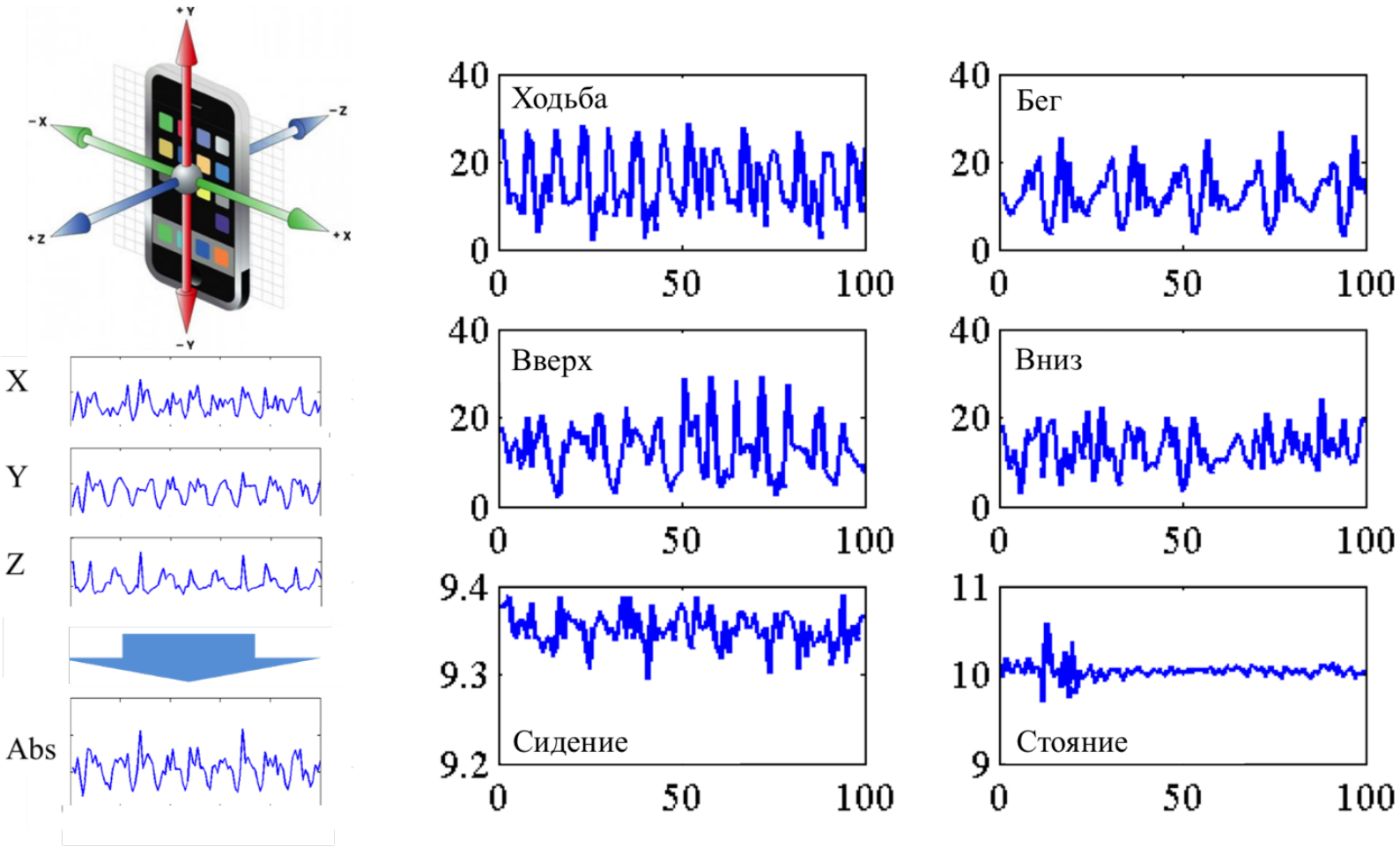
Метрический анализ временных рядов с помощью их динамического выравнивания

Гончаров Алексей Владимирович

Московский физико-технический институт
Кафедра интеллектуальных систем, ФУПМ.
Научный руководитель: д.ф.-м.н. В. В. Стрижов.

Раздел 1. Описание задачи и выбор базовой функции расстояния.

Пример задачи



Абсолютные значения ускорения акселерометра мобильного телефона для различных классов физической активности человека

Цель исследования

Построить функцию расстояния между временными рядами, которая собирает объекты одного класса и разделяет объекты различных.

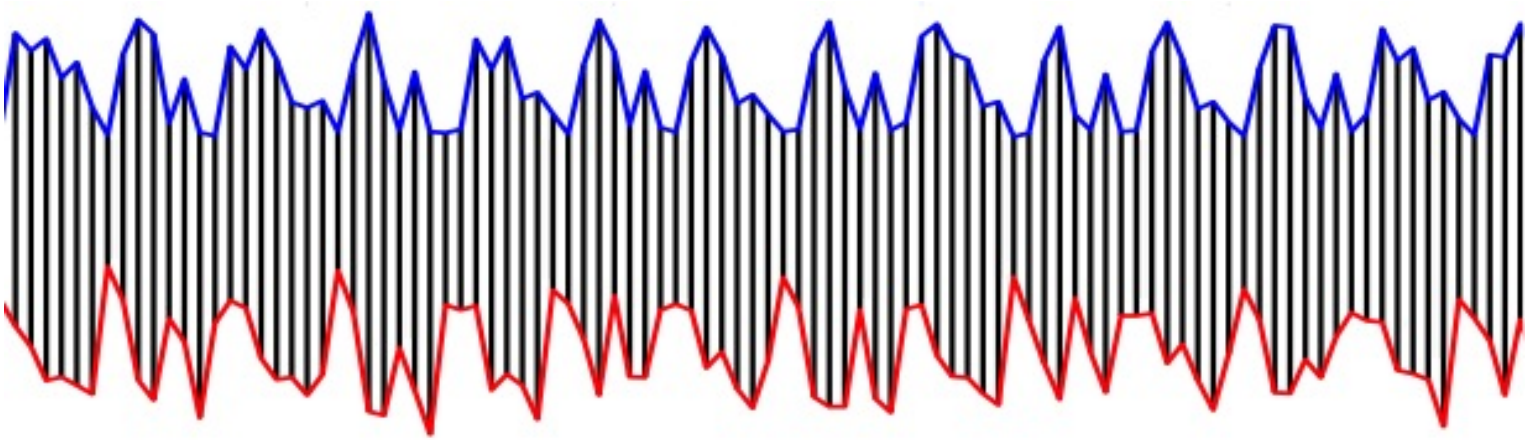
Проблема

- Высокая вычислительная сложность.
- Локальные и глобальные сдвиги временных рядов.

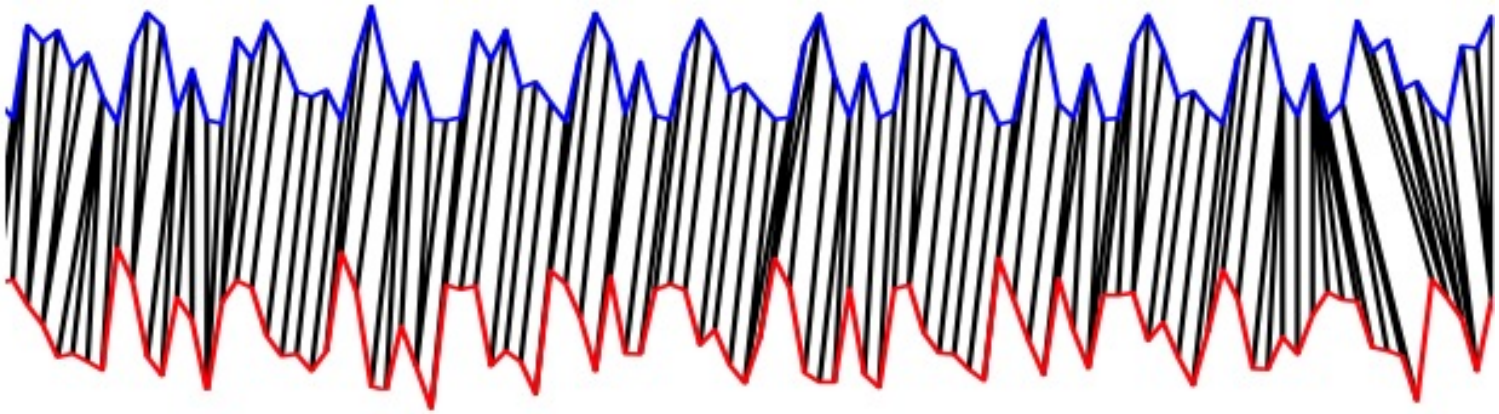
Предлагается

- Строить центроиды классов.
- Модифицировать существующую функцию.

