Министерство образования и науки Российской Федерации
Московский физико-технический институт (государственный университет)
Факультет управления и прикладной математики
Вычислительный центр им. А. А. Дородницына РАН
Кафедра «Интеллектуальные системы»

Кузнецова Маргарита Валерьевна

Классификация временных рядов с использованием инвариантных преобразований

010990 — Математические и информационные технологии

Выпускная квалификационная работа магистра

Научный руководитель:
д. ф.-м. н. Стрижов Вадим Викторович

Москва
2015

Содержание

Аннотация

В данной работе рассматривается задача многоклассовой классификации временных рядов. Модель классификации включает инвариантное преобразование относительно монотонного преобразования шкалы времени и аффинного преобразования шкалы значений. Инвариантное преобразование построено с помощью метода динамического выравнивания временных рядов. Временные ряды, принадлежащие одному классу, преобразуются относительно центроида этого класса. Для нахождения центроида предлагается двухшаговый итерационный подход. Задача классификации решается с помощью метода ближайшего соседа с обученной метрикой (Large Margin Nearest Neighbour). Предложенная модель классификации протестирована на временных рядах, описывающих движение человека и содержащих показания акселерометра. Произведено сравнение с методом ближайшего соседа, не использующем преобразование.

Ключевые слова: метрическая классификация временных рядов, динамическое выравнивание, инвариантное преобразование, центроид.

Введение

Актуальность темы. Современное состояние акселерометров беспроводных устройств позволяет быстро получить большие объемы данных, описывающие физическую активность человека. Анализ временных рядов, содержащих показания этой активности, позволяет решать задачи, связанные с идентификацией движений человека и анализом его поведения.

В данной работе строится модель классификации временных рядов. Метки классов — ходьба, бег, подъем по лестнице, спуск по лестнице, сидение, лежание. Модель включает инвариантное преобразование относительно монотонного преобразования шкалы времени и аффинного преобразования шкалы значении. Данное преобразование строится относительно некоторого эталонного объекта класса физической активности (шаг, бег, ходьба по лестнице). Целью данного преобразования является учет разнообразия изучаемых временных рядов, которое возникает из-за зависимости временного ряда от конституции и возраста человека, его физической подготовки, нерегулярности периодики и различных временных масштабов его движений.

Цель работы. Целью данной работы является построение модели классификации временных рядов, содержащей инвариантное преобразование для учета растяжений и сдвигов временных рядов относительно шкалы времени и значений.

Методы исследования. Для достижения поставленной цели используется метод динамического выравнивания временных рядов (Dynamic Time Warping) [1] относительно центроида класса и алгоритм метрической классификации поиска ближайших соседей с обученной метрикой [2]. Для нахождения центроида класса используется итерационный алгоритм DTW Barycenter Averaging [3].

Основные положения, выносимые на защиту.

1. Двухшаговыи алгоритм классификации временных рядов.

2. Исследование свойств инвариантного преобразования относительно монотонного преобразования шкалы времени и аффинного преобразования шкалы значении.

Научная новизна. Разработан двухшаговый алгоритм классификации временных рядов, содержащих показания физической активности человека. Предложено инвариантное преобразование, учитывающее растяжения и сдвиги временных рядов относительно шкалы времени и значений, и, как следствие,

Практическая значимость. Предложенный в работе алгоритм позволяет проводить идентификацию движений человека, и, как следствие, решать задачи в диагностике его физических отклонений.

Степень достоверности и апробация работы. Достоверность результатов подтверждена экспериментальной проверкой полученных методов на реальных задачах. Результаты работы докладывались и обсуждались на 57-й международной научной конференции МФТИ 24-29 ноября 2014г.

Публикации по теме дипломной работы. Основные результаты по теме диплома изложены в статье:

1. М. В. Кузнецова, В. В. Стрижов Локальное прогнозирование временных рядов с использованием инвариантных преобразований // подано в Информационные технологии.

Обзор литературы.

Для решения задачи классификации временных рядов применяются многие методы: нейронные сети [4], машины опорных векторов [5] и композиции различных классификаторов(Ada Boost, Random Forest) [6], байесовские классификаторы [7], метод k ближайших соседей [8, 9]. Последний метод основан на понятии близости объектов, где близость формализуется введением метрики в пространстве признаковых описаний. В качестве функциий расстояния между временными рядами может быть задана различными способами: используется Евклидово расстояние [10], путь наименьшей стоимости, полученный с помощью динамического выравнивания временных рядов(DTW) [11], расстояние, основанное на нахождение наибольшей общей последовательности [12], Edit Distance with Real Penalty (ERP) [13], Edit Distance on Real sequence (EDR) [14], DISSIM [15], Sequence Weighted Alignment model (Swale) [16],

Spatial Assembling Distance (SpADe) [17] и другие. Во многих работах по классификации временных рядов предлагается использовать метрическое обучение. [18, 19]. Задачу метрического обучения можно сформулировать следующим образом. Задана выборка $\mathfrak{D} = (\mathbf{s}_i, y_i)_{i=1}^N$, где $\mathbf{s}_i = [s_{i1}, \ldots, s_{iT}]$ — временной ряд длины T, $y_i \in Y$ — метка класса. Между объектами выборки вводится метрика:

$$\rho_{\mathbf{A}}(\mathbf{s}_i, \mathbf{s}_j) = \sqrt{(\mathbf{s}_i - \mathbf{s}_j)^\intercal \mathbf{A}^{-1} (\mathbf{s}_i - \mathbf{s}_j)},$$

где $\mathbf{A}$ — положительно полуопределенная матрица. Водятся следующие множества:

$\mathcal{S} = \{(\mathbf{s}_i, \mathbf{s}_j) | y_i = y_j\}$ — пары похожих (similar) описаний объектов,

$\mathcal{D} = \{(\mathbf{s}_i, \mathbf{s}_j) | y_i \neq y_j\}$ — пары непохожих (dissimilar) описаний объектов.

