Федеральное государственное автономное образовательное учреждение
высшего образования
«Московский физико-технический институт
(национальный исследовательский университет)»
Физтех-школа Прикладной Математики и Информатики
Кафедра интеллектуальных систем

Направление подготовки / специальность: 03.03.01 Прикладные математика и физика (бакалавриат)
Направленность (профиль) подготовки: Компьютерные технологии и интеллектуальный анализ данных

АНАЛИЗ СВОЙСТВ ЛОКАЛЬНЫХ МОДЕЛЕЙ В ЗАДАЧАХ КЛАСТЕРИЗАЦИИ ВРЕМЕННЫХ РЯДОВ

(бакалаврская работа)

Студент:
Грабовой Андрей Валериевич

_____
(подпись студента)

Научный руководитель:
Стрижов Вадим Викторович,
д-р физ.-мат. наук

_____
(подпись научного руководителя)

Консультант *(при наличии)***:**
_____
_____
(подпись консультанта)

Москва 2019

Аннотация

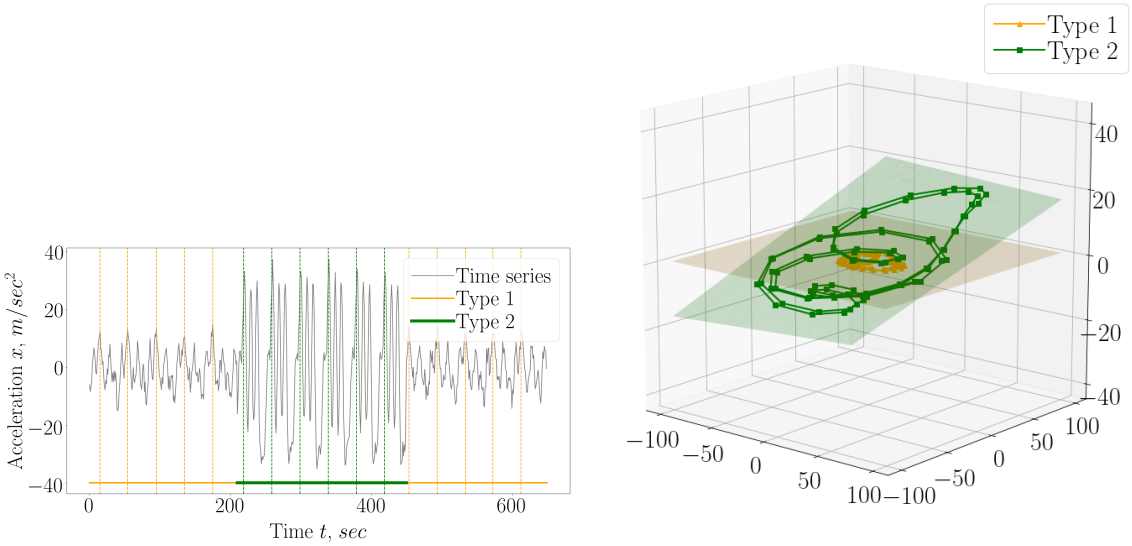
Данная работа посвящена анализу периодических сигналов во временных рядах с целью распознавания физических действий человека с помощью акселерометра. Предлагается метод кластеризации точек временного ряда для поиска характерных квазипериодических сегментов временного ряда. Временные ряды являются объектами сложной структуры, для которых не задано исходное признаковое описание. В качестве признакового описания точек временного ряда рассматриваются главные компоненты локальной окрестности фазовой траектории вблизи данной точки. Для оценки близости двух точек временного ряда вычисляется расстояние между данными точками в построенном пространстве признаков. При помощи матрицы попарных расстояний между точками временного ряда выполняется кластеризация данных точек. Для анализа качества представленного алгоритма проводятся эксперименты на синтетических данных и данных полученных при помощи мобильного акселерометра. Проводится эксперимент с поиском начала квазипериодических сегментов внутри каждого кластера.

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

Содержание

1 Введение

Анализ физической активности человека производится при помощи мобильных телефонов, разумных часов [1, 2]. Эти устройства используют акселерометр, гироскоп и магнитометр. Цель данной работы заключается в разметке и распознавании человеческой активности [3, 4, 5], а также поиска начала каждого действия [6]. Примерами одного сегмента действия служит шаг, шаг бега, приседание, прыжок и др. Исследуются последовательности, которые состоят не менее чем из двух подряд идущих сегментов, которые соответствуют одному и тому жу типу человеческой активности.



