Анализ свойств локальных моделей в задачах кластеризации временных рядов*

А. В. Грабовой[1], **В. В. Стрижов**[2]

Аннотация: Данная работа посвящена поиску периодических сигналов во временных рядах с целью 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 метода главных компонент для локального снижения размерности фазового пространства. Для оценки близости двух периодических сигналов вычисляется расстояние между базисными векторами, которые получены методом главных компонент. Используя матрицу 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.

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

DOI: 00.00000/00000000000000

*Работа выполнена при поддержке РФФИ и правительства РФ.

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

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

1 Введение

Анализ повседневной физической активности человека производится при помощи мобильных телефонов, разумных часов. Портативные устройства используют акселерометр, гироскоп и магнитометр. Цель данной работы заключается в распознавании и разметке человеческой активности во времени.

Временные ряды это объекты сложной структуры, при классификации которых значимую роль играет построение признакового пространства. Для этой цели используются: экспертно заданные базовые функций, гипотеза порождения данных. В [2] рассматривается комбинированное признаковое описание на основе данных методов. В [3] также рассматривается проблема построение признакового пространства и предлагается критерий избыточности выбранных признаков.

В данной работе рассмотрена задача *кластеризации* точек временного ряда. Под *кластеризацией* точек, подразумевается сопоставление каждой точке временного ряда некоторой метке, которая соответствует некоторому *сегменту* временного ряда. *Сегмент* временного ряда это часть временного ряда, которая соответсвует одному характерному физическому действию, например: один шаг при ходьбе, или один шаг при беге. Пример сегментов показан на рис. 1. На рис. 1 также показан пример кластеризации временного ряда, в котором ряд разбит на два характерных физических действия, полученных при помощи акселерометра.

Решение задачи кластеризации состоит из двух этапов. Во-первых предлагается алгоритм *локальной* аппроксимации временного ряда при помощи метода главных компонент [6] для получения признакового описания временного ряда. Под *локальной* аппроксимацией временного ряда подразумевается, что для признакового описания его точки используется не весь ряд, а только некоторая окрестность данной точки. Во-вторых рассматривается метрика в новом пространстве признакового описания. После получения расстояния между точками временного ряда используется метод кластеризации KMeans [5] для кластеризации точек временного ряда.

Для решения задачи кластеризации точек временного ряда вводится ряд предположений о данном ряде. Предполагается, что периоды всех различных сегментов различаются не значительно, причем известен максимальный период сегмента и количество различных сегментов внутри

временного ряда. Также предполгается, что тип активности во времени меняется не часто, а также что фазовые траектории разных сегментов являются различными.

Проверка и анализ метода проводится на синтетической и реальной выборках. Синтетическая выборка построенная при помощи суммы нескольких первых членов ряда Фурье со случайными коэффициентами. Реальные данные получены при помощи мобильного акселерометра, который снимал показания во время некоторой физической активности человека.

2 Постановка

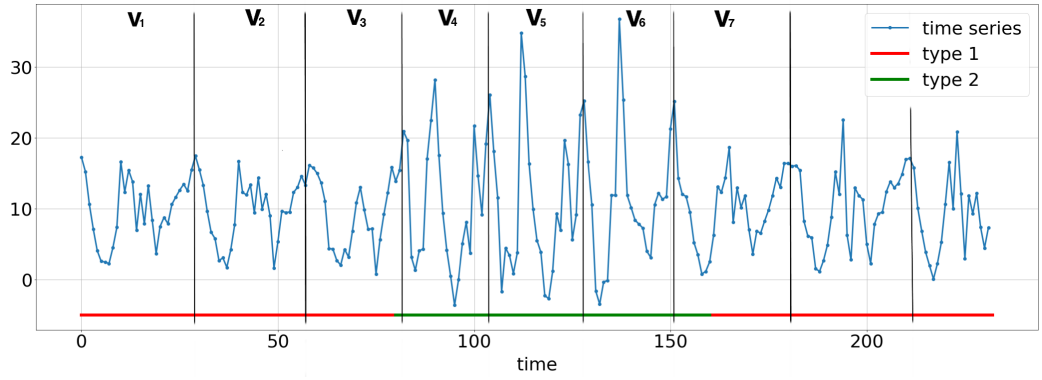


Рис. 1: Временной ряд, с разметкой на кластеры.