Тогда задача оптимизации будет иметь следующий вид:

$$\min_{\mathbf{A}} l(\mathbf{A}, \mathcal{S}, \mathcal{D}),$$

где $l(\mathbf{A}, \mathcal{S}, \mathcal{D})$ — функция потерь.

В работе [20] на множестве временных рядов вводится расстояние Махалонобиса [21]. Матрица $\mathbf{A}$ представляется в виде произведения $\mathbf{A} = \mathbf{L}\mathbf{L}^\intercal$. В обоих случаях происходит максимизация функционала:

$$\mathbf{L} = \arg\max_{\mathbf{L}} tr(\mathbf{L}^\intercal(\mathcal{D} - \mathcal{S})\mathbf{L}), \mathbf{L}^\intercal\mathbf{L} = \mathbf{I}. \quad (1)$$

В работах [22, 23] временные ряды классифицируются с помощью метода Large Margin Nearest Neihbour [2, 24]. В работе [25] предлагается метод метрического обучения, где в качестве расстояния между временными рядами рассматривается расстояние между их параметрами.

Для решения задач классификации важно определить эталонный объект класса, описывающий класс наиболее лучшим образом. Будет называть такой объект центроидом класса. Задача нахождения центроида является задачей степенной сложности. Она решается несколькими способами — иерархической кластеризацией [26], методами k-means [27] и k-medoids [28], методом нахождение наилучшего выравнивания, построенного по пути наименьшей стоимости нескольких временных рядов [3]. Один методов выбора центроида основан на алгоритме DTW Barycenter Averaging [3], который будет более подробно описан в соответствующем разделе.

Для учета растяжений и сдвигов временных рядов относительно шкалы времени вводится преобразование сегментов. В работе [29] рассмотренно понятие инвариантного преобразования и выбор наиболее подходящего преобразования для решения задачи прогнозирования. В работе [30] вводится понятие параметрического инвариантного преобразования. В работе [31,33] с помощью метода динамического выравнивания вводится функция сдвига между кривыми и затем рассматривается выравнивание кривых относительно некоторой эталонной, с целью получения кривой общей формы. В работе [32] рассматривается парное выравнивание кривых относительно друг друга и получение устойчивой оценки для функции деформации.

1 Постановка задачи

Задана выборка $\mathfrak{D} = (\mathbf{s}_i, y_i)_{i=1}^N$, где $\mathbf{s}_i = [s_{i1}, \ldots, s_{iT}]$ — временной ряд длины T, $y_i \in Y$ — метка класса. Выборка разбита на обучающую $\mathfrak{D}_l$ и контрольную $\mathfrak{D}_t$. Решается задача метрической классификации объектов, т.е строится отображение:

$$f : \mathbf{s}_i \mapsto y_i.$$

Введем функцию ошибки классификации:

$$Q(f, \mathfrak{D}_l) = \frac{1}{|\mathfrak{D}_l|} \sum_{i=1}^{|\mathfrak{D}_l|} ([f(\mathbf{s}_i) \neq y_i)])$$

Для решения задачи классификации предлагается следующая процедура:

- выделить центроиды каждого класса, т.е построить отображение

  $$h : \boldsymbol{\mu}_k \mapsto y_i,$$

  где функция h устанавливает взаимно-однозначное соответствие между меткой каждого класса y_i центроидом $\boldsymbol{\mu}_k$.

- выровнять временные ряды каждого класса относительно центроида:

  $$a : \mathbf{s}_i \mapsto \mathbf{x}_i.$$

- ввести на полученном множестве объектов метрику Махаланобиса:

  $$\rho_{\mathbf{A}}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^\intercal \mathbf{A}^{-1} (\mathbf{x}_i - \mathbf{x}_j)},$$

  где $\mathbf{A}$ — положительно полуопределенная матрица.

- решить задачу метрической классификации в полученном пространстве.