(a) (b)

Рис. 1: Временной ряд, с разметкой на кластеры: a) временной ряд с ассесорской разметкой на кластеры и выделением начала квазипериодического сегмента; b) проекция фазовых траекторий на первые две главные компоненты

Временные ряды — это объекты сложной структуры. При их классификации значимую роль играет модель построения признакового пространства. В данной работе объектом анализа и кластеризации является точка на оси времени. Решается задача класте-

ризации точек временного ряда. При *кластеризации* каждой точке временного ряда ставится в соответствие метка из конечного множества меток. Каждая метка соответствует одному характерному физическому действию. *Сегмент* это часть временного ряда, которая соответствует одному характерному физическому действию, например: шаг двумя ногами при ходьбе, или шаг двумя ногами при беге. Последовательность сегментов, которые соответствуют одному физическому действия образуют *цепочку* действий. Предполагается, что цепочка действий образует квазипериодическую последовательность значений временного ряда. Последовательность точек $\{b_t\}_{t=1}^{N}$ назовем *квазипериодической* с периодом T, если для всех t найдется Δ, такое что:

$$b_t \approx b_{t+T+\Delta}, \quad |\Delta| \ll T. \tag{1.1}$$

Пример кластеризации и разбиения ряда на сегменты показан на рис. 1a. Данный ряд разбит на два характерных физических действия, которые обозначаются Type 1 и Type 2. Также данный ряд содержит в себе две квазипериодические цепочки действий.

Решение задачи кластеризации состоит из двух этапов. Во-первых, для получения признакового описания временного ряда предлагается алгоритм локальной аппроксимации временного ряда при помощи метода главных компонент [11]. Под *локальной* аппроксимацией временного ряда подразумевается, что для признакового описания его точки используется не весь ряд, а только некоторая окрестность данной точки. В качестве признакового описания точки временного ряда рассматриваются две главные компоненты *сегмента фазовой траектории* в окрестности данной точки. На рис. 1b показаны две первые главные компоненты *фазовых траекторий*, а также проекция фазовых траекторий на эти компоненты. Они соответствуют разным физическим действиям, которые обозначаются Type 1 и Type 2, внутри одного временного ряда. Как видно плоскости, которые порождены данными главными компонентами не совпадают. Это говорит о том, что наблюдаются различные действия. Во-вторых, вводится функция расстояния в построенном пространстве признакового описания. Данная функция является расстояниям между двумя базиса-

ми некоторых подпространств внутри всего фазового пространства временного ряда. На рис. 1b данная функция является некоторым расстояниям между двумя плоскостями. Получив расстояния между точками временного ряда, выполним кластеризацию данных точек. Задача сегментации внутри каждого кластера решается при помощи метода, который рассмотрен в [6].

Для решения задачи кластеризации точек временного ряда вводятся предположения. Предполагается, что периоды различных сегментов различаются незначительно, причем известны минимальный и максимальный периоды сегмента и число различных сегментов внутри временного ряда. Также предполагается, что тип активности во времени не меняется часто, а также что фазовые траектории разных сегментов являются различными.

Проверка и анализ метода кластеризации проводится на синтетической и реальной выборках. Синтетическая выборка построенная при помощи суммы нескольких первых членов ряда Фурье со случайными коэффициентами. Эксперимент по сегментации временного ряда проводился на простых синусоидальных сигналах с произвольной амплитудой и частотой. Реальные данные получены при помощи мобильного акселерометра, который снимал показания во время некоторой физической активности человека.

2 Анализ литературы

В [1] рассматриается метод построения признакового описания на основе экспертно заданых порождающих функций. В [7] рассматривается метод построения признаков на основе гипотезы порождения данных. В [8] рассматривается комбинированное признаковое описание на основе данных методов. В [9] рассматривается проблема построение признакового пространства и предлагается критерий избыточности выбранных признаков.

Работа [6] является ближайшей работой по данной теме. Она заключается в поиске начала сегмента внутри квазипериодического сигнала, который состоит, только из одной цепочки действий. Этот

метод основан на исследовании фазового пространства, а именно поиска устойчивой гиперплоскости, которая делит фазовое пространство на две равные части. В качестве начала сегмента выбираются точки, которые находятся близко к данной гиперплоскости. В [6] предлагается выполнить проекцию фазового пространства на первые две главное компоненты, после чего провести устойчивую прямую, выделив начала каждого сегмента. Данный метод имеет недостаток в том, что позволяет находить начало только для временного ряда, который состоит из квазипериодического сигнала единственного типа.

Также близкой является работа [5]. Данная работа заключается в поиске периодической структуры внутри ряда при помощи модели LSTM с модифицированным механизмом Attention. Предполагается, что механизм Attention будет давать максимальное значение score в точках, которые удаленны от данной на целое количество периодов.

3 Постановка задачи кластеризации точек временного ряда

Задан временной ряд

$$\mathbf{x} \in \mathbb{R}^N, \tag{3.1}$$

где N число точек временного ряда. Он состоит из последовательности сегментов:

$$\mathbf{x} = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_M], \tag{3.2}$$

где $\mathbf{v}_i$ некоторый сегмент из множества сегментов $\mathbf{V}$, которые встречаются в данном ряде. Причем для всех i либо $[\mathbf{v}_{i-1}, \mathbf{v}_i]$ либо $[\mathbf{v}_i, \mathbf{v}_{i+1}]$ является цепочкой действий. Пусть множество $\mathbf{V}$ удовлетворяет следующим свойствам:

$$|\mathbf{V}| = K, \quad \mathbf{v} \in \mathbf{V} \ \ |\mathbf{v}| \leq T, \tag{3.3}$$

где $|\mathbf{V}|$ число различных действий в множестве сегментов $\mathbf{V}$, $|\mathbf{v}|$ длина сегмента, а K и T это число различных действий во временном ряде и длина максимального сегмента соответсвенно.