Задан временной ряд:

$$\mathbf{x} \in \mathbb{R}^N, \tag{2.1}$$

где N количество точек, которыми задается временной ряд.

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

$$\mathbf{x} = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_M], \tag{2.2}$$

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

$$|\mathbf{V}| = K, \quad \forall \mathbf{v} \in \mathbf{V} \;\; |\mathbf{v}| \leq T, \tag{2.3}$$

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

Рассмотрим отображение:

$$a : x \to \{1, \cdots, K\}, \tag{2.4}$$

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

Требуется, чтобы отображение удовлетворяло следующим свойствам:

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

Пусть задана некоторая асессорская разметка временного ряда:

$$\mathbf{y} \in \{1, \cdots, K\}^N. \tag{2.5}$$

Тогда ошибка, которую совершает алгоритм a на временном ряде $\mathbf{x}$ представляется в следующем виде:

$$S = \frac{1}{N} \sum_{i=1}^{N} [y_i = a(x_i)], \tag{2.6}$$

где x_i — точки временного ряда, y_i асессорская разметка временного ряда.

3 Кластеризация точек

Рассматривается фазовая траектория временного ряда $\mathbf{x}$:

$$\mathbf{H} = \{\mathbf{h}_t | \mathbf{h}_t = [x_{t-T}, x_{t-T+1}, \cdots, x_t],\ T \leq t \leq N\}. \tag{3.1}$$

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

$$\mathbf{S} = \{\mathbf{s}_t | \mathbf{s}_t = [h_{t-2T}, h_{t-2T+1}, \cdots, h_t],\ 2T \leq t \leq N\}. \tag{3.1}$$

Утверждение: Размерность $2T$ фазового подпространства является достаточным для построения локальной аппроксимирующей модели.

Каждое T-мерное пространство $\mathbf{s}_t$ проектируется на подпространство значительно меньшей размерности при помощи метода главных компонент $\mathbf{z}_t = \mathbf{W}_t \mathbf{s}_t$. Получим представление базисных векторов $\mathbf{W}_t$,

а также собственные числа, которые соответствуют данным базисным векторам каждого подпространства $\mathbf{s}_t$ в T-мерном пространстве:

$$\mathbf{W} = \{\mathbf{W}_t | \mathbf{W}_t = [\mathbf{w}_t^1, \mathbf{w}_t^2]\}, \quad \mathbf{\Lambda} = \{\boldsymbol{\lambda}_t | \boldsymbol{\lambda}_t = [\lambda_t^1, \lambda_t^2]\}, \tag{3.3}$$

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 $\mathbf{s}_t$.

Рассмотрим расстояние между элементами $\mathbf{W}$:

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

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

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

Рассмотрим расстояние между элементами $\mathbf{\Lambda}$:

$$\rho\left(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2\right) = \sqrt{\left(\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\right)^{\mathsf{T}} \left(\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\right)}. \tag{3.5}$$

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

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

$$\mathbf{M}_{\mathrm{c}} = [0, 1]^{N \times N}. \tag{3.6}$$

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

$$\mathbf{M}_{\mathrm{l}} = [0, 1]^{N \times N}. \tag{3.7}$$

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

$$\rho\left(t_1, t_2\right) = \rho\left(\mathbf{W}_1, \mathbf{W}_2\right) + \rho\left(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2\right), \quad \mathbf{M} = \mathbf{M}_{\mathrm{l}} + \mathbf{M}_{\mathrm{c}}, \tag{3.8}$$

где $\rho\left(t_1, t_2\right)$ 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:

$$a : x \to \{1, \cdots, K\}, \tag{3.9}$$

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 $(3.6 - 8)$.