Выбор базовой функции расстояния

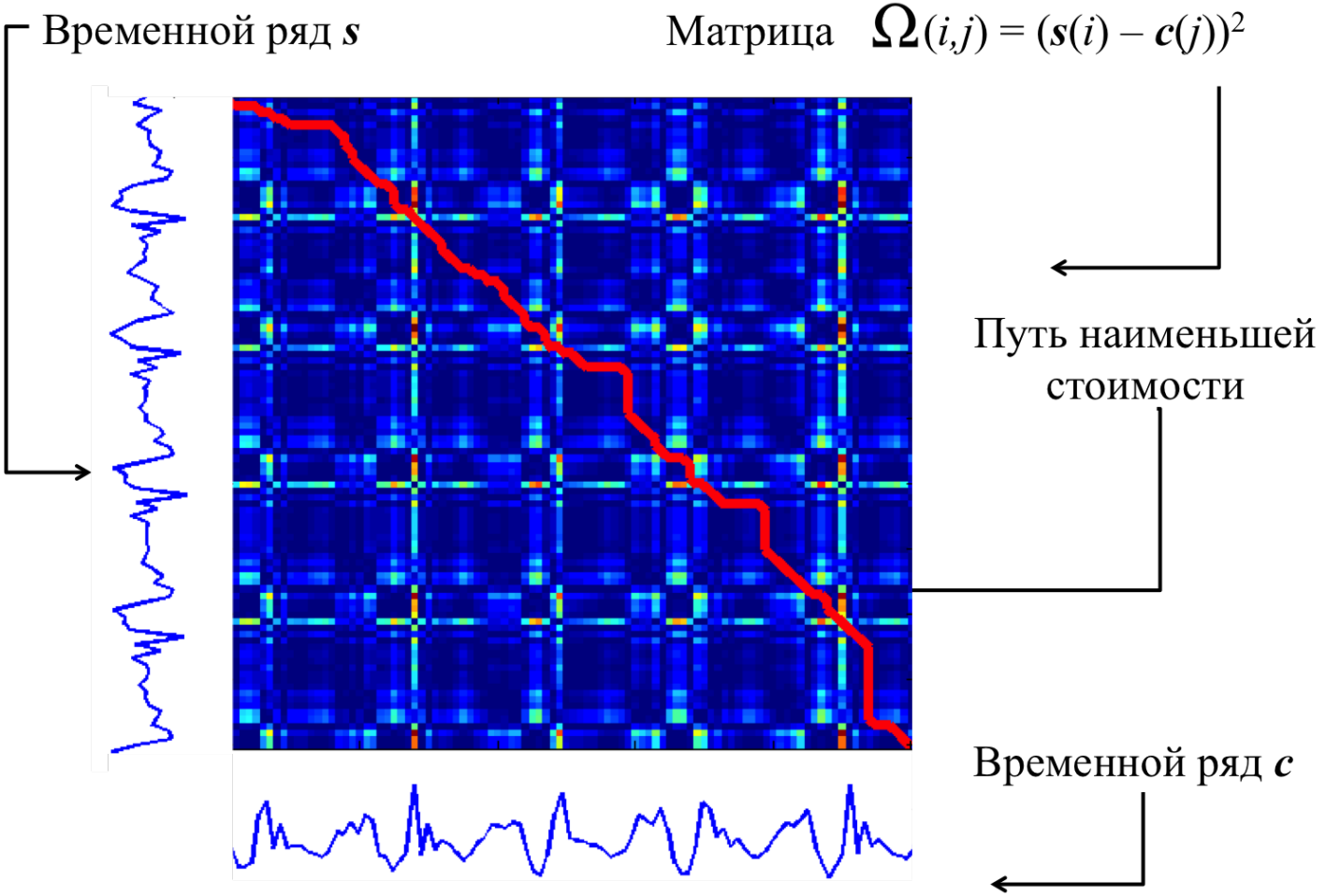


Евклидово расстояние между временными рядами



Выровненное расстояние между временными рядами

Путь наименьшей стоимости



- Матрица расстояний между элементами временных рядов $\mathbf{\Omega}^{n\times n} : \mathbf{\Omega}(i,j) = \big(\mathbf{s}(i) - \mathbf{c}(j)\big)^2$.
- Путь $\boldsymbol{\pi}$ длины K между $\mathbf{s}$ и $\mathbf{c}$:
  $\boldsymbol{\pi} = \{\pi_k\} = \{(i_k, j_k)\}, \quad k = 1, \ldots, K, \quad i, j \in \{1, \ldots, n\}$

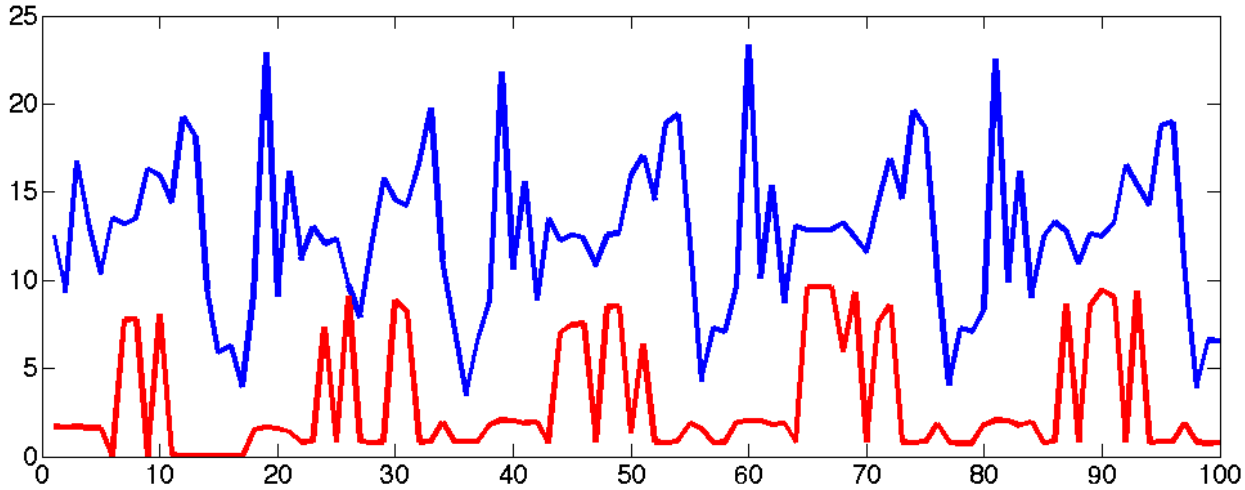
Список литературы

- *Petitjean F., Forestier G., Webb G. I., Nicholson A. E., Chen Y., Keogh E.* Dynamic Time Warping Averaging of Time Series allows Faster and more Accurate Classification // IEEE International Conference on Data Engineering (ICDE), 2014.
- *Гончаров А. В., Попова М. С., Стрижов В. В.* Метрическая классификация временных рядов с выравниванием относительно центроидов классов // Системы и средства информатики, 2015.
- *Berndt D. J., Clifford J.* Using dynamic time warping to find patterns in time series // In KDD Workshop, 1994.
- *Keogh E. J., Ratanamahatana C. A.* Exact indexing of dynamic time warping. // Knowl. Inf. Syst., 2005. Vol. 7, No. 3.
- *Kwapisz J. R.* 2010. Данные из акселерометра.

  http : //sourceforge.net/p/mlalgorithms/TSLearning/data/preprocessedla

Раздел 2. Взвешенное выравнивание относительно центроида.