Рассматривается отображение

$$a : t \to \mathbb{Y} = \{1, \cdots, K\}, \tag{3.4}$$

где $t \in \{1, \cdots, N\}$ некоторый момент времени, на котором задан временной ряд. Требуется, чтобы отображение a удовлетворяло следующим свойствам:

$$\begin{cases} a(t_1) = a(t_2), & \text{если в моменты } t_1, t_2 \text{ совершается один тип действий} \\ a(t_1) \neq a(t_2), & \text{если в моменты } t_1, t_2 \text{ совершаются разные типы действий} \end{cases} \tag{3.5}$$

Пусть задана некоторая асессорская разметка временного ряда:

$$\mathbf{y} \in \{1, \cdots, K\}^N. \tag{3.6}$$

Тогда ошибка алгоритма a на временном ряде $\mathbf{x}$ представляется в следующем виде:

$$S = \frac{1}{N} \sum_{t=1}^{N} [y_t = a\,(t)], \tag{3.7}$$

где t — момент времени, y_t асессорская разметка t-го момента времени для заданого временного ряда.

4 Кластеризация точек

Рассмотрим фазовую траекторию временного ряда $\mathbf{x}$:

$$\mathbf{H} = \{\mathbf{h}_t | \mathbf{h}_t = [x_{t-T}, x_{t-T+1}, \cdots, x_t],\ T \leq t \leq N\}, \tag{4.1}$$

где $\mathbf{h}_t$ — точка фазовой траектории.

Информация об длине максимального сегмента T внутри временного ряда позволяет разбить фазовую траекторию на сегменты из $2T$ векторов:

$$\mathbf{S} = \{\mathbf{s}_t | \mathbf{s}_t = [\mathbf{h}_{t-T}, \mathbf{h}_{t-T+1}, \cdots, \mathbf{h}_{t+T-1}],\ T \leq t \leq N - T\}, \tag{4.2}$$

где $\mathbf{s}_t$ — это сегмент фазовой траектории. Данные сегменты имеют всю локальную информацию об временном ряде, так как содержит всю информацию на периоде до момента времени t и информацию о периоде после момента времени t.

В качестве признакового описания точки временного ряда t рассматриваются главные компоненты $\mathbf{W}_t$ для T-мерных сегментов $\mathbf{s}_t$. Сегмент $\mathbf{s}_t$ проекцируется на подпространство размерности два при помощи метода главных компонент $\mathbf{z}_t = \mathbf{W}_t \mathbf{s}_t$. Получаем:

$$\mathbf{W} = \{\mathbf{W}_t | \mathbf{W}_t = [\lambda_t^1 \mathbf{w}_t^1, \lambda_t^2 \mathbf{w}_t^2]\}, \quad \mathbf{\Lambda} = \{\boldsymbol{\lambda}_t | \boldsymbol{\lambda}_t = [\lambda_t^1, \lambda_t^2]\}, \tag{4.3}$$

где $[\mathbf{w}_t^1, \mathbf{w}_t^2]$ и $[\lambda_t^1, \lambda_t^2]$ это базисные векторы и соответствующие им собственные для сегмента фазовой траектории $\mathbf{s}_t$.

Для кластеризации точек временного ряда рассмотрим функцию расстояния между элементами $\mathbf{W}_{t_1}, \mathbf{W}_{t_2}$:

$$\rho\left(\mathbf{W}_1, \mathbf{W}_2\right) = \max\left(\max_{\mathbf{e}_2 \in \mathbf{W}_2} d_1\left(\mathbf{e}_2\right), \max_{\mathbf{e}_1 \in \mathbf{W}_1} d_2\left(\mathbf{e}_1\right)\right), \tag{4.4}$$

где $\mathbf{e}_i$ это базисный вектор пространства $\mathbf{W}_i$, а $d_i\left(\mathbf{e}\right)$ является расстоянием от вектора $\mathbf{e}$ до пространства $\mathbf{W}_i$.

В случае, когда все подпространства $\mathbf{W}_t$ имеют размерность два, расстояние $\rho\left(\mathbf{W}_1, \mathbf{W}_2\right)$ имеет следующую интерпретацию:

$$\rho\left(\mathbf{W}_1, \mathbf{W}_2\right) = \max_{\{\mathbf{a},\mathbf{b},\mathbf{c}\}\subset \mathbf{W}_1 \cup \mathbf{W}_2} V\left(\mathbf{a}, \mathbf{b}, \mathbf{c}\right), \tag{4.5}$$

где $\mathbf{W}_1 \cup \mathbf{W}_2$ это объединение базисных векторов первого и второго пространства, $V\left(\mathbf{a}, \mathbf{b}, \mathbf{c}\right)$ — объем параллелепипеда построенного на векторах $\mathbf{a}, \mathbf{b}, \mathbf{c}$, которые являются столбцами матрицы $\mathbf{W}_1 \cup \mathbf{W}_2$.