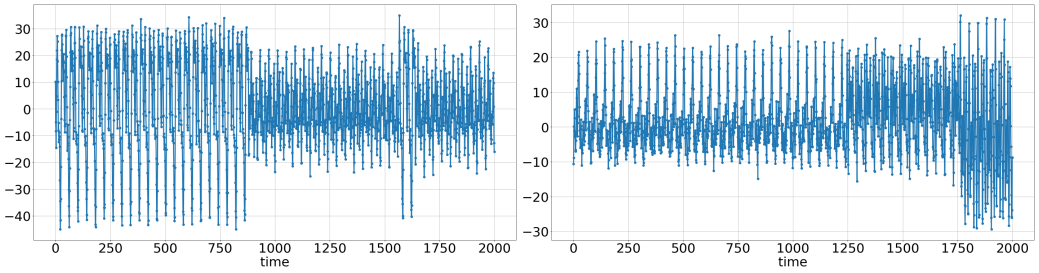
В качестве данных использовались две выборки временных рядов, которые описаны в таблице 1. Выборка Physical Motion это реальные временные ряды полученные при помощи мобильного акселерометра. Синтетические временные ряды были построены при помощи нескольких первых слагаемых ряда Фурье со случайными коэффициентами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

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

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

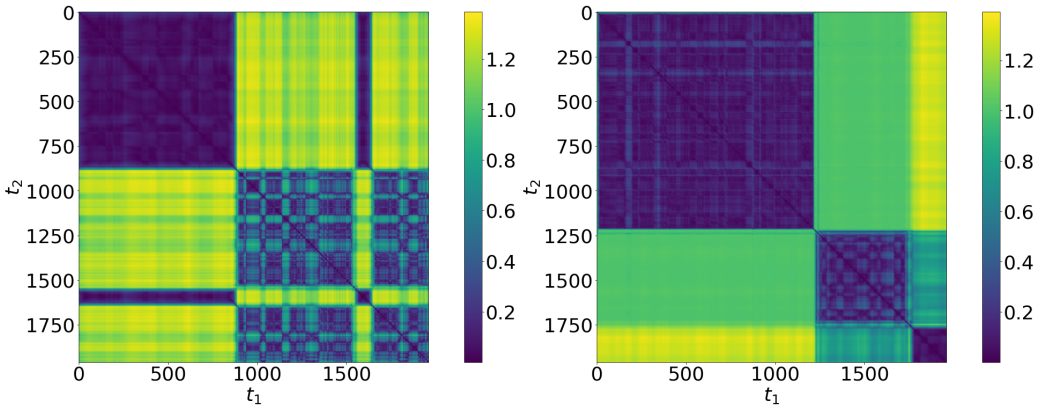
Таблица 1: Описание выборок

Ряд	N	K	T
Physical Motion 1	900	2	30
Physical Motion 2	1000	2	30
Synthetic 1	2000	2	20
Synthetic 2	2000	3	20