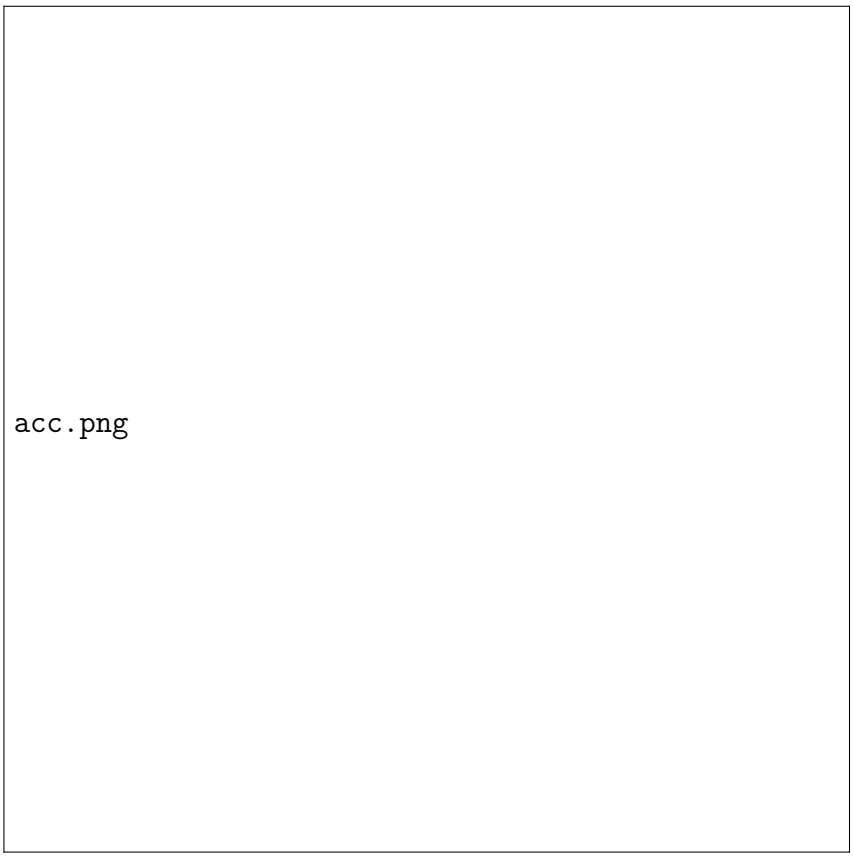


Рис. 1: Примеры временных рядов, рассматриваемых в работе.

2 Выделение эталонных объектов

Для построения инвариантного преобразования предлагается трансформировать каждый временной ряд класса относительно центроида этого класса. Под центроидом класса будем понимать некоторый типичный паттерн этого класса, учитывающий особенности временных рядов, входящих в класс. Введем формальное определение центроида $\boldsymbol{\mu}_k$ класса k из Y. Пусть $\mathfrak{D}_k$ — множество элементов из $\mathfrak{D}$, принадлежащих одному классу k из Y. Центроидом множества векторов $\mathfrak{D}_k = \{\mathbf{s}_i | y_i = k\}_{i=1}^m$ по расстоянию ρ назовем вектор $\boldsymbol{\mu}_k \in \mathbb{R}^n$ такой, что:

$$\boldsymbol{\mu}_k = \arg\min_{\boldsymbol{\mu}_k \in \mathbb{R}^n} \sum_{\mathbf{s}_i \in \mathfrak{D}_k} \rho^2(\mathbf{s}_i, \boldsymbol{\mu}_k).$$

В качестве расстояния ρ для рассчета центроида будем использовать стоимость пути наименьшей стоимости.

2.1 Путь наименьшей стоимости

Введем понятие пути наименьшей стоимости между временными рядами $\mathbf{s}_i$ и $\mathbf{s}_{i'}$:

$$\mathbf{s}_i = [s_{i1}, \ldots, s_{il}]^\intercal,\ \mathbf{s}_{i'} = [s_{i'1}, \ldots, s_{i'l}]^\intercal.$$

Зададим матрицу $\mathbf{\Omega}$ с элементами-парами из $(l \times l)$ — декартова произведения, квадрата множества $\{1, \ldots, l\}^2$. Обозначим путь $\boldsymbol{\pi}$ в матрице $\mathbf{\Omega}$ — последовательность

$$\boldsymbol{\pi} = (\pi_1(1), \pi_2(1)), (\pi_1(2), \pi_2(2)), \ldots, (\pi_1(N), \pi_2(N)),$$

где N — длина пути, которая удовлетворяет условию:

$$l \leq N < 2l - 1.$$

При построении пути учитываются:

1. Граничные условия. Начало и конец пути $\boldsymbol{\pi}$ находятся на диагонали в противоположных углах $\mathbf{\Omega}$, т. е.

$$1 = \pi_1(1) \leq \pi_1(2) \leq \ldots \leq \pi_1(N),$$

$$1 = \pi_2(1) \leq \pi_2(2) \leq \ldots \leq \pi_2(N).$$

2. Непрерывность. В шаге пути $\boldsymbol{\pi}$ участвуют только соседние элементы матрицы (включая соседние по диагонали). $\pi_1(j+1) \leq \pi_1(j)$ и $\pi_2(j+1) \leq \pi_2(j)$.

3. Монотонность. Точки π монотонно перемещаются во времени. $(\pi_1(j+1) - \pi_1(j)) + (\pi_2(j+1) - \pi_2(j)) \geq 1$.

Рассмотрим множество всевозможных путей $\{\boldsymbol{\pi}\}$, удовлетворяющее вышеуказанным ограничениям. В этом множестве требуется найти путь, проходящиий по элементам матрицы $\mathbf{\Omega}$ такой что:

$$\delta_{\boldsymbol{\pi}}(\mathbf{s}_i, \mathbf{s}_{i'}) = \min_{\boldsymbol{\pi} \in \mathbf{\Omega}} \frac{1}{N} \sum_{i=1}^{N} \pi(n). \tag{2}$$

Знаменатель N нужен для того, чтобы учесть длину пути $\boldsymbol{\pi}$.