Рассмотрим расстояние между собственными числами:

$$\rho\left(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2\right) = \sqrt{\left(\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\right)^{\mathsf{T}} \left(\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\right)}. \tag{4.6}$$

Используя выражения (4.5-4.6) введем расстояние между двумя точками t_1, t_2 временного ряда, а также рассмотрим матрицу попарных расстояний $\mathbf{M}$ между точками данного ряда:

$$\rho\left(t_1, t_2\right) = \rho\left(\mathbf{W}_1, \mathbf{W}_2\right) + \rho\left(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2\right), \quad \mathbf{M} = \mathbb{R}^{N \times N}, \tag{4.7}$$

где матрица $\mathbf{M}$ является матрицей попарных расстояний между всеми парами точек t временного ряда $\mathbf{x}$. Используя матрицу попарных расстояний $\mathbf{M}$ выполним кластеризацию моментов времени t временного ряда (3.4):

5 Вычислительный эксперимент

5.1 Кластеризация точек временного ряда

Для анализа свойств предложенного алгоритма кластериизации был проведен вычислительный эксперимент в котором кластеризация точек временного ряда проводилась используя матрицы попарных расстояний (4.7).

В качестве данных использовались две выборки временных рядов, которые описаны в таблице 1. Выборка Physical Motion это реальные временные ряды полученные при помощи мобильного акселерометра. Синтетические временные ряды были построены при помощи нескольких первых слагаемых ряда Фурье со случайными коэффициентами из стандартного нормального распределения. Генерация данных состояла из двух этапов. На первом этапе генерировались короткие сегменты $\mathbf{v}$ для построения множества $\mathbf{V}$. Вторым этапом генерации выборки $\mathbf{x}$ является следующим случайным процессом:

$$\mathbf{x} = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_M] + \boldsymbol{\varepsilon}, \quad \begin{cases} \mathbf{v}_1 \sim \mathcal{U}(\mathbf{V}), \\ \mathbf{v}_i = \mathbf{v}_{i-1}, & \text{с вероятностью } \frac{3}{4}, \\ \mathbf{v}_i \sim \mathcal{U}(\mathbf{V}), & \text{с вероятностью } \frac{1}{4} \end{cases} \quad (5.1)$$

где $\mathcal{U}(\mathbf{V})$ — равномерное распределение на объектах из $\mathbf{V}$, а $\boldsymbol{\varepsilon}$ является шумом из нормального распределения.

Таблица 1: Описание временных рядов в эксперименте кластеризации точек временного ряда

Ряд, $\mathbf{x}$	Длина ряда, N	Число сегментов, K	Длина сегмента, T
Physical Motion 1	900	2	40
Physical Motion 2	900	2	40
Synthetic 1	2000	2	20
Synthetic 2	2000	3	20

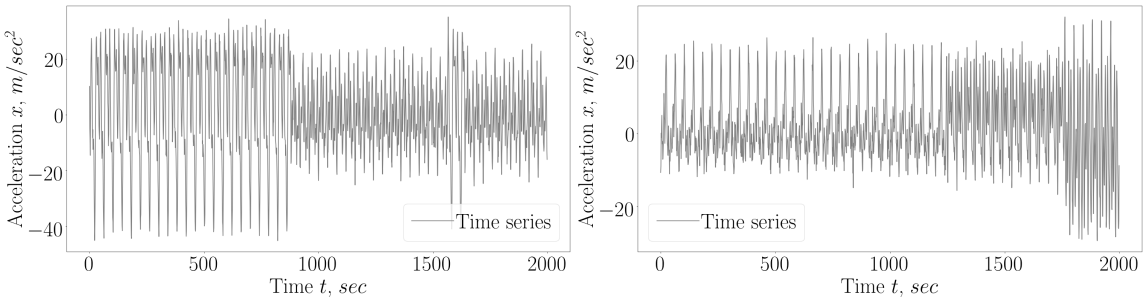


Рис. 2: Пример синтетически построенных временных рядов: a) для временного ряда Synthetic 1; b) для временного ряда Synthetic 2

