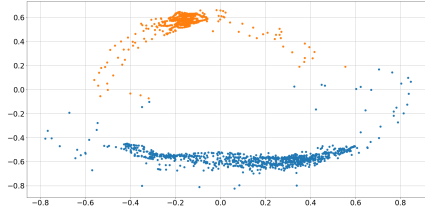
(a) Synthetic 1



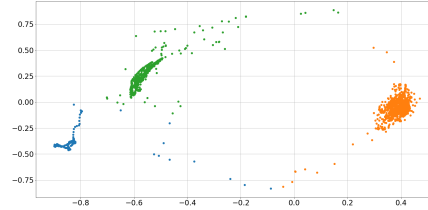
(b) Synthetic 2

Рис. 3: Матрица попарных расстояний \mathbf{M} между точками временного ряда

На рис. 3 проиллюстрированы матрицы попарных расстояний \mathbf{M} между построенными при помощи формулы (3.8). Используя матрицу попарных

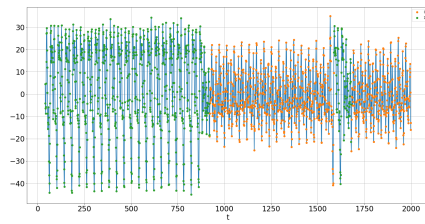


(a) Synthetic 1

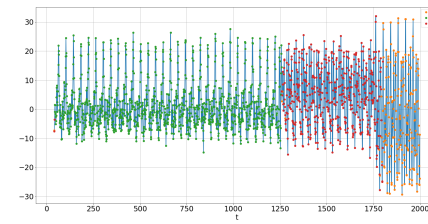


(b) Synthetic 2

Рис. 4: Проекция точек временного на плоскость при помощи матрицы попарных расстояний M



(a) Synthetic 1



(b) Synthetic 2

Рис. 5: Кластеризация точек временного ряда

расстояний и метод Multidimensional Scaling [4] визуализируем точки временного ряда на плоскости. На рис. 4 показана визуализация точек на плоскости и выполнена их кластеризация при помощи метода KMeans [5]. Иллюстрация кластеров точек временного ряда продемонстрирована на рис. 5.

5 Заключение

В работе рассматривалась задача поиска характерных периодических структур внутри временного ряда. Рассматривался метод основанный на локальном снижении размерности фазового пространства. Был предложен алгоритм поиска характерных сигналов, который основывается на методе главных компонент для локального снижения размерности, а так-

же на использовании некоторой функции расстояния между локальными базисами в каждый момент времени, которые интерпретировались как признакового описание точки временного ряда.

В ходе эксперимента было показано, что предложенный метод измерения расстояния между базисами хорошо разделяет точки которые принадлежат различным классам, что приводит к хорошей кластеризации объектов.

Предложенный метод имеет ряд недостатков связанных с большим количеством ограничений на временной ряд. Данные ограничения будут ослаблены в последующих работах.

Список литературы

- [1] *Y. G. Cinar and H. Mirisae* Period-aware content attention RNNs for time series forecasting with missing values // *Neurocomputing*, 2018. Vol. 312. P. 177–186.
- [2] *И. П. Ивкин, М. П. Кузнецов* Алгоритм классификации временных рядов акселерометра по комбинированному признаковому описанию. // *Машинное обучение и анализ данных*, 2015.
- [3] *V. V. Strijov, A. M. Katrutsa* Stresstes procedures for features selection algorithms. // *Schemometrics and Intelligent Laboratory System*, 2015.
- [4] *I. Borg, P. J. F. Groenen* Modern Multidimensional Scaling. — New York: Springer, 2005. 540 p.
- [5] *T. Kanungo, D. M. Mount et al* An Efficient k-Means Clustering Algorithm: Analysis and Implementation. 2000.
- [6] *Д. Л. Данилова, А. А. Жигловский* Главные компоненты временных рядов: метод "Гусеница". — Санкт-Петербургский университет, 1997.
- [7] *A. D. Ignatov, V. V. Strijov* Human activity recognition using quasiperiodic time series collected from a single tri-axial accelerometer. // *Multimedial Tools and Applications*, 2015.

- [8] *A. Olivares, J. Ramirez, J. M. Gorris, G. Olivares, M. Damas* Detection of (in)activity periods in human body motion using inertial sensors: A comparative study. // *Sensors*, 12(5):5791–5814, 2012.