Для получения пути наименьшей стоимости в данной работе используется метод динамического выравнивания шкалы времени, который рекурсивно находит длину

пути наименьшей стоимости по матрице γ с помощью алгоритма динамического программирования:

$$\delta_{\boldsymbol{\pi}}(\mathbf{s}_i, \mathbf{s}_{i'}) = d_{L_p}(\mathbf{s}_i, \mathbf{s}_{i'}) + \min(\gamma_{i,j-1}, \gamma_{i-1,j}, \gamma_{i-1,j-1}), \text{ где } d_{L_p} - L_p \text{ норма.}$$

Таким образом

$$\boldsymbol{\mu}_k = \arg\min_{\boldsymbol{\mu}_k \in \mathbb{R}^n} \sum_{\mathbf{s}_i \in \mathfrak{D}_k} \delta^2(\mathbf{s}_i, \boldsymbol{\mu}_k). \tag{3}$$

2.2 Выбор центроида класса

Для решения задачи (3) предлагается использовать алгоритм DBA (DTW Barycenter Averaging), который минимизирует сумму квадрата расстояний от центроида класса до временных рядов класса. Эта сумма формируется как расстояние от каждой координаты центроида до каждой поставленной ей в соответствие по пути наименьшей стоимости координаты временного ряда. Каждой координате центроида может соответствовать несколько координат временного ряда (следует из определения пути наименьшей стоимости). Таким образом, каждая координата центроида обновляется исходя из вклада тех координат, которые соответствуют ей в рассматриваемом временном ряде (значения по соответствующим координатам усредняются). Таким образом, алгоритм DBA состоит из двух итеративно повторяющихся шагов:

- 1-*шаг.* Рассчитывается путь наименьшей стоимости между всеми временными рядами, принадлежащими определенному классу и начальным приближением центроида.
- 2-*шаг.* Происходит перерасчет каждой координаты центроида в соответствии с соответствующими ей координатами каждого временного ряда.

Алгоритм итеративно повторяется до сходимости. На первом шаге происходит расчет пути наименьшей стоимости между временным рядом класса и текущим приближением центроида. Сложность рассчета пути наименьшей стоимости — $O(l^2)$, где l — длина временного ряда. Тогда сложность алгоритма на первом шаге — $O(ml^2)$, где m — количество временных рядов в классе. На втором шаге каждая координата центроида обновляется по соответствующим ей координатам m временных рядов, таким образом каждой координате центроида ставится в соответствие ml координат.

Сложность алгоритма на втором шаге — $O(ml)$. Таким образом, итоговая сложность алгоритма DBA — $O(Iml^2)$, где I — количество итераций. На рис. 2 показан выделенный алгоритмом DBA центроид класса.

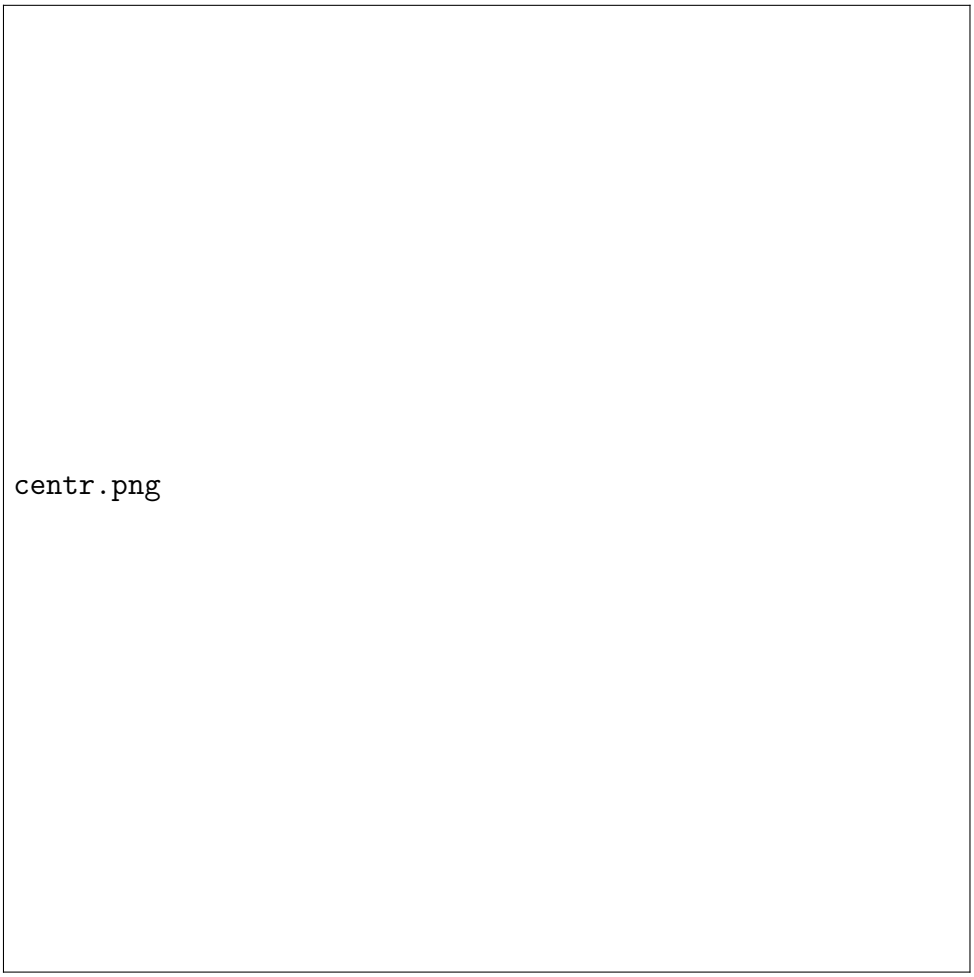


Рис. 2: Пример центроида класса.