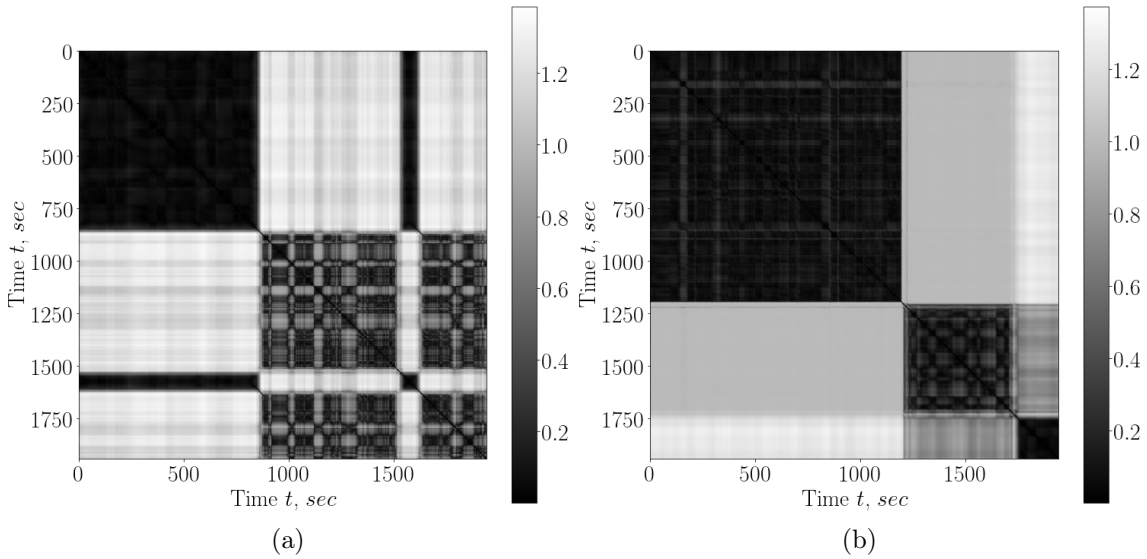


Рис. 3: Матрица попарных расстояний $\mathbf{M}$ между точками временного ряда: a) для временного ряда Synthetic 1; b) для временного ряда Synthetic 2

Синтетические данные. На рис. 2 приведен пример синтетических временных рядов. На рис. 2a показан пример ряда в котором число различных сегментов $K = 2$, а длина каждого сегмента $T = 20$. На рис. 2b показан пример ряда в котором число различных сегментов $K = 3$, а длина каждого сегмента $T = 20$.

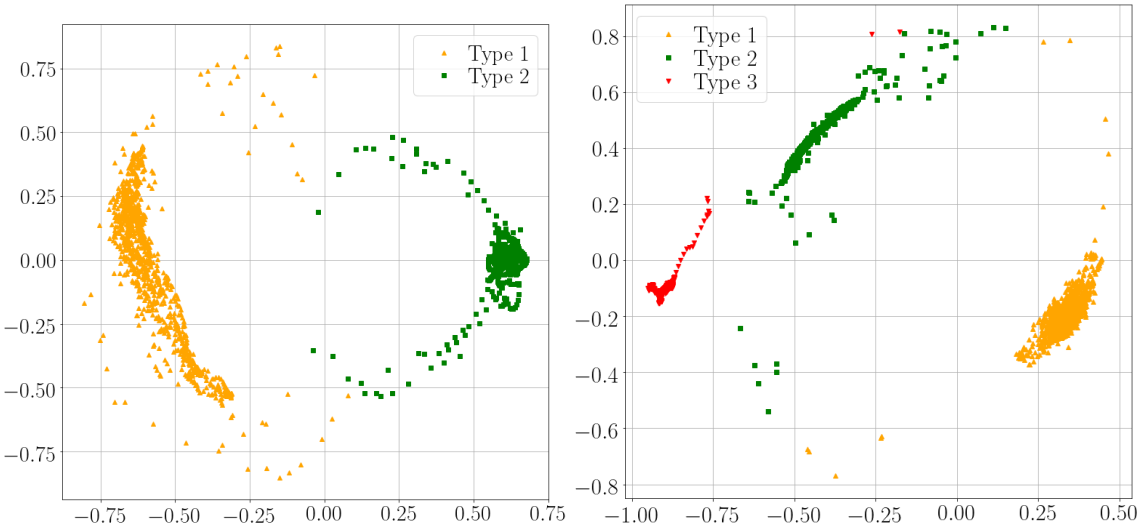


Рис. 4: Проекция точек временного ряда на плоскость при помощи матрицы попарных расстояний $\mathbf{M}$: a) для временного ряда Synthetic 1; b) для временного ряда Synthetic 2

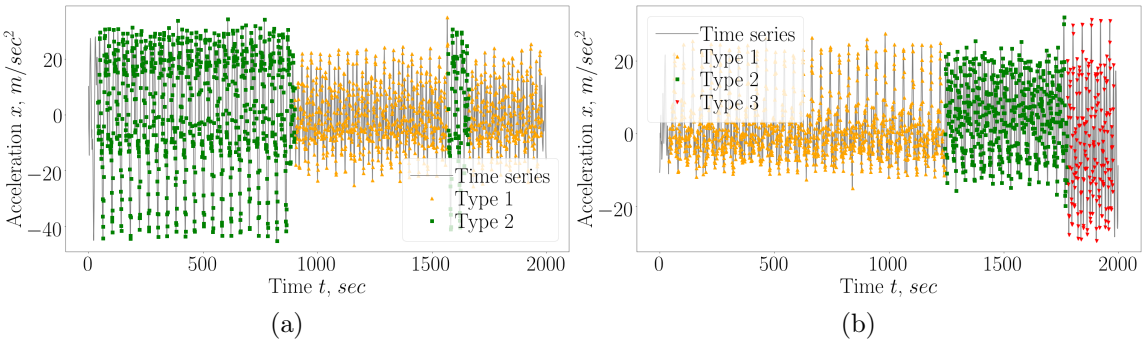


Рис. 5: Кластеризация точек временного ряда: a) для временного ряда Synthetic 1; b) для временного ряда Synthetic 2

Рис. 3 иллюстрирует матрицы попарных расстояний $\mathbf{M}$ между всеми парами точек t временного ряда, которые построены при помощи (4.7). Используя матрицу попарных расстояний и метод Multidimensional Scaling [10] визуализируем точки временного ряда на плоскости. На рис. 4 показана визуализация точек на плоскости и выполнена их

кластеризация при помощи метода иерархической кластеризации. Иллюстрация кластеров точек временного ряда продемонстрирована на рис. 5.

