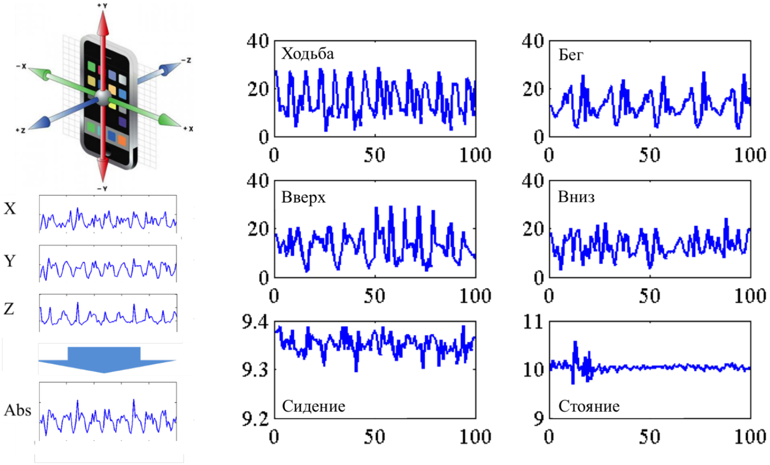
Метрическая классификация временных рядов акселерометра мобильного телефона со взвешенным выравниванием относительно центроидов классов

Гончаров Алексей Владимирович

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

58 научная конференция МФТИ

Задача метрической классификации состояния физической активности человека



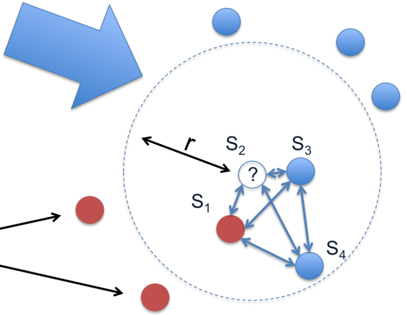
Абсолютные значения ускорения акселерометра мобильного телефона для различных классов физической активности

Метрическая классификация методом k ближайших соседей

Матрица попарных расстояний

	S_1	S_2	S_3	S_4
S_1	0	1	2	3
S_2	1	0	2	4
S_3	2	2	0	5
S_4	3	4	5	0

Пример метода k ближайших соседей при $k = 3$



Временные ряды – объекты классификации

Дана функция расстояния между временными рядами. С ее помощью построена матрица попарных расстояний.

Цель исследования

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

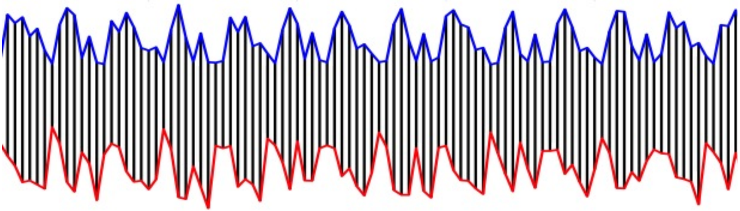
Проблема

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

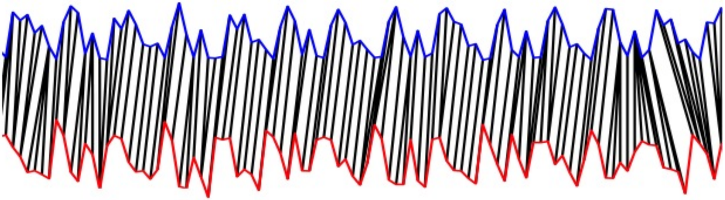
Предлагается

- Строить центроиды классов.
- Модифицировать существующую функцию.