Модификация функции расстояния



В предположении, что отдельные сегменты временного ряда более информативно описывают класс физической активности, предлагается: *взвесить элементы временного ряда.*

Модифицированная функция расстояния

Стоимость пути π: $\mathrm{Cost}(\mathbf{s}_1, \mathbf{s}_2, \pi)$

DTW	vwDTW
$\sum_{(i,j)\in\pi} \mathbf{\Omega}(i,j)$	$\sum_{(i,j)\in\pi} w_e(j)\mathbf{\Omega}(i,j)$

Путь наименьшей стоимости (выравнивающий путь)

$$\hat{\pi} = \underset{\pi}{\operatorname{argmin}}\, \mathrm{Cost}(\mathbf{s}_1, \mathbf{s}_2, \pi).$$

Функция расстояния DTW, vwDTW

$$\rho(\mathbf{s}_1, \mathbf{s}_2) = \mathrm{Cost}(\mathbf{s}_1, \mathbf{s}_2, \hat{\pi}).$$

Постановка задачи построения центроида

Пусть веса $\mathbf{W}$ — фиксированы.

Определение центроида

Центроид множества векторов $\mathfrak{D}_k = \{\mathbf{s}_i | y_i = k\}_{i=1}^m$ по расстоянию ρ — вектор $\mathbf{c} \in \mathbb{R}^n$ такой, что:

$$\mathbf{c}_e = \underset{\mathbf{c}\in\mathbb{R}^n}{\operatorname{argmin}} \sum_{\mathbf{s}_i \in \mathfrak{D}_e} \rho(\mathbf{s}_i, \mathbf{c}).$$

Решение оптимизационной задачи

$$\mathbf{c}_e = \underset{\mathbf{c}\in\mathbb{R}^n}{\operatorname{argmin}} \sum_{\mathbf{s}_i \in \mathfrak{D}_e} \sum_{(t,t')\in\hat{\boldsymbol{\pi}}_i} \mathbf{w}_e(t)\big(\mathbf{s}_i(t') - \mathbf{c}(t)\big)^2,$$

где $\hat{\boldsymbol{\pi}}_i$ — взвешенный выравнивающий путь между временными рядами $\mathbf{s}_i$ и $\mathbf{c}$.

Решение задачи нахождения центроида методом DBA

Теорема 1: Пусть дано множество векторов $\mathfrak{D}_e = \{\mathbf{s}_i | y_i = e\}_{i=1}^m$ одного класса, начальное приближение центроида $\mathbf{c}_e$ и множество выравнивающих путей между каждым рядом и начальным приближением центроида $\{\tilde{\boldsymbol{\pi}}_i\}_{i=1}^m$. Тогда локальный минимум задачи оптимизации при единичном векторе весов в достигается при:

$$\mathbf{c}_e(t) = \frac{1}{N} \sum_{\mathbf{s}_i \in \mathfrak{D}_e} \sum_{t' : (t,t') \in \tilde{\boldsymbol{\pi}}_i} \mathbf{s}_i(t'),$$

$$N = \sum_{\mathbf{s}_i \in \mathfrak{D}_e} \sum_{t' : (t,t') \in \tilde{\boldsymbol{\pi}}_i} 1.$$

Следствие 1: Аналогичное выполняется и для общего случая vwDTW при замене множества путей наименьшей стоимости $\{\tilde{\boldsymbol{\pi}}\}_{i=1}^m$ на множество взвешенных путей наименьшей стоимости $\{\hat{\boldsymbol{\pi}}_i\}_{i=1}^m$.

Оптимизация вектора весов при известном центроиде

Множество центроидов $\mathbf{C}$ фиксировано. Каждому центроиду $\mathbf{c}_e$ из множества $\mathbf{C}$ поставлен в соответствие вектор неотрицательных весов $\mathbf{w}_e$.

$$\mathbf{w}_e = \underset{\mathbf{w}\in\mathbb{R}^n}{\operatorname{argmin}} \sum_{\mathbf{s}_i\in\mathfrak{D}_e} \sum_{(t,t')\in\boldsymbol{\pi}_i} \mathbf{w}_e(t)(\mathbf{s}_i(t') - \mathbf{c}_e(t))^2,$$

при следующих ограничениях:

$$\sum_{t=1}^{T} \mathbf{w}_e(t) = T,$$

$$\mathbf{w}_e(t) \geq a, \quad \mathbf{w}_e(t) \leq b, \quad t \in \{1 \ldots T\}$$

Алгоритм решения задачи классификации

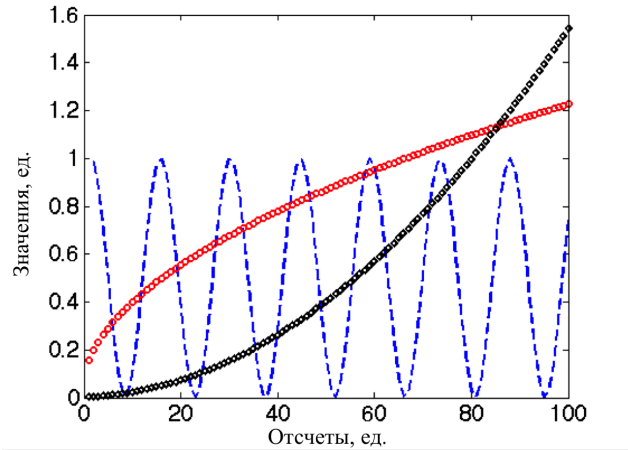
Пусть задано начальное приближение вектора весов центроида и центроид: $\mathbf{w}_e = \mathbf{1}, \quad \mathbf{c}_e = \mathbf{s}_j \in \mathfrak{D}_e \quad e = 1, \ldots, E.$

Шаг 1. Вычисление центроида $\mathbf{c}_e$ при фиксированном векторе весов $\mathbf{w}_e$ центроида.

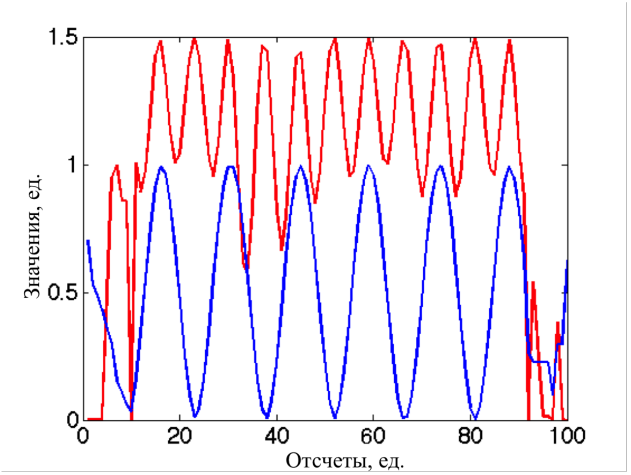
Шаг 2. Вычисление вектора весов $\mathbf{w}_e$ центроида при фиксированном центроиде $\mathbf{c}_e$.

Шаг 3. Использование полученных параметров функции расстояния между временными рядами для решения задачи классификации методом ближайшего соседа.