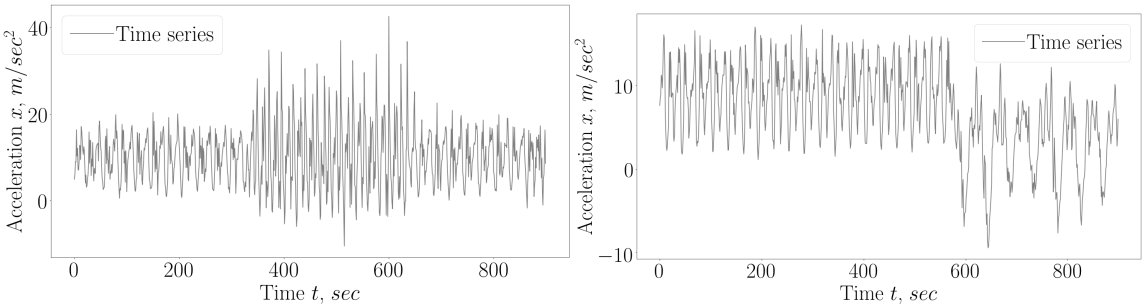


Рис. 6: Пример синтетически построенных временных рядов: a) для временного ряда Physical Motion 1; b) для временного ряда Physical Motion 2

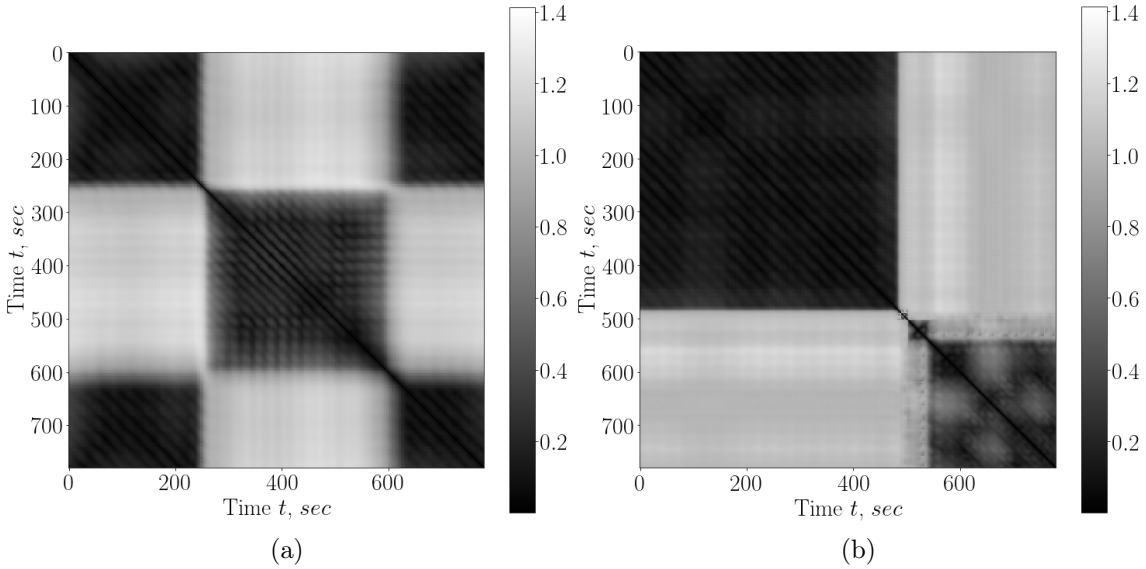


Рис. 7: Матрица попарных расстояний $\mathbf{M}$ между точками временного ряда: a) для временного ряда Physical Motion 1; b) для временного ряда Physical Motion 2