(a) Synthetic 1 (b) Synthetic 2

Рис. 2: Пример синтетически построенных временных рядов

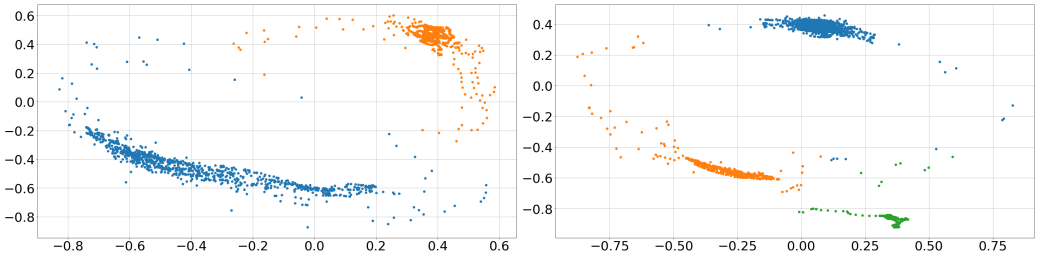


(a) Synthetic 1 (b) Synthetic 2

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

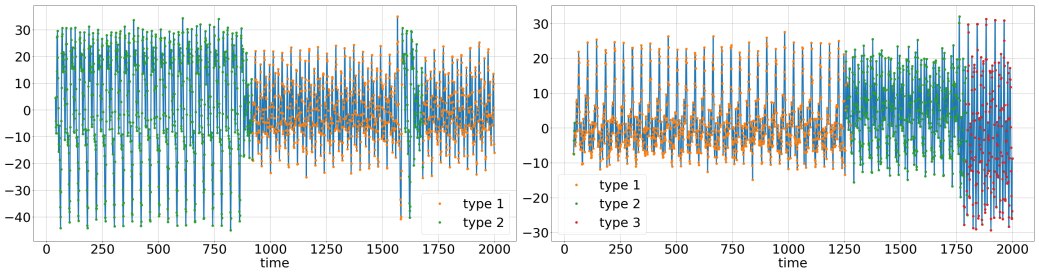
Синтетические данные. На рис. 2 приведен пример синтетически построенных временных рядов. На рис. 2a показан пример ряда в котором количество сигналов $K = 2$, а длина каждого сигнала $T = 20$. На рис. 2b показан пример ряда в котором количество сигналов $K = 3$, а длина каждого сигнала $T = 20$.

На рис. 3 проиллюстрированы матрицы попарных расстояний $\mathbf{M}$ между построены при помощи формулы (3.8). Используя матрицу попарных расстояний и метод Multidimensional Scaling [4] визуальзуализируем точки временного ряда на плоскости. На рис. 4 показана визуализация точек на плоскости и выполнена их кластеризация при помощи метода



(a) Synthetic 1 (b) Synthetic 2

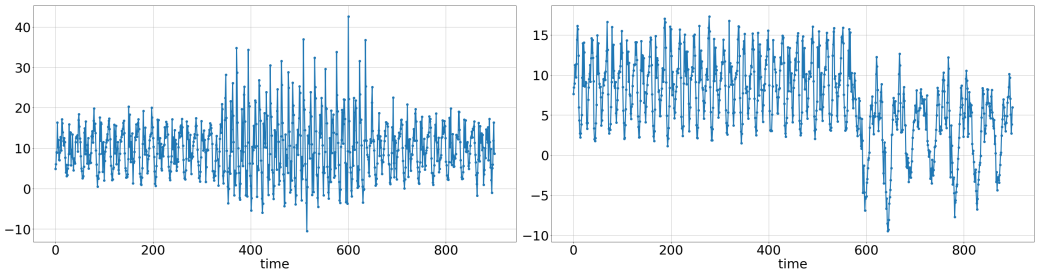
Рис. 4: Проекция точек временного ряда на плоскость при помощи матрицы попарных расстояний $\mathbf{M}$



(a) Synthetic 1 (b) Synthetic 2

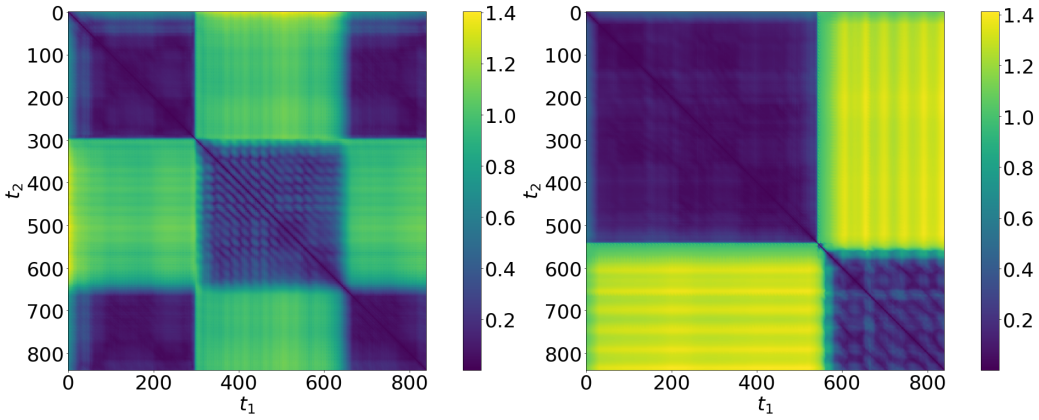
Рис. 5: Кластеризация точек временного ряда

KMeans [5]. Иллюстрация кластеров точек временного ряда продемонстрирована на рис. 5.