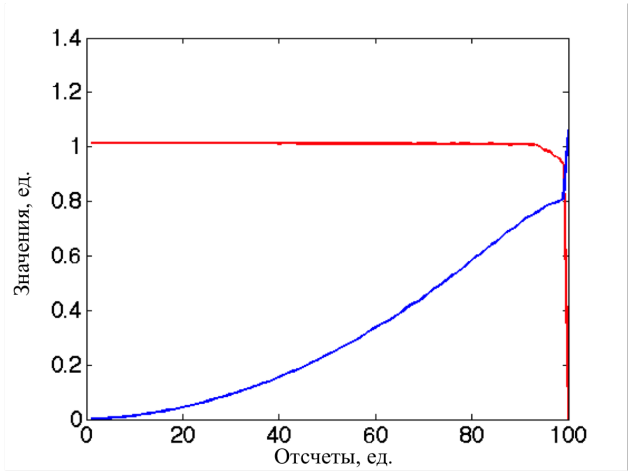
Анализ вектора весов на синтетических данных



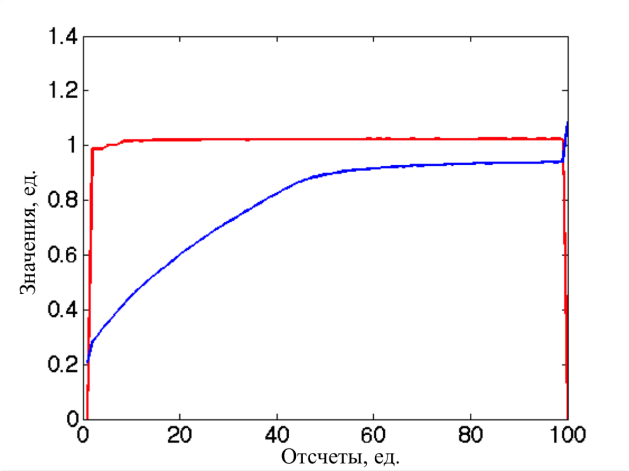
Выборка



Веса для $sin(x)$

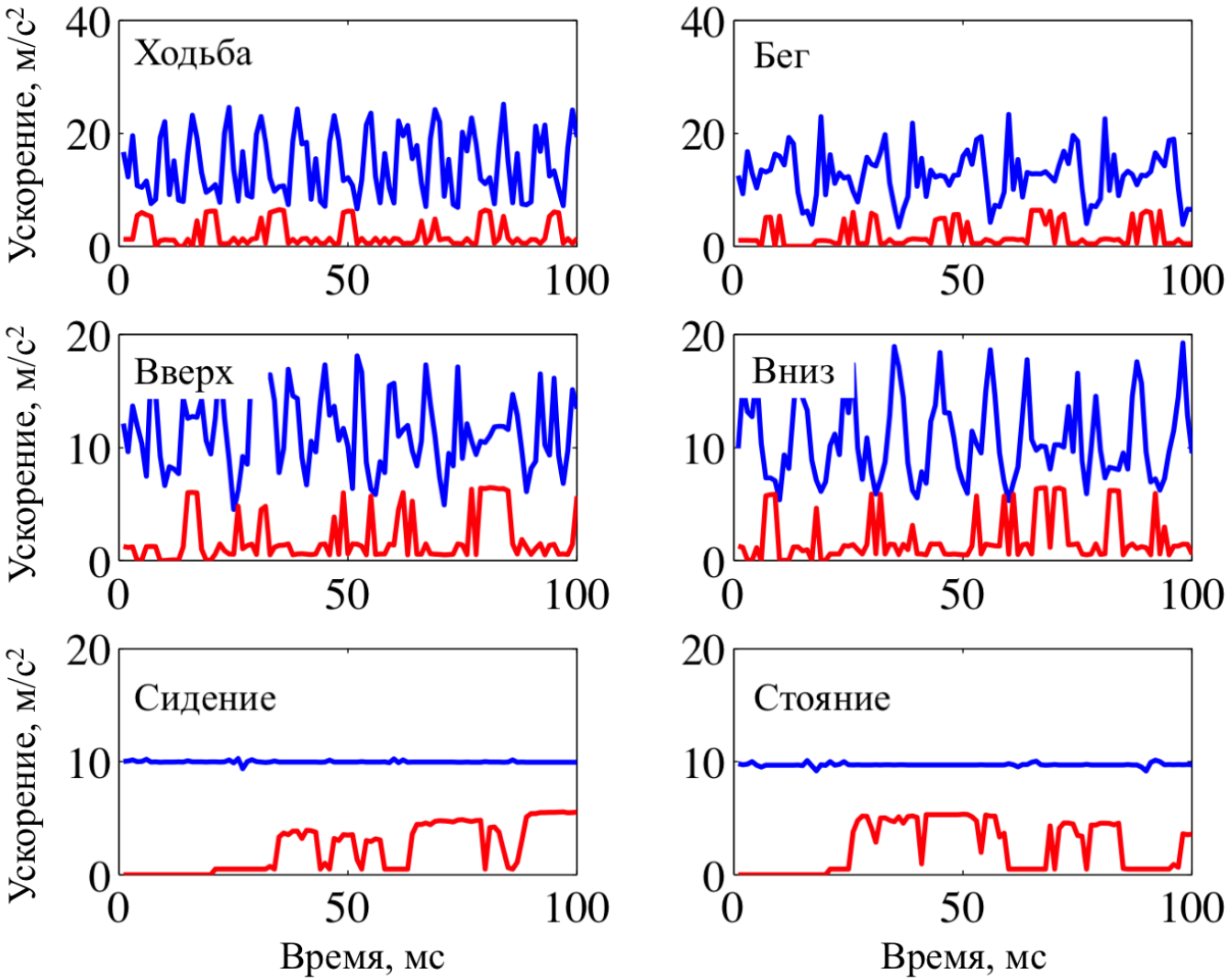


Веса для x^2



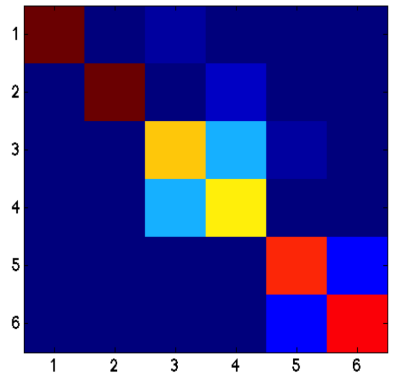
Веса для $\sqrt{x}$

Вектор весов и центроид для разных классов физической активности

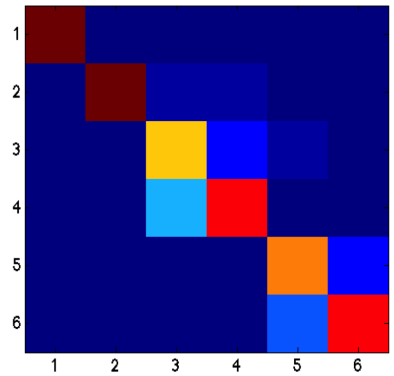


Результаты классификации

Матрица несоответствий (confusion matrix) для классификации с помощью DTW, vwDTW и mwDTW. Ходьба —1, бег —2, вверх —3, вниз — 4, сидение — 5, стояние — 6.



DTW
качество 80%



vwDTW
качество 86%

Раздел 3. Случай непрерывных объектов.

Цель исследования

Предложить эффективный способ работы с временными рядами, частота измерений которых различна.

Предлагается

- Перейти к непрерывным аналогам временных рядов.
- Построить функцию расстояния между ними, аналогичную существующей DTW.

Проблемы

- Что такое непрерывный временной ряд?
- Определения функции расстояния между непрерывными объектами.
- Задача поиска выравнивающего пути не решается полным перебором в непрерывном случае.

Определение временного ряда

Определение 1. Дискретный случай.

В дискретном случае временной ряд **s** представляет собой упорядоченную во времени последовательность измерений какой-либо величины $\{s_i\}_{i=1}^{T}$.

Определение 1. Непрерывный случай.

Непрерывный временной ряд, определенный на участке времени $\widehat{T} = [0; T]$, — непрерывная функция $s^c(t) : \widehat{T} \to \mathbb{R}$.

Определение пути между временными рядами

Определение 2. Дискретный случай.