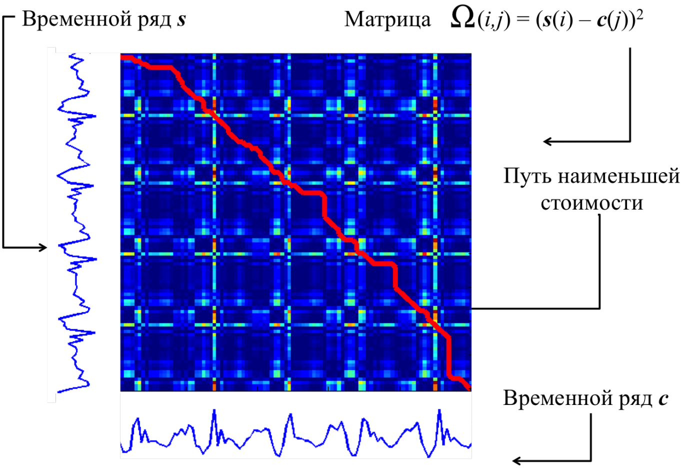
Выбор базовой функции расстояния



Евклидово расстояние между временными рядами

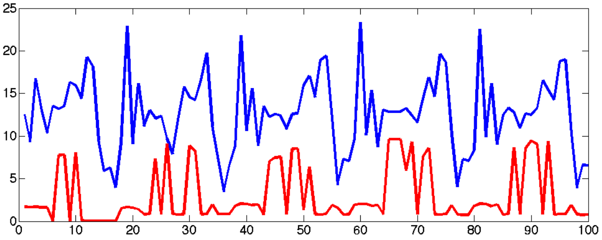


Выровненное расстояние между временными рядами



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

Модификация функции расстояния



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

1. взвесить элементы временного ряда,
2. взвесить пары элементов временных рядов.

1. *Petitjean F., Chen Y., Keogh E.*, ICDE, 2014. — метод DBA вычисления центроида класса.
2. *Гончаров А. В., Стрижов В. В.* Системы и средства информатики, 2015. — обоснования применения функции расстояния DTW в задаче классификации временных рядов.
3. *Kwapisz J. R.* 2010. — Измерение физической активности человека с помощью акселерометра. http://sourceforge.net/p/mlalgorithms/TSLearning/data /preprocessedlarge.csv

Постановка задачи классификации

Дано

$\mathfrak{D} = \{(\mathbf{s}_i, y_i)\}_{i=1}^m, \quad \mathbf{s}_i \in \mathbb{R}^n, y_i \in \{1, \dots, E\}$ — множество временных рядов и метки классов. E — число классов. Выборка разбита на обучающую $\mathfrak{D}_l$ и контрольную $\mathfrak{D}_t$.

Модель классификации f

Параметрическая функция $f(\mathbf{s}_i)$, приближающая целевую зависимость y_i. Параметры:

- множество центроидов $\mathbf{C} = \{\mathbf{c}_e\}_{e=1}^E$,
- множество векторов весов центроидов $\mathbf{W} = \{\mathbf{w}_e\}_{e=1}^E$.

Функция ошибки S модели f: частота несовпадений прогнозируемого класса и фактического,

$$S(f, \mathfrak{D}_t) = \frac{1}{|\mathfrak{D}_t|} \sum_{i=1}^{|\mathfrak{D}_t|} [f(\mathbf{s}_i) \neq y_i] \to \min_{\mathbf{C}, \mathbf{W}}.$$

Модифицированная функция расстояния

Стоимость пути π: $\mathrm{Cost}(s_1, s_2, \pi)$

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

Путь наименьшей стоимости (выравнивающий путь)

$$\hat{\boldsymbol{\pi}} = \underset{\boldsymbol{\pi}}{\mathrm{argmin}}\, \mathrm{Cost}(s_1, s_2, \boldsymbol{\pi}).$$

Функция расстояния DTW, vwDTW и mwDTW

$$\rho(s_1, s_2) = \mathrm{Cost}(s_1, s_2, \hat{\boldsymbol{\pi}}).$$

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 $\mathbf{w}_e$.

$$\mathbf{w}_e = \operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^n} \sum_{\mathbf{s}_i\in\mathfrak{D}_e} \sum_{(t,t')\in\boldsymbol{\pi}_i} \mathbf{w}_e(t)(\mathbf{s}_i(t') - \mathbf{c}_e(t))^2 + \| \mathbf{w}_e \|_2,$$

при следующих ограничениях:

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

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

Алгоритм решения задачи классификации

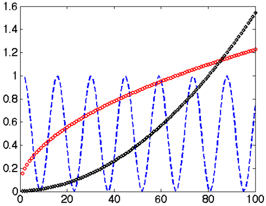
Пусть задано начальное приближение вектора весов центроида и центроид: $\mathbf{w}_e = \mathbf{1}, \quad \mathbf{c}_e = \mathbf{s}_j \in \mathfrak{D}_e \quad e = 1, \dots, E.$

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

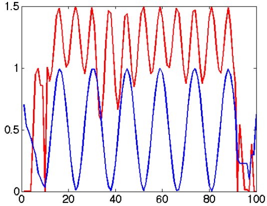
Шаг 2. Вычисление вектора весов $\mathbf{w}_e$ центроида при фиксированном центроиде $\mathbf{c}_e$.