3 Инвариантная трансформация временного ряда

Инвариантным преобразованием назовем такое преобразование a временного ряда $\mathbf{s}_i$, которое сохраняет эквивалентность на классах, т.е. если $\mathbf{s}_i \in y_k$, то и $a(\mathbf{s}_i) \in y_k$. Преобразование строится по пути наименьшей стоимости. Для формализации определения преобразования введем определение проекции пути наименьшей стоимости π_x, π_y — это отображение подмножества $\boldsymbol{\pi}$ декартова произведения $L^2 = \{1, \ldots, l\}^2$ в его сомножители L. Определим следующие множества:

- $L = \{1 \dots l\}$ — множество индексов временного ряда $\mathbf{s}_i$,

- $P = \{1 \dots N\}$ — множество индексов пути наименьшей стоимости $\boldsymbol{\pi}$.

Определим f и g — сюръективные отображения элементов множества P в элементы множества L

$$f, g : P \to L.$$

Определим $F : L \to 2^P$ как функцию, возвращающую все прообразы индексов пути

$$F(i) = \{p \mid f(p) = i\}, \text{ для любого } i \in L : F(i) \neq \emptyset.$$

Определим сюръективное отображение $G : 2^P \to L, G(j) = \{g(p) \mid p \in j\}$.

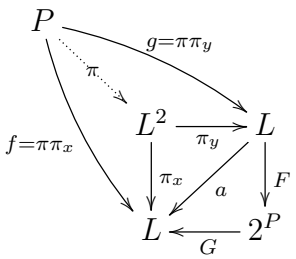
Определим функцию усреднения по значениям полученных индексов

$$avg : 2^L \to \mathbb{R}, \text{ где } avg(G) = \frac{(\sum_{j \in G} s_2(j))}{|G|}.$$

Таким образом, итоговое преобразование a временного ряда $\mathbf{s}_i$ — композиция:

$$a = avg \cdot G \cdot F : \mathbf{s}_i \mapsto \mathbf{x}_i.$$

Покажем преобразование a на коммутативной диаграмме:



На рис. 3 изображен исходный временной ряд $\mathbf{s}_i$ и преобразованный $\mathbf{x}_i$.

Рис. 3: Пример исходного и трансформированного временного ряда.

4 Метрическая классификация

4.1 Обучение метрики

На полученном множестве $\{\mathbf{x}_i, y_i\}_{i=1}^N$,будем решать задачу метрического обучения методом Large Margin Nearest Neighboor (LMNN). Зная метки классов y_i и соседей в каждом классе, LMNN подбирает такую матрицу $\mathbf{L}$ для расстояния Махалонобиса:

$$\rho(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^\intercal \mathbf{L}^\intercal \mathbf{L}(\mathbf{x}_i - \mathbf{x}_j),$$

которая улучшает классификацию методом k ближайших соседей. Индикатор $\eta_{ij} \in \{0, 1\}$ показывает, лежит ли $\mathbf{x}_j$ в одном классе с $\mathbf{x}_i$. Таким образом, $\mathbf{L}$ должно быть

выбрано так, чтобы минимизировать:

$$\psi_1(\mathbf{L}) = \sum_{ij} \eta_{ij} \|\mathbf{L}(\mathbf{x}_i - \mathbf{x}_j)\|^2.$$

Кроме того, расстояние от $\mathbf{x}_i$ до его "соседа"$\mathbf{x}_j$ должно быть меньше, чем расстояние до его "ложного"соседа $\mathbf{x}_j$. Для этого минимизируется следующий функционал:

$$\psi_2(\mathbf{L}) = \sum_{ijl} (1 - y_{il})[1 + \|\mathbf{L}(\mathbf{x}_i - \mathbf{x}_j)\|^2 - \|\mathbf{L}(\mathbf{x}_i - \mathbf{x}_l)\|^2]_+,$$

где $z_+ = \max(z, 0)$. Краевая ошибка будет равна 0, только тогда $\rho(\mathbf{x}_i, \mathbf{x}_l) \geq 1 + \rho(\mathbf{x}_i, \mathbf{x}_j)$, т.е ширина разделяющей полосы между классами равна 1. Итоговая функция ошибки определяется выражением:

$$\psi(\mathbf{L}) = \psi_1(\mathbf{L}) + c\psi_2(\mathbf{L}), \text{ где } c \text{ — константа.} \tag{4}$$

Так как матрица $\mathbf{M} = \mathbf{L}^\intercal\mathbf{L}$ — положительно полуопределенная, то можно свести (4) к задаче выпуклой оптимизации:

$$\min \sum_{ij} \eta_{ij}(\mathbf{x}_i - \mathbf{x}_j)^\intercal \mathbf{M}(\mathbf{x}_i - \mathbf{x}_j) + c\sum_{ij}(1 - y_{il})\xi_{ijl}, \text{ где } \xi_{ijl} \text{ — свободная переменная;}$$

$$(\mathbf{x}_i - \mathbf{x}_l)^\intercal \mathbf{M}(\mathbf{x}_i - \mathbf{x}_l) - (\mathbf{x}_i - \mathbf{x}_j)^\intercal \mathbf{M}(\mathbf{x}_i - \mathbf{x}_j) \geq 1 - \xi_{ijl};$$

$$\xi_{ijl} \geq 0;$$

$$\mathbf{M} \succeq 0.$$

4.2 Алгоритм классификации

Для вновь поступившего временного ряда контрольной выборки $\mathfrak{D}^t$ $\mathbf{s}_q$ вычислим выравнивания ко всем центроидам и выберем в качестве эталонного объекта для этого ряда то, на котором будет достигаться наилучшее значение выравнивания:

$$\hat{\boldsymbol{\mu}_k} = \arg \min_{k=1,\ldots,K} Q(\boldsymbol{\mu}_k, \mathbf{s}_q).$$