Путь $\boldsymbol{\pi}$ между дискретными временными рядами $\mathbf{s}_1$ и $\mathbf{s}_2$ — упорядоченное множество пар индексов:

$$\boldsymbol{\pi} = \{\pi_r\} = \{(i_r, j_r)\}, \quad r = 1, \ldots, R, \quad i, j \in \{1, ..., n\},$$

удовлетворяющее условиям непрерывности, монотонности и граничным условиям:

$$\pi_r = (p_1, p_2), \pi_{r-1} = (q_1, q_2), r = 2, ..., R, \Rightarrow p_1 - q_1 \leq 1, p_2 - q_2 \leq 1,$$

$$\pi_r = (p_1, p_2), \pi_{r-1} = (q_1, q_2), r = 2, ..., R, \Rightarrow p_1 - q_1 \geq 1, p_2 - q_2 \geq 1,$$

$$\pi_1 = (1, 1), \quad \pi_R = (n, n).$$

Определение пути между временными рядами

Определение 2. Непрерывный случай.

Путь π^c между двумя непрерывными временными рядами — монотонно возрастающая, непрерывная функция $\pi^c : t_1 \to t_2$, удовлетворяющая начальным условиям:

$$\pi^c \in C_{[0;T]},$$

$$t_1 > t_1' \Rightarrow \pi^c(t_1) > \pi^c(t_1'),$$

$$\pi^c(0) = 0, \quad \pi^c(T_1) = T_2.$$

Определение стоимости пути

Определение 3. Дискретный случай.

Стоимость $\mathrm{Cost}(\mathbf{s}_1, \mathbf{s}_2, \boldsymbol{\pi})$ пути $\boldsymbol{\pi}$ длины R между дискретными временными рядами $\mathbf{s}_1$ и $\mathbf{s}_2$:

$$\mathrm{Cost}(\mathbf{s}_1, \mathbf{s}_2, \boldsymbol{\pi}) = \frac{1}{R} \sum_{(i,j)\in\boldsymbol{\pi}} |\mathbf{s}_1(i) - \mathbf{s}_2(j)|.$$

Определение 3. Непрерывный случай.

Стоимость $\mathrm{Cost}(s_1^c(t_1), s_2^c(t_2), \pi^c)$ пути π^c между непрерывными временными рядами $s_1^c(t_1)$ и $s_2^c(t_2)$:

$$\mathrm{Cost}(s_1^c(t_1), s_2^c(t_2), \pi^c) = \frac{1}{L} \int_{t_1} |s_1^c(t_1) - s_2^c(\pi^c(t_1))| dt_1,$$

где L — длина кривой, задающейся графиком функции $\pi^c(t), \quad t \in [0, T]$.

Определение выравнивающего пути

Определение 4. Дискретный случай.

Путь наименьшей стоимости (выравнивающим путем) $\widehat{\boldsymbol{\pi}}$ между дискретными временными рядами $\mathbf{s}_1$ и $\mathbf{s}_2$ — путь, имеющий наименьшую стоимость среди всех возможных путей:

$$\widehat{\boldsymbol{\pi}} = \underset{\boldsymbol{\pi}}{\operatorname{argmin}} \operatorname{Cost}(\mathbf{s}_1, \mathbf{s}_2, \boldsymbol{\pi}).$$

Определение 4. Непрерывный случай.

Путь наименьшей стоимости (выравнивающий путь) $\widehat{\pi}^c$ между непрерывными временными рядами $s_1^c(t_1)$ и $s_2^c(t_2)$ — функция $\widehat{\pi}^c$, для которой значение интеграла из определения 2 для непрерывного случая является наименьшим:

$$\widehat{\pi}^c = \underset{\pi^c}{\operatorname{argmin}} \operatorname{Cost}(s_1^c(t_1), s_2^c(t_2), \pi^c).$$

Определение стоимости выравнивающего пути

Определение 5. Дискретный случай.

Стоимость пути наименьшей стоимости, или расстояние DTW между дискретными временными рядами:

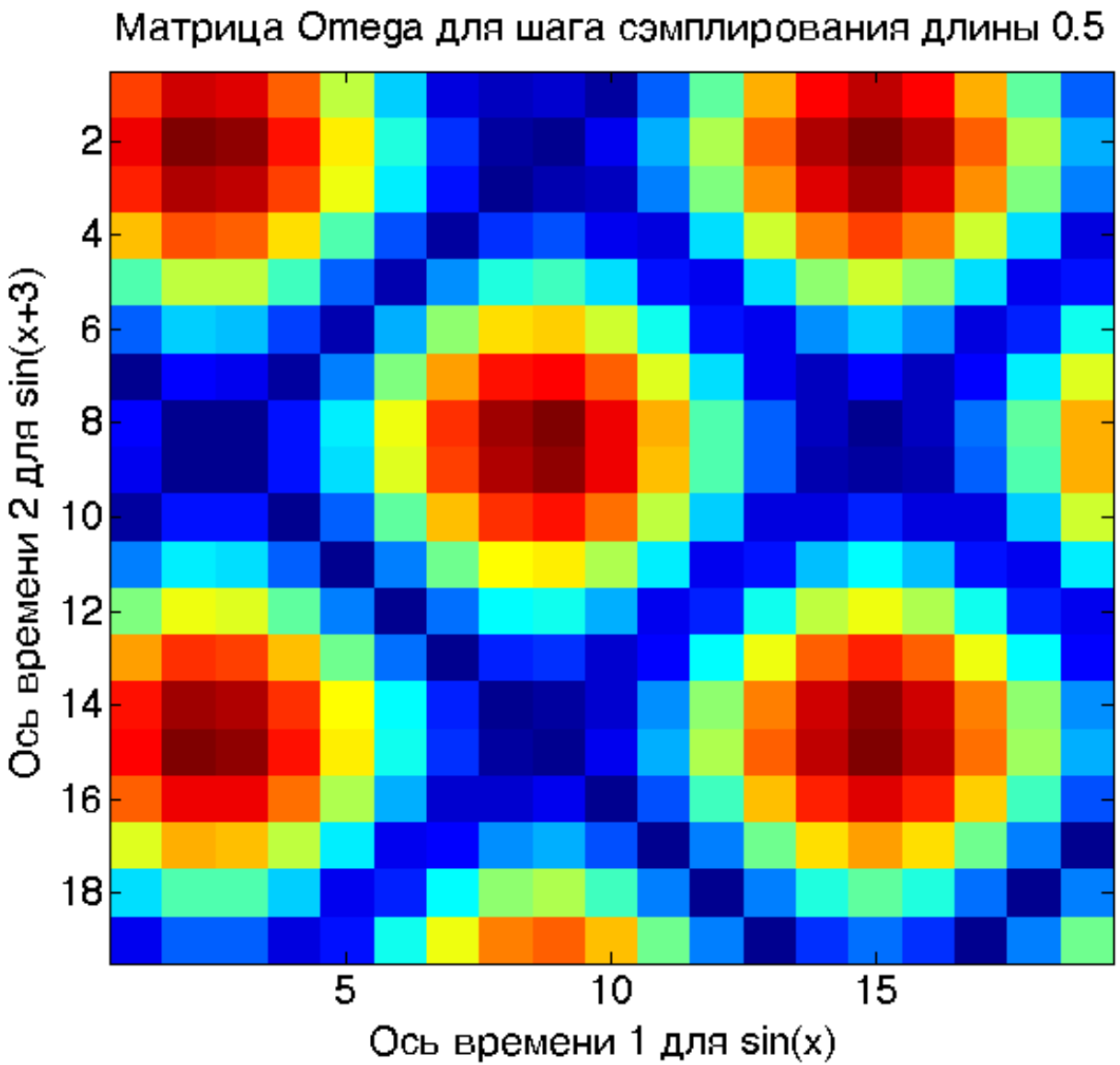
$$DTW(\mathbf{s}_1, \mathbf{s}_2) = \text{Cost}(\mathbf{s}_1, \mathbf{s}_2, \widehat{\pi}).$$

Определение 5. Непрерывный случай.