Шаг 3. Использование полученных параметров функции расстояния между временными рядами для решения задачи классификации методом k ближайших соседей.

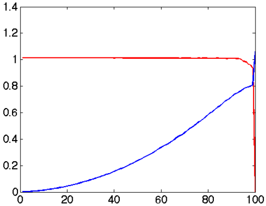
Анализ вектора весов на синтетических данных



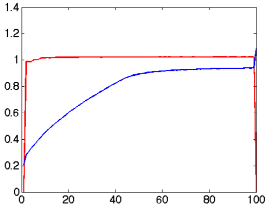
Выборка



Веса для $sin(x)$

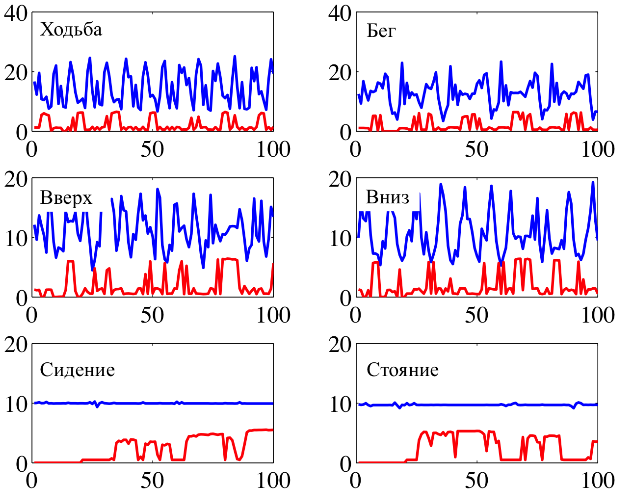


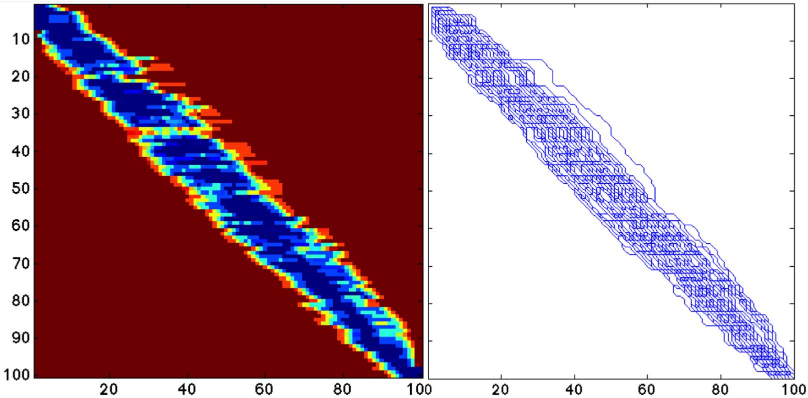
Веса для x^2



Веса для $\sqrt{x}$

Вектор весов и центроид для разных классов физической активности

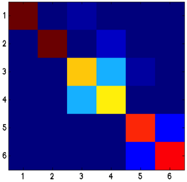




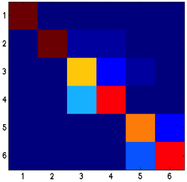
Матрица весов центроида и множество выравнивающих путей между объектами и центроидом.

Результаты классификации

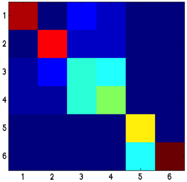
Матрица несоответствий (confusion matrix) для классификации с помощью DTW, vwDTW и mwDTW. Ходьба —1, бег —2, вверх —3, вниз — 4, сидение — 5, стояние — 6.



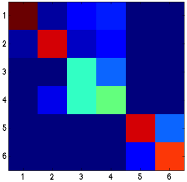
DTW



vwDTW



DTW



mwDTW

Заключение

- Использование процедуры выравнивания временных рядов позволяет значимо увеличить качество классификации состояний физической активности.
- Предложен беспереборный способ построения множества временных рядов и векторов весов. Процедуры их построения представлены в виде задачи оптимизации.

Планируется:

- Предложить набор предметных ограничений задачи построения центроидов
- Предложить метод параметрической аппроксимации пути наименьшей стоимости, имеющий физическую интерпретацию.