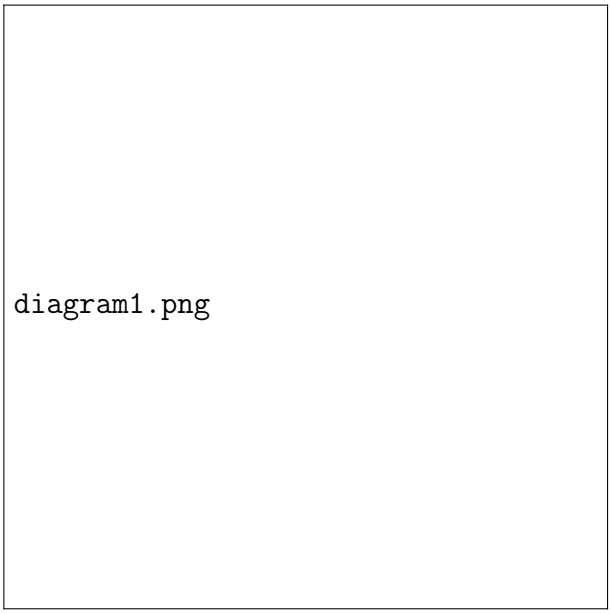
Преобразуем $\mathbf{s}_q$ к выбранному центроиду $\boldsymbol{\mu}_k$: $\mathbf{x}_q = a(\mathbf{s}_q)$. Решим задачу классификации методом k ближайших соседей с обученной метрикой. Расположим элементы обучающей выборки $\mathbf{x}_1, ..., \mathbf{x}_l$ в порядке возрастания расстояний до $\mathbf{x}_q$:

$$\rho(\mathbf{x}_q, \mathbf{x}_{\mathbf{x}_q}^{(1)}) \leq \rho(\mathbf{x}_q, \mathbf{x}_{\mathbf{x}_q}^{(2)}) \leq \ldots \leq \rho(\mathbf{x}_q, \mathbf{x}_{\mathbf{x}_q}^{(l)}).$$

Будем относить объект $\mathbf{x}_q$ к тому классу, элементов которого окажется больше среди k его ближайших соседей $\mathbf{x}_{\mathbf{x}_q}^{(i)}$:

$$\hat{y}_{\mathbf{x}_q}(\mathbf{x}_q, \mathfrak{D}^l, k) = \underset{y \in Y}{\operatorname{argmax}} \sum_{i=1}^{k} [y_{\mathbf{x}_q}^{(i)} = y].$$

Таким образом, всю предложенную процедуру классификации можно изобразить на схеме:



5 Вычислительный эксперимент

Для иллюстрации работы предложенных алгоритмов проведем вычислительный эксперимент на данных акселерометра мобильного телефона.

Для эксперимента использовался набор данных USC-HAD. В наборе содержатся показания акселерометра и гироскопа MotionNode, закрепленного на передней стороне правого бедра участников, снятые во время выполнения ими 12 видов двигательной активности: 1) ходьба вперед; 2) ходьба по кругу налево; 3) ходьба по кругу направо; 4) ходьба вверх по лестнице; 5) ходьба вниз по лестнице; 6) бег вперед; 7) прыжки на месте; 8) небольшие движения в положении сидя; 9) небольшие движения в положении стоя; 10) небольшие движения в положении лежа; 11) небольшие движения в положении стоя при подъеме на лифте; 12) небольшие движения в положении стоя при спуске на лифте. Всего участвовало 14 человек, каждый из которых

выполнил по 5 подходов каждого вида. Показания снимались с трех осей акселерометра и трех осей гироскопа по отдельности. Направления осей по отношению к телу участников были постоянны. Таким образом, объектом в данном случае выступает совокупность шести временных рядов. В существующих работах по классификации активности нет однозначной позиции, рассматривать ли показания каждой из осей отдельно, или агрегировать все оси в один временной ряд. В данной работе используется сумма квадратов осей акселерометра — величина результирующего ускорения. Диапазон измерений используемого акселерометра — $\pm$ 6g, диапазон частот — 100 Hz.

Разделим данные на тестовую и тренировочную выборки. Для этого случайным образом выберем 25 процентов исходных объектов, которые не будут участвовать в обучении. Оптимальное значение параметра k определяется по критерию скользящего контроля с исключением объектов по одному (leave-one-out, LOO).

$$E(k, \mathfrak{D}^l) = \sum_{i=1}^{l} [y_{\mathbf{x}_j}(\mathbf{x}_j, {}^l \setminus \{\mathbf{x}_j\}, k) \neq y_i] \to \min_{k \in \mathbb{N}}.$$

На рис. 4 показана величина ошибки, вычисленной при помощи скользящего контроля, в зависимости от различных k. Ошибка растет при увеличении числа соседей.