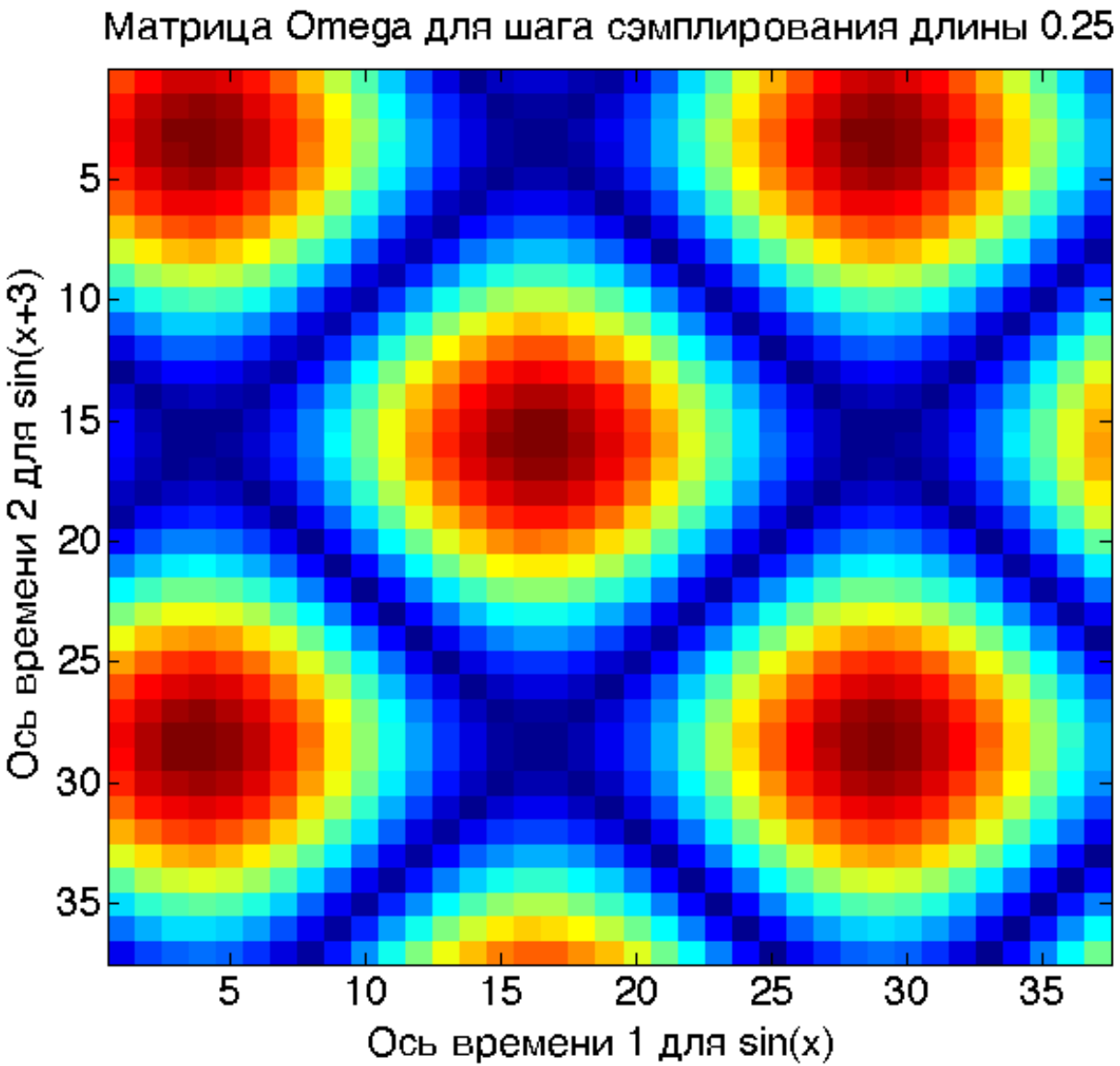
Стоимость пути наименьшей стоимости, или расстояние DTW между непрерывными временными рядами:

$$DTW(s_1^c(t_1), s_2^c(t_2)) = \text{Cost}(s_1^c(t_1), s_2^c(t_2), \widehat{\pi}^c).$$