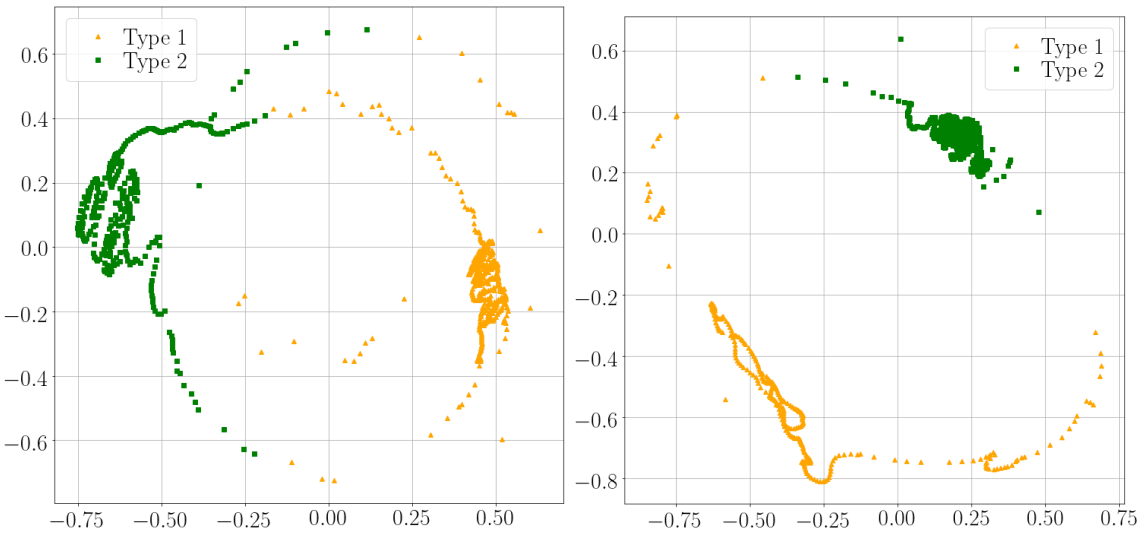


Рис. 8: Проекция точек временного на плоскость при помощи матрицы попарных расстояний $\mathbf{M}$: a) для временного ряда Physical Motion 1; b) для временного ряда Physical Motion 2

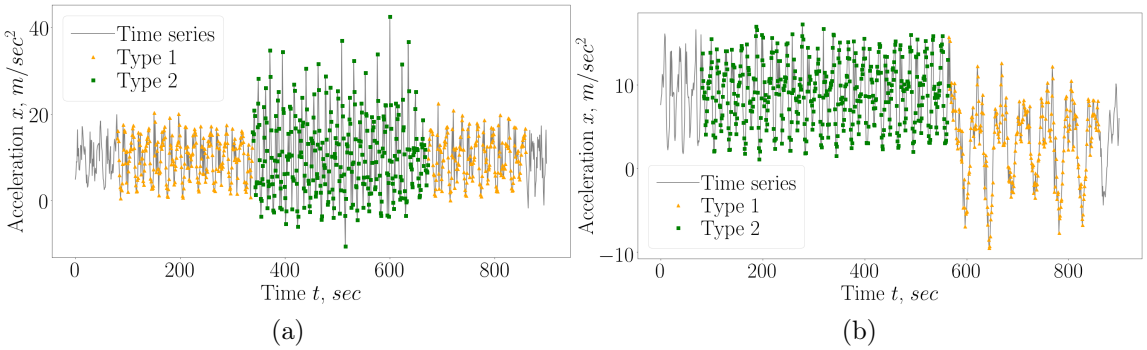


Рис. 9: Кластеризация точек временного ряда: a) для временного ряда Physical Motion 1; b) для временного ряда Physical Motion 2

Реальные данные. На рис. 6 приведен пример реальных временных рядов полученных при помощи взятия одной из координат мобильного акселерометра.

Рис. 7 иллюстрирует матрицы попарных расстояний $\mathbf{M}$ между всеми парами точек t временного ряда, которые построены при помо-

щи (4.7). Используя матрицу попарных расстояний и метод Multidimensional Scaling [10] визуализируем точки временного ряда на плоскости. На рис. 8 показана визуализация точек на плоскости и выполнена их кластеризация при помощи метода иерархической кластеризации. Иллюстрация кластеров точек временного ряда продемонстрирована на рис. 9.

5.2 Сегментация временный рядов

Сегментация временных рядов проводится на синтетических и реальных данных. Для данного эксперимента в качестве синтетического ряда рассматривается ряд построенный из двух синусов с произвольной частотой и амплитудой. Описание временных рядов, которые используются в данном эксперименте представлены в таблице 2.

Сегментация проводится при помощи метода, который представлен в работе [6]. Данный метод применяется для каждого действия внутри временного ряда по отдельности.

Таблица 2: Описание временных рядов в эксперименте сегментации временных рядов

Ряд, $\mathbf{x}$	Длина ряда, N	Число сегментов, K	Длина сегмента, T
Simple 1	1000	2	100
Physical Motion 2	900	2	40

Синтетические данные. На рис. 10 показан результат работы сегментации для временного ряда Simple 1. Данный алгоритм хорошо выделил начала сегментов. Также на рис. 10 показаны проекции фазовых пространств для обеих кластеров на их первые две главные компоненты.

Реальные данные. На рис. 11 показан результат работы сегментации для временного ряда Physical Motion 2. Данный алгоритм хо-