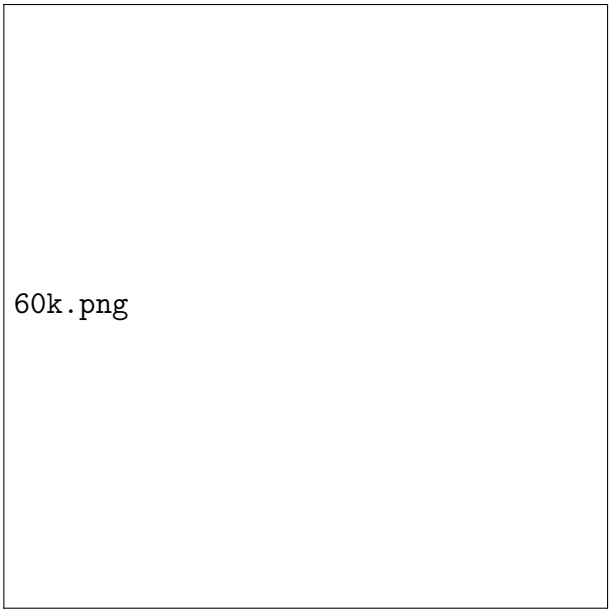


Рис. 4: Зависимость ошибки от величины k.

Видно, что минимум достигается при $k = 2$. Величина ошибки считалась на преобразованных объектах.

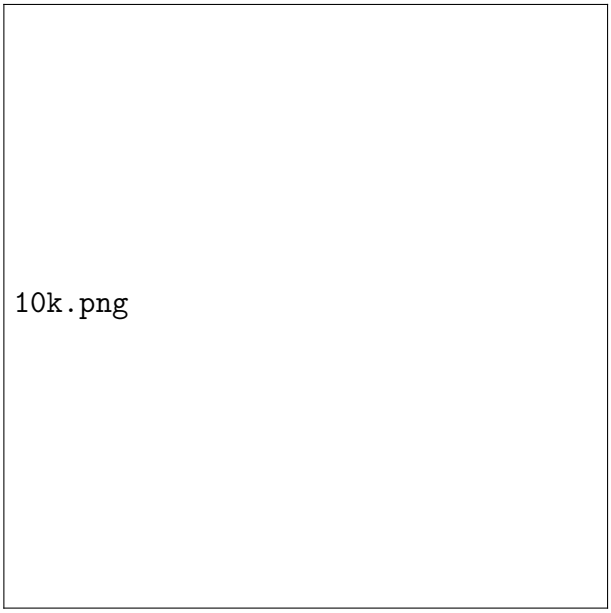


Рис. 5: Зависимость ошибки от величины k.

Отсюда можно сделать вывод, что выборка хорошо разделима и объекты хорошо классифицируются по небольшому количеству соседей. В вычислительном эксперименте исследовалась точность предлагаемой процедуры классифкации. В таблице 1 оценка эффективности предложенного алгоритма в сравнении с kNN.

Таблица 1: Сравнение результатов работы алгоритмов.

Точность, %	Бег	Ходьба	Вверх	Вниз	Сидение	Стояние
Предл. алгоритм	89	85	79	81	94	93
kNN	81	80	69	71	87	88

Выведем матрицы невязок (confusion matrix) для оценки ошибок классификации в случае с kNN без использования преобразований и с его использованием.

m1.png	m2.png

Рис. 6: Матрицы невязок без использования преобразования(слева) и с его использованием(справа).

По ним можно видеть, как сокращается количество ошибок попадания в другие классы при использовании преобразования. Наиболее спорными классами являются подъем вверх и вниз по лестнице, так как временные ряды этих классов довольно сильно похожи на временные ряды классов бег и ходьба. Однако, при использовании преобразования, качество распознавания этих классов улучшается и становится меньше ошибок классификации в других классах.

Заключение

В работе предложен двухшаговый алгоритм классификации временных рядов, содержащих показания физической активности человека. Алгоритм включает инвариантное преобразование относительно монотонного преобразования шкалы времени и аффинного преобразования шкалы значении. Введение данного преобразования позволяет учесть разнообразия изучаемых временных рядов, которое возникает из-за зависимости временного ряда от конституции и возраста человека, его физической подготовки, нерегулярности периодики и различных временных масштабов его движений. Преобразование строится относительно эталонного объекта класса — центроида этого класса. Предложен двухшаговыи итерационныи подход нахождения центроида. Исследованы свойства инвариантного преобразования. Исследовано влияние выделения центроидов класса на алгоритм классификации.

Работа предложенных алгоритмов проиллюстрирована на данных, содержащих показания акселерометра. Произведено сравнение результаты работы предложеного алгоритма с алгоритмом поиска ближайших соседей, не использующего предложенное преобразование.

6 Литература

Список литературы

[1] Keogh E. J., Pazzani M. J. Derivative dynamic time warping // In First SIAM International Conference on Data Mining, 2001.

[2] Weinberger K. and Saul L. Distance Metric Learning for Large Margin Nearest Neighbor Classification. // The Journal of Machine Learning Research, 2009. Vol. 10, Pp. 207–244.

[3] Petitjean F., Ketterlin A., Gançarski P. A Global Averaging Method for Dynamic Time Warping, with Applications to Clustering. // Pattern Recognition, 2011. Vol. 44, No. 3, Pp. 678–693.

[4] Nanopoulos A., Rob Alcock R., Manolopoulos Y. Feature-based Classification of Time-series Data. // International Journal of Computer Research, 2001. Vol. 10, Pp. 49–61.

[5] Rodriguez J., Alonso C., Maestro J. Support vector machines of interval-based features for time series classification. // Knowledge-Based Systems, 2005. Vol. 18, No. 4, Pp. 171–178.

[6] Rodriguez J., Kuncheva L. Time series classification: Decision forests and SVM on interval and DTW features. // Proceedings of the Workshop on Time Series Classification, 13th International Conference on Knowledge Discovery and Data Mining. 2007.