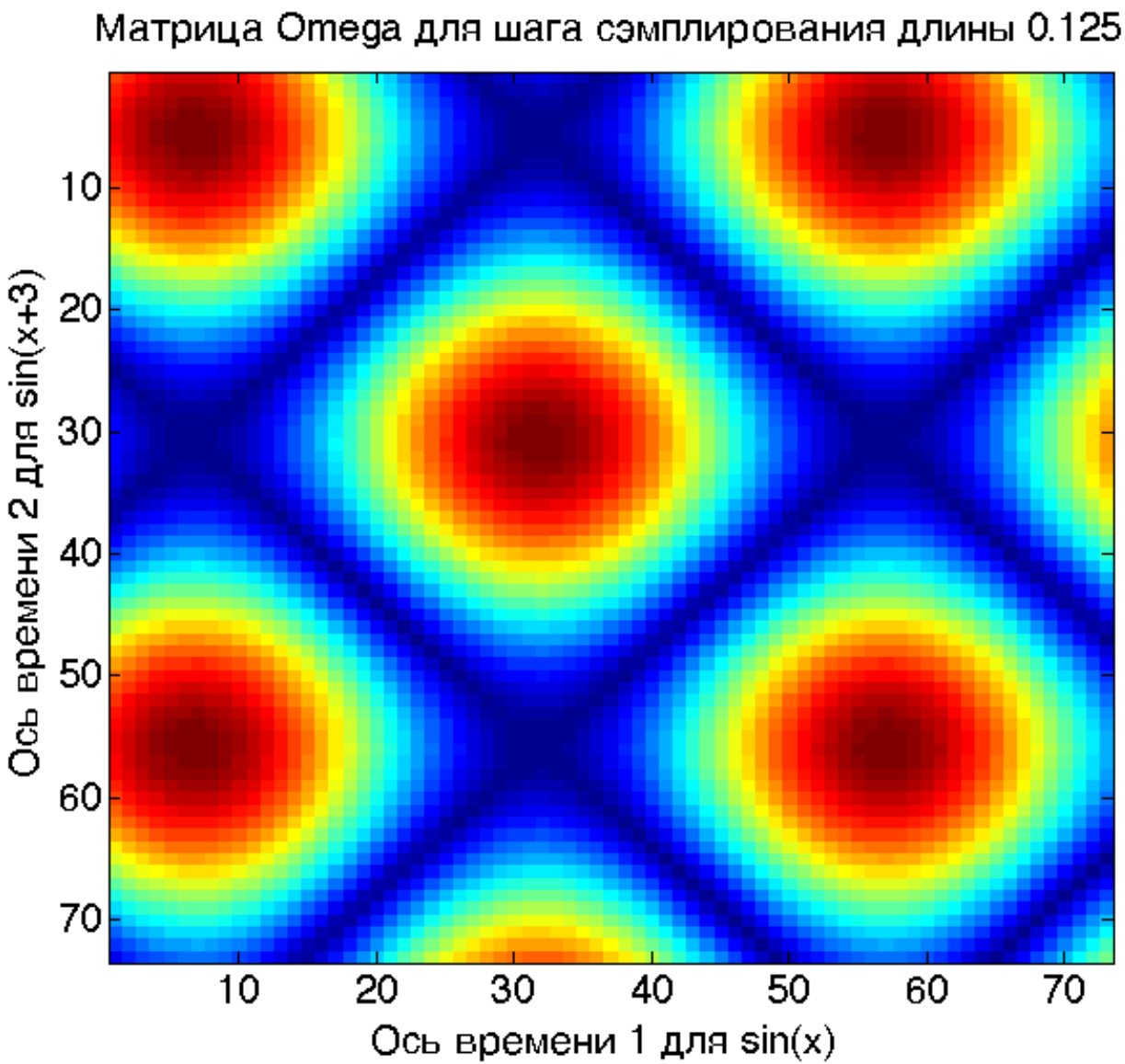
Матрица отклонений Omega



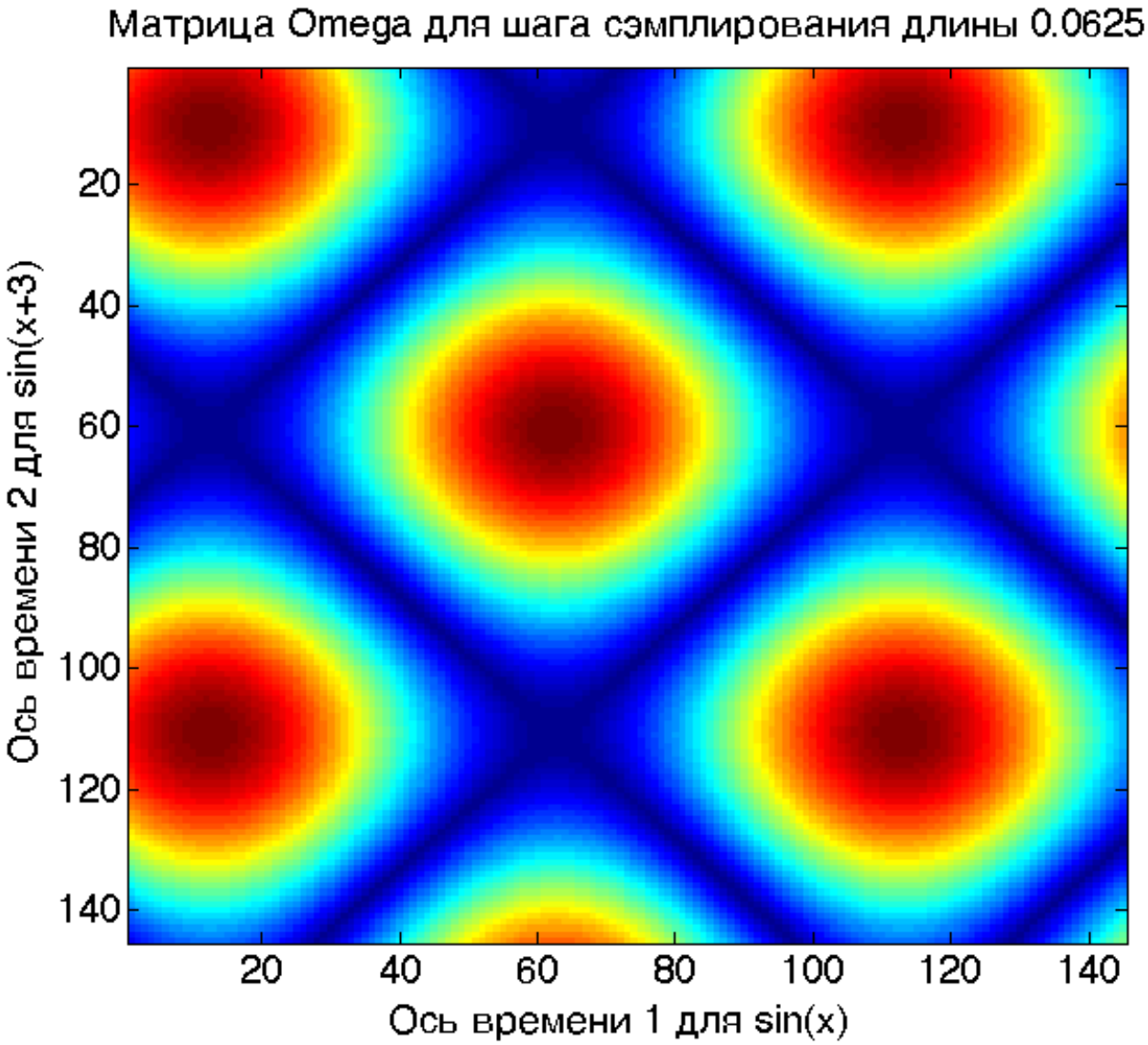
Матрица отклонений Omega



Матрица отклонений Omega



Матрица отклонений Omega



Свойства выравнивающего пути в непрерывном случае

Лемма 1.

Предположим, что $s_1(t)$ и $s_2(t)$ – два временных ряда, $\widehat{\pi}^c : t_1 \to t_2$ — выравнивающий путь между ними. При малых изменениях пути, его стоимость изменяется слабо, то есть:

$$\| \widehat{\pi}^c - \pi^c \|_C \leq \epsilon \quad \Rightarrow \quad |Cost(s_1, s_2, \widehat{\pi}^c) - Cost(s_1, s_2, \pi^c)| \leq \epsilon TL,$$

где L — константа Липшица для $s_1(t)$ и $s_2(t)$, T — граница области определения временного ряда, $\epsilon > 0$.

Свойства выравнивающего пути в непрерывном случае

Лемма 2.

Предположим, что $s_1(t)$ и $s_2(t)$ – два временных ряда, $\widehat{\pi}^c : t_1 \to t_2$ — выравнивающий путь между ними. При малых изменениях одного из временных рядов, стоимость пути изменяется слабо, то есть:

$$\| \widehat{s}_2 - s_2 \|_C \leq \epsilon \quad \Rightarrow \quad |Cost(s_1, \widehat{s}_2, \widehat{\pi}^c) - Cost(s_1, s_2, \widehat{\pi}^c)| \leq \epsilon TL,$$

где L — константа Липшица для $s_1(t)$ и $s_2(t)$, T — граница области определения временного ряда, $\epsilon > 0$.

Свойства выравнивающего пути в непрерывном случае

Предположение 1.

Выдвигается предположение об устойчивости выравнивающего пути к небольшому изменению начальных данных, то есть:

$$\forall \epsilon_1 > 0 \quad \exists \epsilon_2(\epsilon_1), \quad \forall \widehat{s_2}(t): \quad \| \widehat{s_2}(t) - s_2(t) \|_C \leq \epsilon_2 \quad \Rightarrow$$

$$\| \pi^c - \widehat{\pi^c} \|_C \quad \leq \quad \epsilon_1,$$

где π^c и $\widehat{\pi^c}$ — выравнивающие пути между $s_1(t), s_2(t)$ и $s_1(t), \widehat{s_2}(t)$ соответственно.

Задача построения пути наименьшей стоимости

Поиск выравнивающего пути

- Требуется построить решение оптимизационной задачи из определения 3:

$$\widehat{\pi}^c = \underset{\pi^c}{\operatorname{argmin}} \operatorname{Cost}(s_1^c(t_1), s_2^c(t_2), \pi^c).$$

- Предлагается искать не точное решение задачи, а его аппроксимацию среди параметрических функций.
- Сформулируем задачу в таком виде:

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmin}}\, Cost(s_1, s_2, \theta) = \underset{\theta}{\operatorname{argmin}} \int_{t_1} |s_1(t_1) - s_2(F(\theta)(t_1))| dt_1$$

где $F(\theta)$ является отображением из пространства параметров в выбранное ранее пространство параметрических функций.

Вычислительный эксперимент

Вычислительный эксперимент проводился с использованием временных рядов акселерометра мобильного телефона для 6 видов физической активности человека.

План эксперимента

- Для каждого класса строился центроид при помощи DBA.
- Для каждого временного ряда строился непрерывный аналог.
- Вычислялось расстояние между временными рядами и центроидами как в дискретном, так и в непрерывном случае.