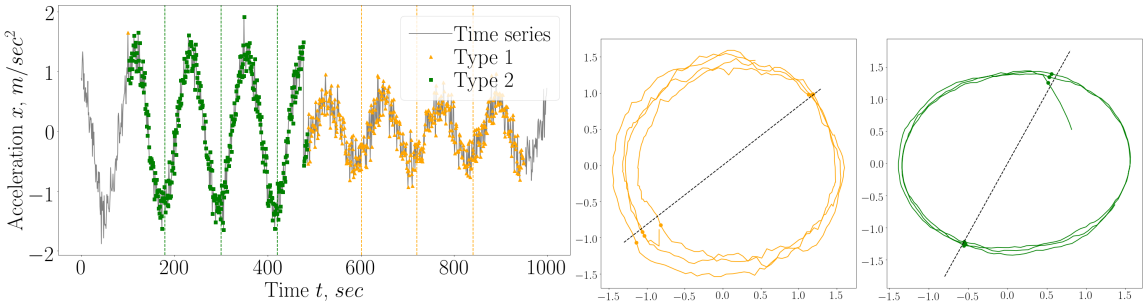
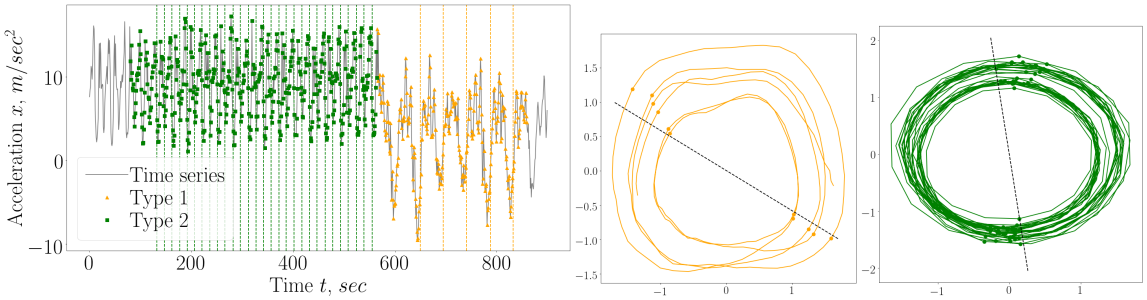


Рис. 10: Сегментация точек временного ряда Simple 1: a) сегментация временного ряда; b) проекция фазового пространства на первые две главные компоненты для первого кластера; c) проекция фазового пространства на первые две главные компоненты для второго кластера

рошо выделил начала сегментов для Type 1 и плохо для Type 2. Также на рис. 11 показаны проекции фазовых пространств для обеих кластеров на их первые две главные компоненты. Видно, что в случае проекции фазового пространства для части ряда, который относится к Type 2 получаем, что фазовая траектория имеет самопересечение внутри одного сегмента, что влечет нахождения ложного начала сегмента.



(a) (b) (c)

Рис. 11: Сегментация точек временного ряда Physical Motion 2: a) сегментация временного ряда; b) проекция фазового пространства на первые две главные компоненты для первого кластера; c) проекция фазового пространства на первые две главные компоненты для второго кластера

6 Заключение

Таблица 3: Результаты работы алгоритма

Ряд, $\mathbf{x}$	Длина ряда, N	# сегментов, K	Длина сегмента, T	Ошибка,
Phys. Motion 1	900	2	40	0.06
Phys. Motion 2	900	2	40	0.03
Synthetic 1	2000	2	20	0.04
Synthetic 2	2000	3	20	0.03

В работе рассматривалась задача поиска характерных периодических структур внутри временного ряда. Рассматривался метод основаный на локальном снижение размерности фазового пространства. Был предложен алгоритм поиска характерных сегментов, который основывается на методе главных компонент для локального снижения размерности. Также введена функция расстояния между локальными базисами в каждый момент времени, которые интерпретировались как признаковое описание точки временного ряда.

В ходе эксперимента, на реальных показаниях акселерометра, а

также на синтетических данных, было показано, что предложенный метод измерение расстояния между базисами хорошо разделяет точки которые принадлежат различным действиям, что приводит к хорошей кластеризации объектов. Результаты работы показаны в таблице 3. Также в эксперименте была проведена полная сегментация временных рядов при помощи метода [6] для каждого кластера по отдельности.

Предложенный метод имеет ряд недостатков связанных с большим число ограничений на временной ряд. Данные ограничения будут ослаблены в последующих работах. Планируется решить задачу нахождения и описания замкнутой фазовой траектории, которая относится к одному квазипериодическому сегменту.

Список литературы

[1] *J. R. Kwapisz, G. M. Weiss, S. A. Moore* Activity Recognition using Cell Phone Accelerometers // Proceedings of the Fourth International Workshop on Knowledge Discovery from Sensor Data, 2010. Vol. 12. P. 74–82.

[2] *W. Wang, H. Liu, L. Yu, F. Sun* Activity Recognition using Cell Phone Accelerometers // Joint Conference on Neural Networks, 2014. P. 1185–1190.

[3] *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.

[4] *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.

[5] *Y. G. Cinar and H. Mirisaee* Period-aware content attention RNNs for time series forecasting with missing values // Neurocomputing, 2018. Vol. 312. P. 177–186.

[6] *A. P. Motrenko, V. V. Strijov* Extracting fundamental periods to segment biomedical signals // Journal of Biomedical and Health Informatics, 2015, 20(6). P. 1466 - 1476.

[7] *Y. P. Lukashin* Adaptive methods for short-term forecasting // Finansy and Statistik, 2003.

[8] *И. П. Ивкин, М. П. Кузнецов* Алгоритм классификации временных рядов акселерометра по комбинированному признаковому описанию. // Машинное обучение и анализ данных, 2015.

[9] *V. V. Strijov, A. M. Katrutsa* Stresstes procedures for features selection algorithms. // Schemometrics and Intelligent Laboratory System, 2015.

[10] *I. Borg, P. J. F. Groenen* Modern Multidimensional Scaling. — New York: Springer, 2005. 540 p.

[11] *Д. Л. Данилова, А. А. Жигловский* Главные компоненты временных рядов: метод "Гусеница". — Санкт-Петербурскиий университет, 1997.