[7] Povinelli R., Johnson M., Lindgren A., Ye J. Time Series Classification Using Gaussian Mixture Models of Reconstructed Phase Spaces. // IEEE Transactions on Knowledge and Data Engineering, 2004. Vol. 16, No. 6, Pp. 779–783.

[8] Cover T., Hart P. Nearest neighbor pattern classification. // IEEE Transactions on Information Theory, 1967. Vol. 13, No. 1, Pp. 21–27.

[9] Воронцов К. В. Лекции по метрическим алгоритмам классификации. http://www.ccas.ru/voron/download/MetricAlgs.pdf (дата обращения: Июнь 22, 2015)

[10] Faloutsos C., Ranganathan M., Manolopoulos Y. Fast Subsequence Matching in Time-series Databases. // SIGMOD '94 Proceedings of the 1994 ACM SIGMOD international conference on Management of data. 1994. Vol. 23, No. 12, Pp. 419–429.

[11] Keogh E., Ratanamahatana C. A. Exact indexing of dynamic time warping. // Knowledge Information System, 2005. Vol. 7, No. 3, Pp. 358–386.

[12] Vlachos M., Gunopulos D., Kollios G. Discovering similar multidimensional trajectories. // Proceedings of the 18th International Conference on Data Engineering, 2002. Vol. 7, No. 3, Pp. 673–684.

[13] Chen L., Raymond N. On the Marriage of Lp-norms and Edit Distance. // Proceedings of the Thirtieth International Conference on Very Large Data Bases.Toronto, Canada. 2004. Vol. 30, Pp. 792–803.

[14] Chen L., Özsu M., Vincent O. Robust and Fast Similarity Search for Moving Object Trajectories. // Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data.Baltimore, Maryland. 2005. Vol. 30, Pp. 491–502.

[15] Frentzos E., Gratsias K., Theodoridis Y. Index-based most similar trajectory search. // Proceedings of the VLDB Endowment. 2008. ICDE 2007. Pp. 816-825.

[16] Morse M., Patel J. An efficient and accurate method for evaluating time series similarity$_{\text{ffl}}$. $//Proceedingsofthe2007ACMSIGMODInternationalConferenceonManagementofD$ -580.

[17] Chen Y., Nascimento M., Ooi B, Tung A. On Shape-based Pattern Detection in Streaming Time Series. // IEEE Transactions on Knowledge Data Engineering. 2007. Vol. 24, No. 2, Pp. 265-278.

[18] Bellet A., Habrard A., Sebban M. A Survey on Metric Learning for Feature Vectors and Structured Data. Available at www.arxiv.org (accessed: May 18, 2015).

[19] Yang L., Jin R. Distance Metric Learning: A comprehensive survey . Available at www.arxiv.org (accessed: May 18, 2015).

[20] Yang L., Wayne Z., Hongfei Y., Xiaoming L. A Metric Learning Based Approach to Evaluate Task-specific Time Series Similarity. // Proceedings of the 14th International Conference on Web-Age Information Management.Beidaihe, China. 2013. Pp. 314–325.

[21] Mahalanobis P. C. On the generalised distance in statistics. // Proceedings National Institute of Science.India. 1936. Vol. 2, No. §. Pp. 49–55.

[22] Prekopcsak Z., Lemire D. Time Series Classification by Class-Specific Mahalanobis Distance Measures. // Advances in Data Analysis and Classification , 2012. Vol. 6, No. 3. Pp. 185–200.

[23] Garreau D., Lajugie R., Bach F., Arlot S. Metric Learning for Temporal Sequence Alignment // Advances in Neural Information Processing Systems. 2014. Vol. 27. Pp. 1817–1825.

[24] Vandenberghe L., Boyd B. Semidefinite Programming. // SIAM REVIEW, 2004. Vol. 38, Pp. 49–95.

[25] Киселев А. Н., Стрижов В. В. Расстояние между временными рядами как расстояние между функциями распределения параметров их моделей. Available at http://rlu.ru/5K7n (accessed: June 22, 2015).

[26] Wei D., Jiang Q., Wei Y., Wang S. A novel hierarchical clustering algorithm for gene sequences. // BMC Bioinformatics, 2013. Vol. 13, Pp. 174–189.

[27] Solovyov A., Lipkin WI. A Centroid based clustering of high throughput sequencing reads based on n-mer counts. // BMC Bioinformatics, 2013. Vol. 14, Pp. 268–289.

[28] Liao T., Ting C.-F., Chang P.-C. An adaptive genetic clustering method for exploratory mining of feature vector and time series data. // International Journal of Production Research, 2006. Vol. 44, Pp. 2731—2748.

[29] Цыганова С. В. Локальные методы прогнозирования с выбором преобразования // Машинное обучение и анализ данных, 2012. Т. 1, 3. С. 311–317.

[30] Кононенко Д. С. Оценка параметров инвариантных преобразований в задачах прогнозирования временных рядов. Магистерская диссертация, Московский физико-технический институт, 2013, 24 с. http://rlu.ru/5K7g (дата обращения: Июнь 22, 2014).

[31] Wang K., Gasser T. Aligment of curves by dynamic time warping // The Annals of Statistics, 1997. Vol. 25, No. 3. Pp. 1251–1276.

[32] Rong Tang, Hans-Georg Muller. Pairwise curve synchronization for functional data // Biometrika, 2008. Vol. 95, No. 4. Pp. 875–889.

[33] Ana Arribas-Gil, Hans-Georg Muller. Pairwise dynamic time warping for event data // Computational Statistics Data Analysis, 2014. Vol. 69. Pp. 255–268.