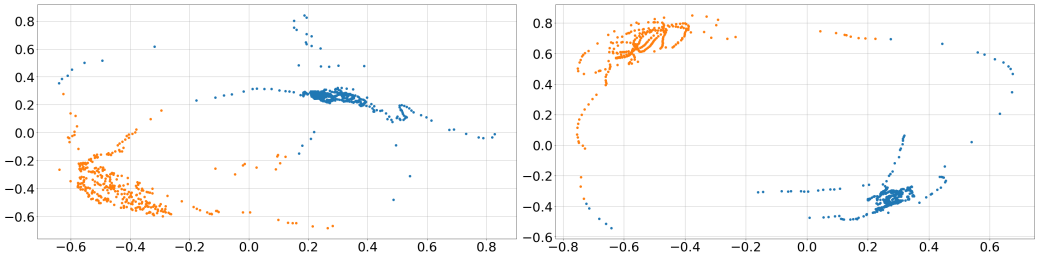
(a) Physical Motion 1 (b) Physical Motion 2

Рис. 6: Пример синтетически построенных временных рядов



(a) Physical Motion 1 (b) Physical Motion 2

Рис. 7: Матрица попарных расстояний $\mathbf{M}$ между точками временного ряда

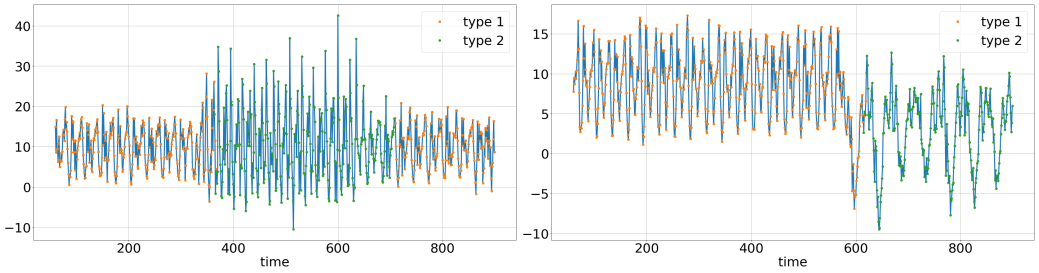


(a) Physical Motion 1 (b) Physical Motion 2

Рис. 8: Проекция точек временного на плоскость при помощи матрицы попарных расстояний $\mathbf{M}$

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

На рис. 7 проиллюстрированы матрицы попарных расстояний $\mathbf{M}$ между построены при помощи формулы (3.8). Используя матрицу попарных расстояний и метод Multidimensional Scaling [4] визуальзуализируем точки временного ряда на плоскости. На рис. 8 показана визуализация точек на плоскости и выполнена их кластеризация при помощи метода KMeans [5]. Иллюстрация кластеров точек временного ряда продемон-



(a) Physical Motion 1 (b) Physical Motion 2

Рис. 9: Кластеризация точек временного ряда

стрирована на рис. 9.

5 Заключение

Таблица 2: Результаты работы алгоритма

Ряд	N	K	T	S
Physical Motion 1	900	2	30	0.06
Physical Motion 2	1000	2	30	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 сигналов, который основывается на методе главных компонент для локального снижения размерности, а также на использовании некоторой функции расстояния между локальными базисами в каждый момент времени, которые интерпретировались как признакового 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 классам, что приводит к хорошей кластеризации объектов. Результаты работы, показаны в таблице 2.

Предложеный метод имеет ряд недостаткров связаных с большим количеством ограничей на временной ряд. Данные ограничения будут ослаблены в последнующих работах.

Список литературы

[1] *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.

[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-Петербурскиий 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.