Вычислительный эксперимент

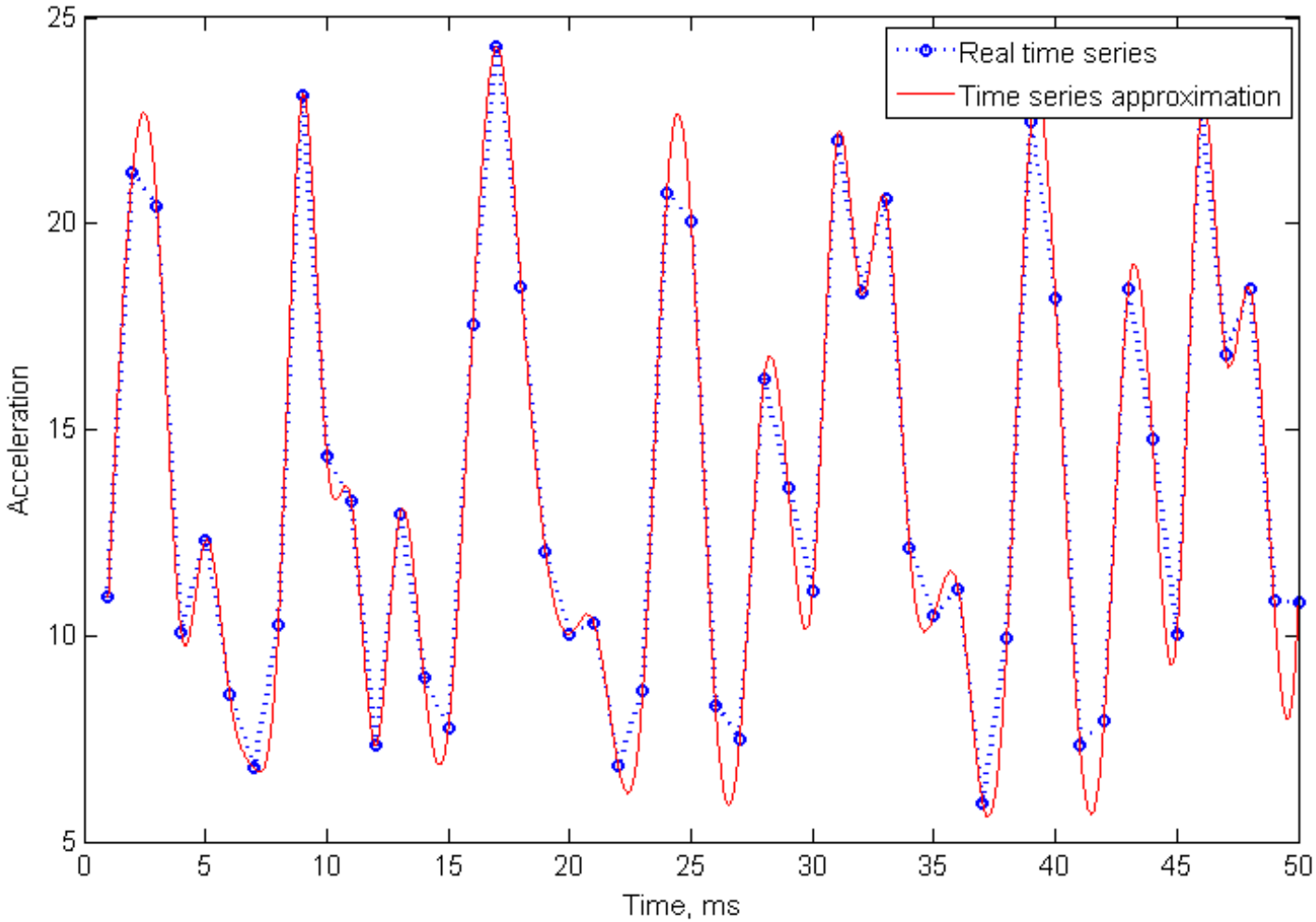


Рис.: Аппроксимация дискретного временного ряда

Вычислительный эксперимент

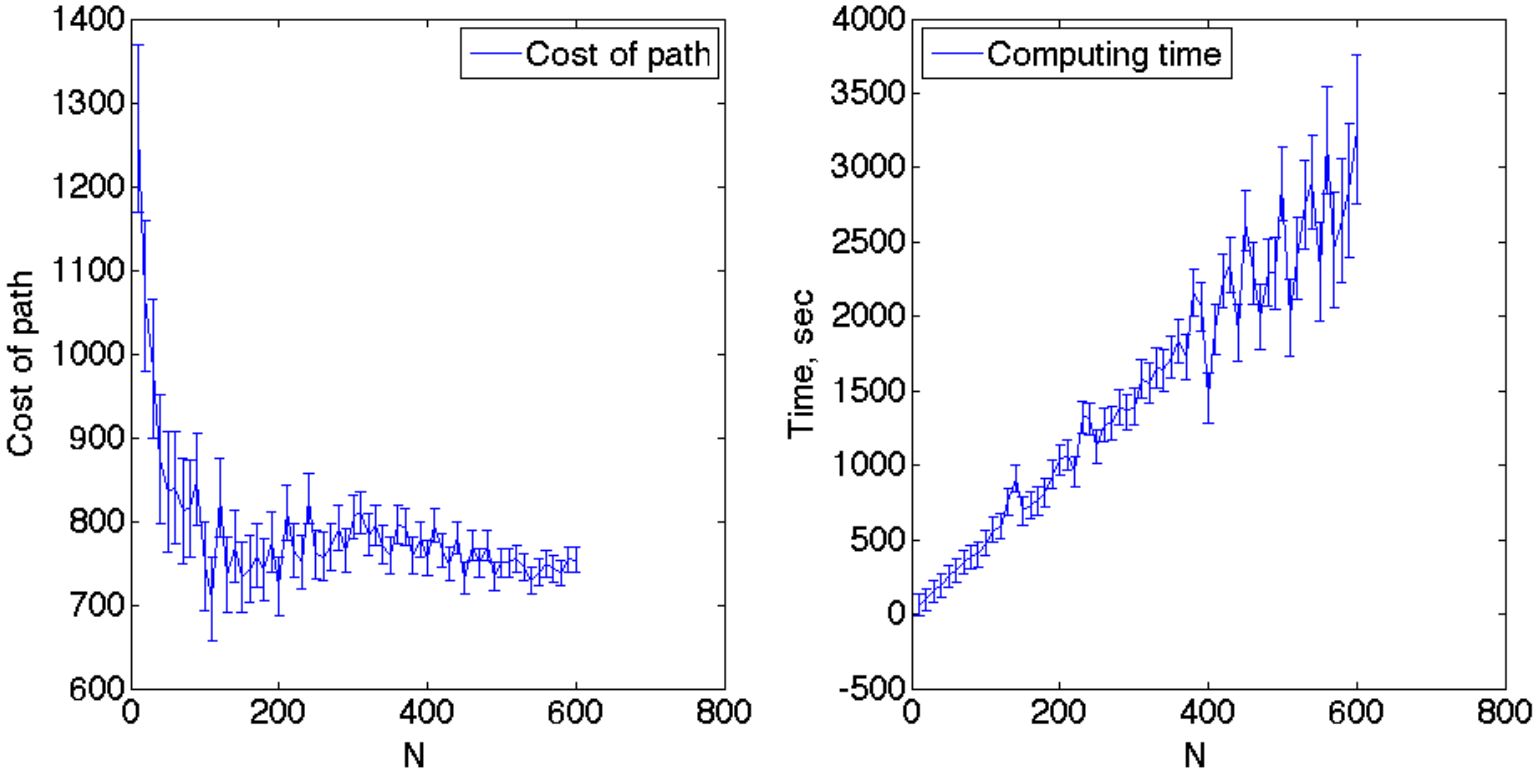


Рис.: Зависимости времени работы и стоимости пути от количества узлов сплайна

Матрица попарных расстояний

Средние значения функции расстояния между рядами класса и центроидами различных классов

Таблица: Средние расстояния между объектами различных классов и центроидами этих классов для непрерывного случая.

	Бег	Ходьба	Вверх	Вниз	Сидение	Стояние
Бег	**693**	803	811	733	1165	1143
Ходьба	676	**498**	696	610	946	927
Вверх	714	739	**696**	701	1038	1021
Вниз	591	601	653	**464**	836	804
Сидение	516	465	434	400	**6**	42
Стояние	508	441	454	366	105	**79**

Заключение

- Векторно-взвешенное выравнивание повышает качество классификации и демонстрирует важные элементы временного ряда.
- Использование непрерывных временных рядов решает задачу ресэмплирования выборки.
- Функция расстояния DTW в непрерывном случае обладает схожими со стандартной DTW свойствами.

Планируется:

- Сделать непрерывную модификацию векторно-взвешенной функции расстояния
- Использовать стохастические методы оптимизации с мультистартом как для нахождения вектора весов, так и для нахождения пути.
- Адаптировать различные ограничения классического DTW для случая непрерывных временных рядов.